INFERENCE OF PHYLOGENY USING DISCRETE MORPHOLOGICAL CHARACTERS

By
Basanta Khakurel

A Thesis Submitted to the Faculty
of Southeastern Louisiana University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Biology

Southeastern Louisiana University

Hammond, Louisiana

June 2023

INFERENCE OF PHYLOGENY USING DISCRETE MORPHOLOGICAL CHARACTERS

By
Basanta Khakurel

Approved:

April M. Wright Associate Professor of Biological Sciences (Director of Thesis)	Brian I. Crother Professor of Biological Sciences (Committee Member)
Kyle R. Piller Professor of Biological Sciences (Committee Member)	Gary Shaffer Professor of Biological Sciences (Committee Member)
Justin R. Anderson Department Head of the Department of Biological Sciences	Daniel R. McCarthy Dean of the College of Science and Technology

Name: Basanta Khakurel

Previous degrees: B.S., Southeastern Louisiana University, 2021 (Biology)

Date of Current Degree: July 31, 2023

Institution: Southeastern Louisiana University

Major Field: Biological Sciences

Major Professor: Dr. April M. Wright

Title of Study: INFERENCE OF PHYLOGENY USING DISCRETE MORPHOLOGICAL CHARACTERS

Pages in Study: 94

Candidate for Degree of Master of Science

Morphology is essential in phylogenetic studies specially in taxa involving deep relationships in the past. As molecular sequences are unavailable for extinct taxa, morphology is the only source for them. Morphological characters are most often analysed using parsimony methods, which minimizes the number of character changes across the tree. But recent development of probabilistic methods has enabled the use of models in the inference of phylogeny using both molecular or morphological data. Although there has been significant development of models in molecular tree inference, morphologists typically use the Mk model and its variations in phylogenetic tree inference. The first chapter gives an introduction to the use of morphological characters in tree estimation. In the second chapter I have highlighted the influence of observer bias in estimating phylogenetic trees using the Mk model. The Q-Matrix which is an integral part of the model during estimation of trees using Bayesian methods could be incorrectly partitioned due to the lack of understanding in how many character states actually exists. This may lead to incorrect partitioning of the Q-Matrix. This is a simulation study in which we incorrectly partitioned simulated

datasets and inferred trees using the model.

I further examine the use of another model for tree inference using morphological data in the third chapter. The model is named Site-Heterogeneous Discrete Morphology which allows to model different character frequencies for each character within one dataset. This allows researchers to assume different rates for different character states with respect to among character frequency variation. I have explained the model and compared it to the widely used Mk model using Stepping stones method, posterior predictive simulation and reversible jump Markov chain Monte Carlo using empirical dataset of ants.

Key words: fossilized birth-death, model testing, phylogenetic posterior prediction, Bayesian inference, morphological characters, Mk model.

Dedication

I would like to dedicate this thesis to my dearest ones. My parents Bhim and Shanta Khakurel, my brother Barshat Khakurel, and my lovely girlfriend Dipika Ghimire for believing in me and supporting me. My family is 8500 miles away but they always made me feel like I'm with them. With their support, I was able to do what I wished.

Acknowledgements

First of all, I would like to thank Dr. April Wright for believing in me and letting me continue in the lab as a graduate student. Her guidance, support has inspired me in every step of this journey.

Second, I would also like to thank my committee members Dr. Brian Crother and Dr. Kyle Piller for their invaluable guidance and wisdom.

Additionally, I would like to thank Dr. Orlando Schwery and Dr. Brenen Wynd for their guidance that helped me progress in my research. I would also like to thank all the current and former lab members from the Wright lab.

Finally, I would like to thank all the friends I made in Louisiana who made me feel like I was never out of home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, I have executed some experiments that highlight the use and importance of model choice when using morphological data in phylogenetic studies. In this section, I provide an overview of the use of morphology and models of character evolution in phylogenetic estimation. In the rest of the thesis, I demonstrate avenues to improve the robustness of phylogenetic analyses of morphological data.

Phylogenetic trees are essential as they help us establish a historical context in which to study organismal form and function. For almost all deep-time phylogenetic analysis the only source of information is the fossil record. The oldest DNA that is sequenced dates back to 1 million years (Bailleul and Li, 2021) but for organisms that have existed beyond that time the source of information are fossils that cannot be sequenced. Thus, morphological character data is our only way to directly incorporate the fossil record in building a phylogenetic tree. Morphological character data has thus been an irreplaceable resource in building phylogenetic trees. Despite that, there is no well-defined mechanistic model of evolution for morphological characters. This

is because unlike molecular sequences, there might be different number of character states for different characters. There can also be character states that are found in one clade but not the others, often corresponding to gain of derived states or loss of ancestral ones. Whole morphological characters can also be lost in clades, rendering them inapplicable across parts of the tree (Brazeau, 2011; Hopkins and St. John, 2021). This complicates the derivation of a general morphological evolution model.

Phylogenetic trees have been long built looking at the morphological similarities between taxa. With the development of sequencing techniques such as Sanger sequencing and most recently with Next Generation Sequencing, molecular techniques for phylogenetic tree inference gained traction. During that time, many mechanistic models of evolution were developed involving molecular data. For example, JC69 (Jukes and Cantor, 1969), F81 (Felsenstein, 1981), HKY (Hasegawa et al., 1985), GTR (Tavaré, 1986). However, there has been a recent push to re-incorporate morphological data with molecular data to understand the evolutionary history of a group of taxa. Methods such as divergence time estimation have become important (O'Reilly et al., 2015; Smith et al., 2018; Barido-Sottani et al., 2020a). These involve the understanding of historical biology, and morphology is the predominant source of information for understanding the ecology and evolution of organisms in the past. Such methods require morphological data to date nodes or tips to obtain accurate estimation of the ages. One of the models that has been widely used for tip-dating methods is the Fossilized Birth-Death model (Heath et al., 2014). For data sources such as DNA, RNA and amino acids there has been a significant development in models and expectations but morphology has lagged in phylogenetic estimation (Spencer and Wilberg, 2013). With development of advanced total-evidence methods, the necessity of a

robust model of character evolution for morphology has increased.

1.1 Morphological Data

Molecular sequence alignment is the process of assessing homology within a DNA, Amino Acid, or protein sequence across two or more organisms, such that each column in a sequence alignment can be assumed to be homologous. A column of molecular sequence alignment is a single site that is positioned with the sequenced DNA. Homology is assigned using multiple sequence alignment in such a manner that all nucleotides in each column can be assumed to be the descendants of an ancestral site. This is similar to morphological data. Each column in a morphological data matrix is a character and a row is a character state, the trait of an individual organism. For example, in discrete characters, a state '1' could mean a structure is present in one individual and a state of '0' could mean that the state is absent in an individual. A column in a data matrix is similar to a column in an alignment. For a molecular character, the character states are the ACGTs (or amino acids) in that column.

Morphological characters are generally collapsed into discrete character states rather than using continuous character data. Around 60% of the morphological matrix used in phylogenetic studies are binary with the characters 0 and 1 (Barido-Sottani et al., 2020b). Due to arbitrary coding of the characters, the magnitude of change between the characters are very unlikely to be equal. For example, the number of teeth in a mammal's mouth and the shape of a mammal's ear. The number of teeth can vary greatly from as few as 20 in a platypus to as many as 44 in the hippopotamus. Also, the shape of a mammal's ear can also vary greatly, from rounded ears of

a mouse to the pointed ears of a bat. If we were to code these two characters using a binary system, we would have to assign the value 0 to *few teeth* and 1 to *many teeth*. Similarly, for the ears, we might assign 0 for *rounded* and 1 for *pointed*. The magnitude of change between the character states can vary greatly. A change from few teeth to many teeth is likely to be a much larger change than the change from rounded ear to pointy ear. We could model the evolution of characters in a trait to be more likely to another with this knowledge *a priori*. But in practice, most studies model the changes in all traits to be equal which might not be the biological reality. This modeling issue could lead to inaccuracy in estimating phylogenetic relationships.

Even though the magnitude of change can be taken into account while coding for characters, we need to ensure that the characters are modeled correctly with respect to how they evolve.

In addition to that, the state space for morphological data is subjective. Considering the similar example from above, for the number of teeth in the mouth of a given mammal, there can be a finite number of possibilities. But for the ear there could be infinite number of possible shapes which would lead to difficultly in defining each state precisely. This subjectivity of state space makes it difficult in determining the relationships between different taxa. This can result in model misspecification, which is further discussed in Chapter Two.

Morphological data remains important in phylogenetic studies as most organisms of the past are extinct and to understand the relationship between these organisms, we need to rely on morphology. Traditionally to include fossils in a tree, calibration points (fossil ages) were used but recent advancements have allowed the use of morphological characters obtained from fossils as well (Pyron, 2011; Ronquist et al., 2012; Heath

et al., 2014). These methods have proven to be important in empirical studies and one of the major component is the morphological data.

1.2 Use of Morphological Data

Morphology is a essential in time estimation studies, but we do not have a mechanistic model of morphological evolution. In molecular studies the state space (DNA [ACGT], RNA [ACGU] or amino acids) are known. The researcher knows all the possible molecular character states for every character. This allows researchers to make assumptions about the relationships about how these states relate to each other. With biochemical experiments, we now know that the rates of transitions (purine to purine or pyrimidine to pyrimidine) and transversions (purine to pyrimidine and vice versa) are different (Lanave et al., 1986; Lydeard and Roe, 1997). With this knowledge, we can model the Q-matrix accordingly, allowing for more transition than transversions. In contrast to that, morphological data cannot rely on this knowledge (Brazeau, 2011). Morphological data are often obtained from fossils and the sampling density of the taxa of interest determines the ability to correctly identify the number of character state present for a trait. There have been various studies about the model of character evolution in morphological phylogenetics (Wright and Hillis, 2014; Wright et al., 2016; Bapst et al., 2018; Klopfstein et al., 2019) which suggest that character coding plays an important role in a model's plausibility for a dataset. The number of character states in a character determines the number of possible transitions a character can make. For example, a change from state 1 to state 2 is impossible is character 2 does not exist at all. In likelihood based models,

the rate of change between character states is coded in the Q-matrix, which contains the number and relative probabilities of change between different character states. There are various ways the data can be coded and such different methods could make a statement about the process of evolution. How characters are coded changes the models that may be considered for the data even before a model of evolution is chosen in an analysis.

Models of evolution further make assumptions about character evolution. In most molecular and morphological analyses, a transition rate matrix called Q-matrix is used to model the exchangeabilities between the character state (Felsenstein, 1981; Lewis, 2001). The Q-matrix can range from very simple assumptions to complex assumptions about the process of evolution. For example, for molecular studies, the Jukes-Cantor (JC69) model (Jukes and Cantor, 1969) is the simplest model containing just 1 parameter and assumes equal rate of transition between any character state. Similarly, the most complex model for molecular studies, the General Time Reversible model (Tavaré, 1986) contains 10 parameters and even the rates between each nucleotide can be changed per the data. Only JC69 model has been applied to morphological data retaining all the assumptions: the equilibrium frequencies of all characters are the same, and that all changes between characters states are equally likely. Complex models of molecular sequence evolution such as HKY (Hasegawa et al., 1985) and GTR (Tavaré, 1986) has yet to be applied in morphology.

The use of morphological data in phylogenetic studies using probabilistic methods was pioneered by the development of the Mk model (Lewis, 2001). This model, as mentioned before, is a generalization of the Jukes-Cantor model of DNA sequence evolution (Jukes and Cantor, 1969), and assumes an equal rate of transition between

A

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} \\ \mu_{10} & -\mu_1 \end{pmatrix}$$

B

$$Q = \begin{pmatrix} -\mu_0 & \mu_{01} & \mu_{02} & \mu_{03} \\ \mu_{10} & -\mu_1 & \mu_{12} & \mu_{13} \\ \mu_{20} & \mu_{21} & -\mu_2 & \mu_{23} \\ \mu_{30} & \mu_{31} & \mu_{32} & -\mu_3 \end{pmatrix}$$

C

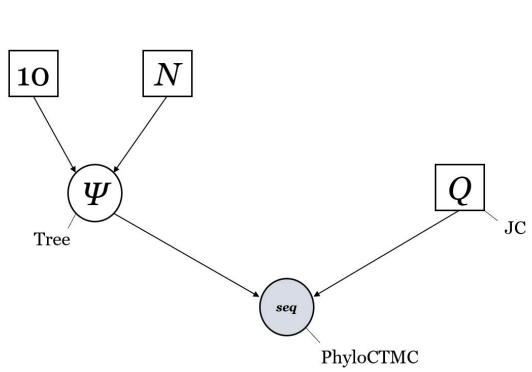


Figure 1.1: A. A Q-Matrix for Binary characters under Mk model. B. A Q-Matrix under Mk model for Multistate characters. C. A graphical representation of the Mk model (Figure referred from Höhna et al. (2016))

any two character states. The Mk model contains one parameter, the rate of transition between the character states. The model can be extended to an arbitrary number of k character states. Due to the simple nature of the model assumptions, some researchers have raised concerns about the realism of the Mk model (Goloboff et al., 2018b,a) and much of the work has been focused on comparing the model to parsimony methods (Wright and Hillis, 2014; Brown et al., 2017; Puttick et al., 2017; Schrago et al., 2018).

1.3 The Fossilized Birth Death Model

The Fossilized Birth-Death model can be readily separated into its component models: the substitution model, the clock model, and the tree model (Warnock and Wright, 2020). To jointly estimate the topology and the node ages, researchers generally assume a tripartite model of evolution: one that describes the evolution of character

data, second that describes the evolutionary rates across the tree and the third that describes the diversification process leading to the observed tree.

The benefit of this type of approach is that the researchers can treat each of the component as a discrete inferential module and also provides the flexibility to combine models according to the data in use (sequence data, character data or both). This type of approach also allows for the use of fossil taxa in addition to molecular and/or morphological character information.

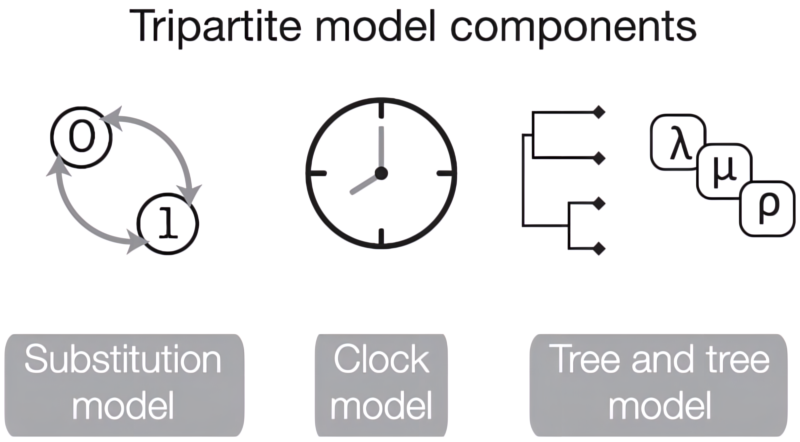


Figure 1.2: The key components of a Fossilized Birth-Death model. Figure from Warnock and Wright (2020)

In this model, different parts (components) are informed by different parts of the data. The character model is informed by the characters in a matrix and the tree model is informed by the ages of fossils. The next portion of the model are the model components themselves as shown in 1.2. As this study deals with morphology only, I have highlighted some important aspects of it below, for a more detailed explanation of other model components see Warnock and Wright (2020).

The wider availability of models as such and its incorporation in various phylo-

genetic softwares has spurred interest in fuller incorporation of morphological data (Barido-Sottani et al., 2022; Mongiardino Koch et al., 2021; Gavryushkina et al., 2017). But modeling morphological characters come with some challenges. Molecular characters have same nucleotide bases ([A,C,G,T] for DNA) across all taxa that has existed but for morphological characters, different groups of organisms have different traits that are coded in the character matrix. As morphological data do not have a common meaning across a range of taxa like the molecular data does, morphological characters have largely been confined to the use of Mk model (Lewis, 2001) for discrete character evolution. This model is the morphological counterpart of the JC69 model which assumes that the rate of change between the character states are equal. There has been several variations of this model which allows researchers to relax the assumptions of the rates of changes (Nylander et al., 2004; Wright et al., 2016). In addition to the study of discrete morphological data, continuous characters have also been getting traction in phylogenetic inference (Goloboff et al., 2006; Parins-Fukuchi, 2018). The evolution of continuous characters can be modeled under Brownian motion (Felsenstein, 1973, 1985; Gingerich, 1993), or Ornstein-Uhlenbeck (Beaulieu et al., 2012; Hansen, 1997; Butler and King, 2004). These models allow for changes to accumulate continuously along branches.

As morphological characters are being used more in time estimation studies as FBD, it is important to understand the mechanisms of modeling morphology. This study highlights some important aspects of modeling morphology. In the second chapter, I performed experiments to simulate incorrect partitioning of the Q-matrix. This revealed that it is certainly important to correctly size the Q-matrix or it would lead to inaccurate estimation of phylogenetic relationships. In the third chapter, I

introduce a novel model for estimating phylogenies using morphological data using beta and Dirichlet priors. This is similar to Wright et al. (2016) in which the author has used beta prior for binary characters. In addition to that, I have performed an in-depth look at model choice to compare this model with widely used Mk model.

Bibliography

Bailleul, A. M. and Z. Li. 2021. Dna staining in fossil cells beyond the quaternary: Reassessment of the evidence and prospects for an improved understanding of dna preservation in deep time. Earth-Science Reviews 216:103600.

Bapst, D. W., H. A. Schreiber, and S. J. Carlson. 2018. Combined analysis of extant rhynchonellida (brachiopoda) using morphological and molecular data. Systematic Biology 67:32–48.

Barido-Sottani, J., J. A. Justison, A. M. Wright, R. C. Warnock, W. Pett, and T. A. Heath. 2020a. Estimating a time-calibrated phylogeny of fossil and extant taxa using revbayes.

Barido-Sottani, J., A. Pohle, K. De Baets, D. Murdock, and R. C. Warnock. 2022. Putting the f in fbd analyses: tree constraints or morphological data? bioRxiv Pages 2022–07.

Barido-Sottani, J., N. M. Van Tiel, M. J. Hopkins, D. F. Wright, T. Stadler, and R. C. Warnock. 2020b. Ignoring fossil age uncertainty leads to inaccurate topology and divergence time estimates in time calibrated tree inference. Frontiers in Ecology and Evolution 8:183.

Beaulieu, J. M., D.-C. Jhwueng, C. Boettiger, and B. C. O'Meara. 2012. Modeling stabilizing selection: expanding the ornstein–uhlenbeck model of adaptive evolution. Evolution 66:2369–2383.

Brazeau, M. D. 2011. Problematic character coding methods in morphology and their effects. Biological Journal of the Linnean Society 104:489–498.

Brown, J. W., C. Parins-Fukuchi, G. W. Stull, O. M. Vargas, and S. A. Smith. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on puttick et al. Proc. R. Soc. B 284:20170986.

Butler, M. A. and A. A. King. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. The American Naturalist 164:683–695.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet 25:471–92.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Felsenstein, J. 1985. Phylogenies and the comparative method. The American Naturalist Pages 1–15.

Gavryushkina, A., T. A. Heath, D. T. Ksepka, T. Stadler, D. Welch, and A. J. Drummond. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Systematic biology 66:57–73.

Gingerich, P. D. 1993. Quantification and comparison of evolutionary rates. American Journal of Science 293:453.

Goloboff, P. A., C. I. Mattoni, and A. S. Quinteros. 2006. Continuous characters analyzed as such. Cladistics 22:589–601.

Goloboff, P. A., M. Pittman, D. Pol, and X. Xu. 2018a. Morphological data sets fit a

common mechanism much more poorly than dna sequences and call into question the mkv model. Systematic Biology Page syy077.

Goloboff, P. A., A. Torres, and J. S. Arias. 2018b. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. Cladistics 34:407–437.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160–174.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:E2957–E2966.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology 65:726–736.

Hopkins, M. J. and K. St. John. 2021. Incorporating hierarchical characters into phylogenetic analysis. Systematic Biology 70:1163–1180.

Jukes, T. and C. Cantor. 1969. Evolution of protein molecules. Mammalian Protein Metabolism 3:21–132.

Klopfstein, S., R. Ryer, M. Coiro, and T. Spasojevic. 2019. Mismatch of the morphology model is mostly unproblematic in total-evidence dating: insights from an extensive simulation study. BioRxiv Page 679084.

Lanave, C., S. Tommasi, G. Preparata, and C. Saccone. 1986. Transition and transversion rate in the evolution of animal mitochondrial dna. Biosystems 19:273–283.

Lewis, P. O. 2001. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. Systematic Biology 50:913–925.

Lydeard, C. and K. J. Roe. 1997. The phylogenetic utility of the mitochondrial cytochrome b gene for inferring relationships among actinopterygian fishes. Molecular systematics of fishes 1:285–303.

Mongiardino Koch, N., R. J. Garwood, and L. A. Parry. 2021. Fossils improve phylogenetic analyses of morphological characters. Proceedings of the Royal Society B 288:20210044.

Nylander, J. A., F. Ronquist, J. P. Huelsenbeck, and J. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47–67.

O'Reilly, J. E., M. Dos Reis, and P. C. Donoghue. 2015. Dating tips for divergence-time estimation. Trends in Genetics 31:637–650.

Parins-Fukuchi, C. 2018. Use of continuous traits can improve morphological phylogenetics. Systematic Biology 67:328–339.

Puttick, M. N., J. E. O'Reilly, A. R. Tanner, J. F. Fleming, J. Clark, L. Holloway, J. Lozano-Fernandez, L. A. Parry, J. E. Tarver, D. Pisani, et al. 2017. Uncertain-

tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. Proc. R. Soc. B 284:20162290.

Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology Page syr047.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–542.

Schrago, C. G., B. O. Aguiar, and B. Mello. 2018. Comparative evaluation of maximum parsimony and bayesian phylogenetic reconstruction using empirical morphological data. Journal of evolutionary biology 31:1477–1484.

Smith, S. A., J. W. Brown, and J. F. Walker. 2018. So many genes, so little time: a practical approach to divergence-time estimation in the genomic era. PloS one 13:e0197433.

Spencer, M. R. and E. W. Wilberg. 2013. Efficacy or convenience? m odel-based approaches to phylogeny estimation using morphological data. Cladistics 29:663–671.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Some Mathematical Questions in Biology: DNA Sequence Analysis 17:57–86.

Warnock, R. C. and A. M. Wright. 2020. Understanding the tripartite approach to Bayesian divergence time estimation. Cambridge University Press.

Wright, A. M. and D. M. Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS One 9:e109210.

Wright, A. M., G. T. Lloyd, and D. M. Hillis. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. Systematic Biology 65:602–611.

Chapter 2

The fundamental role of character coding in Bayesian morphological phylogenetics [1]

Abstract

Phylogenetic trees establish a historical context for the study of organismal form and function. Most phylogenetic trees are estimated using a model of evolution. For molecular data, modeling evolution is often based on biochemical observations about changes between character states. For example, there are four nucleotides, and we can make assumptions about the likelihood of transitions between them. By contrast, for morphological characters, we may not know *a priori* how many characters states there are per character, as both extant sampling and the fossil record

[1] This chapter was submitted as Khakurel et. al to Systematic Biology.

may be highly incomplete, which leads to an observer bias. For a given character, the state space may be larger than what has been observed in the sample of taxa collected by the researcher. In this case, how many evolutionary rates are needed to even describe transitions between morphological character states may not be clear, potentially leading to model misspecification. We simulated character data with varying numbers of character states per character. We then used the data to estimate phylogenetic trees using models of evolution with the correct number of character states and an incorrect number of character states. The results of this study indicate that this observer bias may lead to phylogenetic error, particularly in the branch lengths of trees.

2.1 Introduction

Molecular phylogenetics relies on known state spaces (DNA [ACGT], RNA [ACGU] or amino acids). In this case, the researcher knows all molecular character states that are possible at a character. As I will discuss below, the ability to know the number of character states per character enables researchers to make a variety of assumptions about how these states relate to each other, character change rates, and character change probabilities. Morphological data cannot necessarily rely on this knowledge (Brazeau, 2011). Much data is recovered from fossils, where the density of our sampling affects our ability to correctly identify how many states are present for a character. For example, we simply may not observe certain character states if we have few complete samples recovered from the fossil's range. Or, perhaps a character state occurs in a clade that has not been sampled, or sampled from complete enough

specimens to find the character (Fig. 2.1). This can lead to misleading estimates of phylogeny and diversification metrics from trees in the fossil record (Wagner, 2000; Ciampaglio et al., 2001; Flannery Sutherland et al., 2019). Additionally, observer bias, a phenomenon when the limitations or prior expectations of the observer (i.e., an individual coding morphological characters) colors the observations produced, may obscure the correct number of character states. This may occur, for example, if a character is somewhat cryptic to human eyes, such as infrared coloration in butterflies (Stavenga and Arikawa, 2006), resulting in under-reporting of variation. Alternatively over-splitting of variation that is more recognizable to us as human observers has also been documented (Keating, 1985). In this study, I aim to understand the effects of inappropriate understanding of character state spaces on phylogenetic inference.

While much has been written about the role of the model of character evolution in morphological phylogenetics (Wright and Hillis, 2014; Wright et al., 2016; Bapst et al., 2018; Klopfstein et al., 2019), character coding plays a role in which character models are plausible for a dataset. The number of possible state transitions a character can make is determined by how many states are present for that character. For example, a change from a 0 state for a character to a 2 state is simply impossible if the 2 character state does not exist (Fig. 2.2). In a likelihood-based model, possible changes between character states will be codified in the Q-matrix, which explicates the number and relative probabilities of different character-to-character changes (Fig. 2.2). It is assumed in most models that the number of states (often called k) is known without error.

Assumptions about character states determine whether the transition rates between the character states are similar or different, whether a heterogeneous rate is

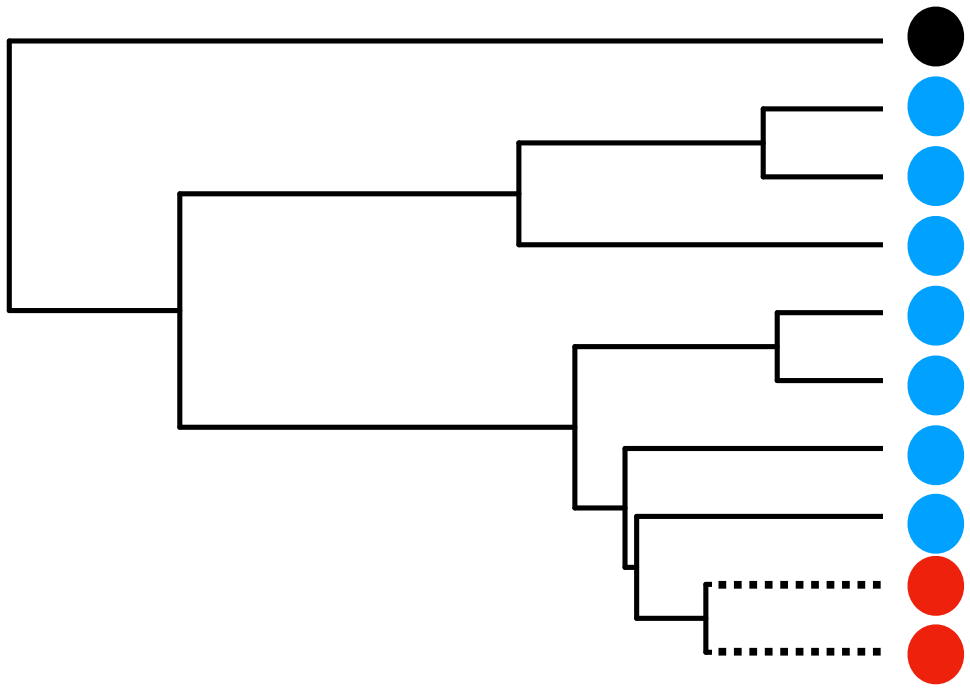


Figure 2.1: This figure displays a fundamental difficulty with characterizing a morphological state space. Unsampled lineages are indicated with dotted edges. In this case, there is a single character with three states (black circles, blue circles, and red circles). As the lineage containing red circles is unsampled, one may assume that the state space only includes two states, and thus any Q-matrix generated by a researcher from the sampled data will not appropriately represent the character state space.

allowed for different characters, and whether the character state is conserved or not. For example, if a character state is lost on a branch, then observed in the descendants of that branch and coded as the same character state, it will be assumed to be a reversal or regain of that character state. If the researcher codes the reversal as a new state, as one might do for a Dollo process, this is no longer a regain of the character state, but the innovation of a new character state (Gould, 1970; Goldberg and Igić, 2008). In this case, the state space of the phylogenetic model must be larger, implying a model with more possible changes between characters (Fig. 2.2). In this way, choices made about the homology statements of a character implicitly make a statement about the process of evolution. How characters are coded changes the models that may be considered for the data, even before a model of evolution is chosen in an analysis.

As an example of this, imagine a character, such as egg-laying in reptiles. This character is often coded as a two-state character (oviparity and viviparity), with the root of the tree generally assumed to be oviparous (Wright et al., 2015). Therefore, any regain of oviparity in a clade that is viviparous is considered a re-evolution of the oviparity character state, rather than a potentially new character state. In this case, the number of transitions possible will be that of a binary character, as opposed to a multistate character. However, if the researcher has chosen to code the character as a multistate, polar character (Stevens, 1980), in which states are expected to be ordered, or a Dollo character, which is expected not to reverse, then a simple binary model of substitution is no longer adequate. In these cases, the reappearance of oviparity in a viviparous clade must be coded as a new character state, necessitating a Q-matrix with a larger state space. This can be visualized on Figure 2.2. As shown

Figure 2.2: At left is a multistate character for which only two character states are included in the model. This is how we would construct a Q-matrix for the trait in Fig. 1. In the case of an unordered model, it is assumed that backwards and forwards transitions are allowed between all states. In the case where one state is not observed, in this case state 3, transitions to and from that character are not considered under the model. In this case, over half of the possible character state changes are removed by failing to sample the third state.

on Figure 2.3, this can lead to misestimation of branch lengths.

Models of evolution then make further assumptions about character evolution. In most modern molecular and morphological analyses, a transition rate matrix — also called Q-matrix— is set up to model changes between the different character states (Felsenstein, 1981; Lewis, 2001). This Q-matrix, at minimum, specifies the exchangeabilities between character states. A Q-matrix can range from making very simple assumptions about the process of evolution. For example, the Jukes-Cantor (JC) model of sequence evolution (Jukes and Cantor, 1969), is the simplest models assuming equal rates of transitions between any character state, and is used for both molecular sequence data and morphological data. Let us focus on molecular data first. In a nucleotide dataset, the JC model assumes that all the bases (A, T, G, C) have the same frequency and the rates for their transition is the same. That is, there are

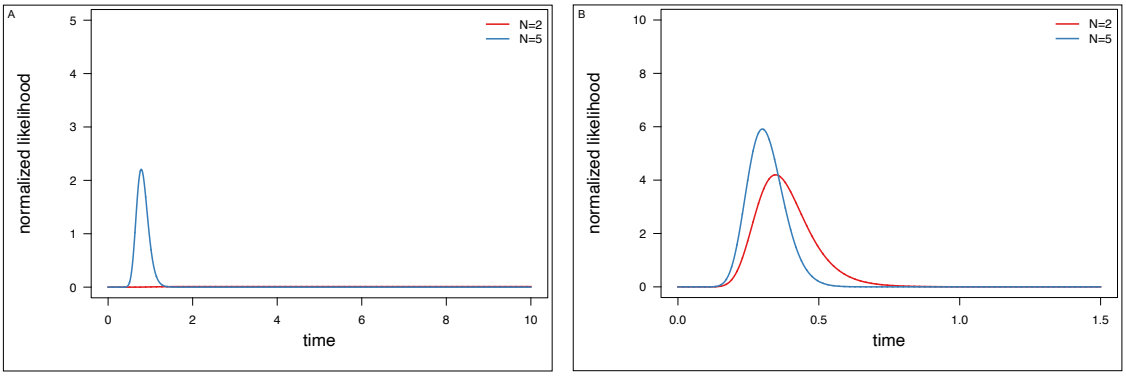


Figure 2.3: Likelihoods of branch lengths given a number of mismatches between the state space and the Q-matrix. In the graphic at left, there are 75 characters for which the Q-matrix is correctly-parameterized (either 2 or 5 characters) and 25 for which it is not. At right, there are 50 characters for which the model is correctly specified and 50 for which it is not. In these cases, branch lengths (represented by time) tend to be wrongly estimated.

the same number of each base type, and each base is equally likely to change to any other base type. When this was applied to morphological data (Lewis, 2001), these assumptions were retained: that the equilibrium frequencies of all characters are the same, and that all changes between character states are equally likely. More complex models, such as the Felsenstein 81 model (Felsenstein, 1981) have been applied to morphological data (Nylander et al., 2004; Wright et al., 2016), and assume characters may have differential change rates as a function of their frequencies. Models such as the General Time Reversible model (GTR) (Tavaré, 1986), which is among the more complex models, have not been - and should not be - applied to morphological data. This is because coding by human interpretation of state is inherently arbitrary, and likelihoods of the morphology models should be invariant to how the states are coded.

The Q-matrix is a core component of the phylogenetic model, specifying the transition rates of different types of evolutionary changes in the observed dataset.

Therefore, we might expect that error in correctly-sizing the Q-matrix could lead to problems in estimating a phylogeny correctly. There are several ways this error could arise. As covered above, sampling error could lead to misunderstanding of the state space. Additionally, for molecular data, the state space can be assumed to be constant across sites. It is generally presumed that any specific nucleotide can occur at any site, whether or not it is observed to do so. For amino acids, mixture models such as the CAT model (Lartillot and Philippe, 2004) can be used to virtually reduce the state space for sites. This is not the case in morphological data, where different characters, by their nature will have different numbers of states. Some may be presence/absence, others may be multistate. Therefore, the Q-matrix cannot be treated as invariant across characters, and the dataset may need to be split up according to the state space of the character. Without doing so, this may lead to characters being modeled under incorrect Q-matrices.

In this study, I used simulations to assess two issues: The first being assuming an inappropriately-small Q-matrix. This simulates the effect in Fig. 2.1, observer bias in the number of character states. The second is failing to account for Q-matrix heterogeneity by not breaking up data matrices by character state space. This will lead to the assumption that all characters evolved using to the largest state space. For many characters, this will mean the state space is overly-large. For example, if a character is binary, but the largest number of character states in the matrix is 7, the model will assume there are additional 5 character states for the binary characters that simply have not been observed. This would imply far more evolutionary transitions are possible than truly are. On the other hand, if we have a too-small state space, we can end up underestimating the number of evolutionary transitions. We might

expect to see this affect branch lengths or topology. Finally, I looked at a set of simulations under conditions consistent with long-branch attraction (LBA). In this chapter, I have highlighted the consequences of the observer bias in phylogenetic tree reconstruction.

2.2 Methods

2.2.1 Simulations

I simulated datasets with different numbers of character states possible per character. The datasets were simulated using an empirical tree. This tree comes from a fairly averaged-sized paleontological dataset of 41 taxa and 42 characters (Barden and Grimaldi, 2016). These small dataset sizes are fairly standard for morphological character matrices (Wright et al., 2016; Barido-Sottani et al., 2019). Characters were simulated under the Mk model of morphological evolution (Lewis, 2001) using the software RevBayes (Höhna et al., 2014; Höhna et al., 2016). I simulated two dataset sizes, 44 characters (the size of the true Barden and Grimaldi dataset) and 100 characters.

I simulated the data with four categories of Gamma-distributed among-site rate variation (ASRV), as most empirical studies include this. To examine the effect of the base Q-matrix size, I simulated data under Q-matrices with either 2, 3, 4 or 5 states, in varying proportions as described in the three following sections. I simulated 100 datasets for each dataset and Q-matrix size.

Inclusive Sampling

Under the inclusive character sampling scheme, I simulated characters given a specific Q-matrix, but did not remove characters that didn't have the maximum size in the Q-matrix. For a dataset with Q-matrix state-space 4, it would be possible to have 2-state, 3-state and 4-state characters. The resulting character matrix would be partitioned into 3 datasets, corresponding to three state spaces. I also estimated trees using a mis-specified model. In the mis-specified model, all the characters were left in a single partition, with the Q-matrix sized according to the character with the largest state number. This will specify too many possible transitions for many of the characters in such a matrix, though the exact proportion of mis-specified characters will vary among simulations. In this simulation scheme, I simulated under state spaces of 2 (no misspecification possible), 3, 4 and 5.

Rejection Sampling

Under the rejection sampling scheme, I simulated matrices with defined numbers of characters from each Q-Matrix size. For each dataset, either 50% or 75% of the dataset was binary. The remainder would be made of characters with either 3, 4 or 5 character states. For each matrix size, I rejected and re-simulated any characters that did not have the maximal value of states. I did this because when simulating under a Q-matrix with size 5, it is possible to simulate a 4-state character. Under the rejection sampling scheme, this character would be removed and re-simulated in order to maintain maximal model mis-specification. I estimated trees from the simulated data using a partitioned model. In this model both the binary and the larger character states are parameterized with a correctly-sized Q-matrix. I also estimated

trees using an unpartitioned model as described in section Inclusive Sampling, in which all characters are put together in a single partition using the Q-matrix of the largest state. In the unpartitioned model, I am therefore guaranteed that some proportion of the dataset will have a mis-specified model. For datasets with binary and trinary characters, this mis-specification may be small. For datasets split between binary and five-state characters, it should be larger.

Incomplete Character Sampling

In order to examine the effect of unsampled character states, I ran a set of missing data simulations in which I replaced the largest character state with missing data. For example, if the largest state in the matrix was 4, all 4s were replaced with missing data. This simulated the effect shown in Fig. 2.1, in which one character state is unsampled in the focal clade, and therefore unrepresented in the analysis. In this case, the researcher is unaware of all the possible character states for a character and cannot specify the Q-matrix correctly. For example, if a character has three possible states, but only two have been sampled, the researcher will think that a binary model describes the trait best. For these, I estimated the tree using correctly-sized Q-matrices for the datasets without the unsampled character removed. Then, I re-estimated the tree with the unsampled character missing, and the size of the Q-matrix decremented by one (i.e., reflecting the observed state space).

Long-Branch Attraction

I also produced a set of simulations that approximate long-branch attraction. In these simulations, I used a four-taxon tree in which there are two regular branches and

two branches of three times their length. For these trees, I simulated 100 datasets of either 42 or 100 characters, using the method described in section Inclusive Sampling. As described in section **Incomplete Character Sampling**, I also did a round of simulations in which the largest character state had been removed.

2.2.2 Phylogenetic Estimation

Estimations were performed in RevBayes, under the same model the data were simulated under, except for the Q-matrix as described above. I ran the Markov chain for each dataset for up to 1,000,000 generations and assessed for convergence using Tracer (Rambaut and Drummond, 2011). I also checked for convergence using the R package Convenience (Fabreti and Höhna, 2022), which analytically calculates the effective sample sizes to reach convergence. This is a more objective convergence assessment diagnostic. The simulations were performed on the Louisiana Optical Network Initiative (LONI) High Performance Computing managed by Louisiana State University at Baton Rouge, LA.

2.2.3 Phylogeny Processing

I used the symmetric difference measure and the Robinson-Foulds distance (Robinson and Foulds, 1979, 1981) as implemented in DendroPy (Sukumaran and Holder, 2010) to compare the empirical tree (tree under which the data were simulated) with the trees estimated from this study. These two measures are often conflated, but provide two different data points: The symmetric difference compares the tree in topology, providing a whole-number measure of the number of differences between two or more trees under comparison. The Robinson-Foulds metric shows the topological distance

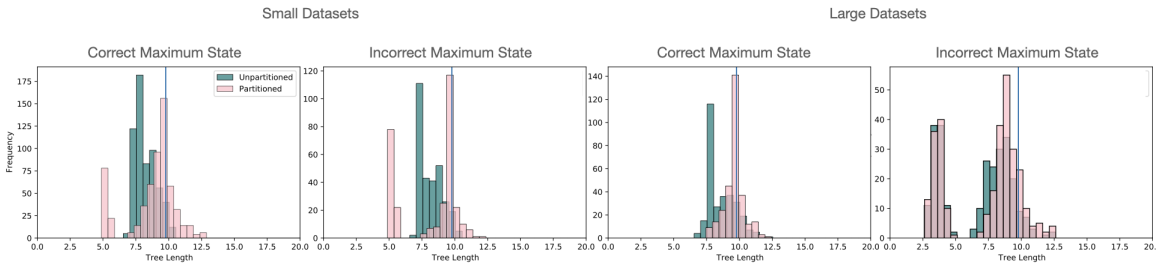


Figure 2.4: This figure shows the distribution of tree-lengths for each set of simulation conditions. Pink bars represent the correctly-specified Q-matrices (partitioned by number of character states); teal bars represent incorrectly-specified models (not partitioned by character states). In the correct maximum state models, all character states are sampled. In the incorrect maximum state models, the maximum state is unsampled (i.e., the true character state space is not known; analogous to the left-hand panel of Fig. 2.2). The true tree length (9.77) is indicated by the blue bar.

scaled by the branch lengths on the trees. Finally, tree length is the sum of branch lengths on a tree, providing a measure of total number of expected substitutions across the tree. I visualized the results with the Pandas (McKinney et al., 2011) and Seaborn (Waskom, 2021) Python libraries.

2.3 Results

2.3.1 Empirical Tree

Inclusive Sampling

For inclusive sampling, there was not a strong signal of difference between correctly-specified models and unpartitioned models (Fig. S1). In these datasets, the distribution of symmetric difference scores is distributed roughly the same for both the misspecified and correct models.

However, in the branch lengths (Fig. 2.4), there is an effect of partitioning. Branch

lengths are recovered more correctly more frequently for models with correct partitioning. Branches from the unpartitioned tree are often shorter than expected, an effect that is stronger in smaller datasets. The tree length of the mis-specified tree results in a shorter tree than the true empirical tree (Fig. 2.4). In most cases, the correctly-partitioned models are able to achieve a distribution of tree lengths that is centered on the true tree length (9.77 units) regardless of the dataset size. I see a significant effect between the missing and complete character states. Particularly for large datasets, it seems that if the largest possible character state is incorrect, this can lead to trees that are much shorter than the true tree, regardless of whether or not the remaining characters are correctly partitioned. As shown on Fig. 2.2, eliminating one character state greatly reduces the number of possible transitions per the Q-matrix. This can lead to an underestimate of the total number of expected changes per site. In this case, correct partitioning leads to a distribution of tree lengths that is closer to the true tree length.

Rejection Sampling

In the rejection sampled trees, the effect of partitioning is more strongly seen in the topology of the tree. As can be seen on Fig. 2.5, if 75% of the dataset consists of misspecified Q-Matrix and the remaining 25% has the correct Q-Matrix, the overall phylogenetic error is higher. In the datasets in which 50% of characters have a correctly-sized Q-matrix and 50% have the correct Q-matrix, the effect is lessened. Thus, phylogenetic error grows as more characters are misspecified.

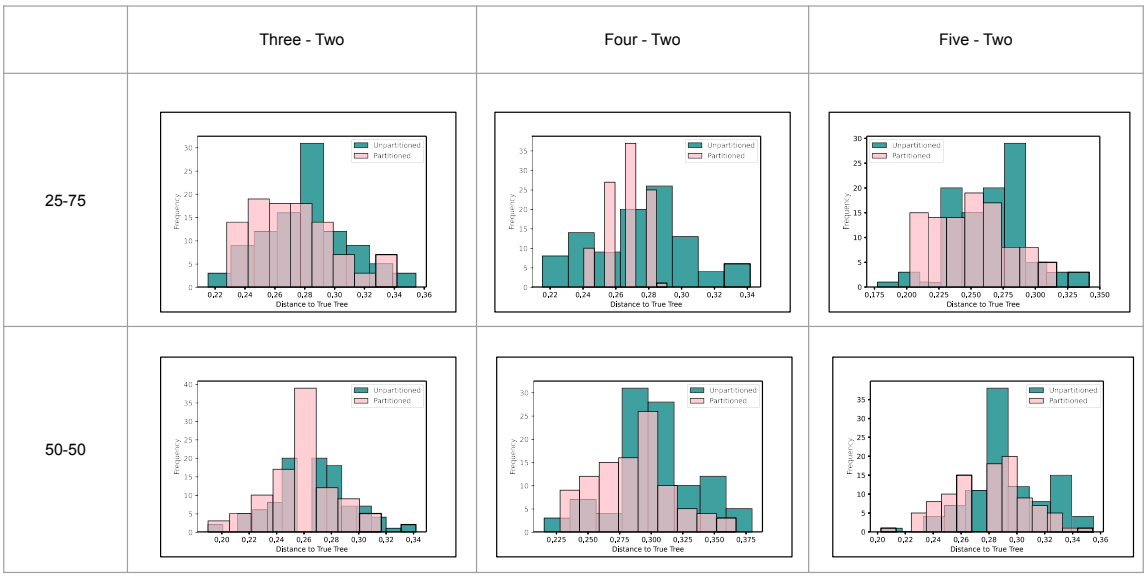


Figure 2.5: On the top panel of this figure is shown simulations in which 25% of characters come from a state space larger than binary, and 75% come from a binary matrix. Across the top are labeled which state states - Three-Two, for example corresponds to 25% of characters having three states. In this case, an not partitioning means *most* characters are being analysed under a misspecified model. On the bottom row are datasets in which 50% of characters will have a misspecified model.

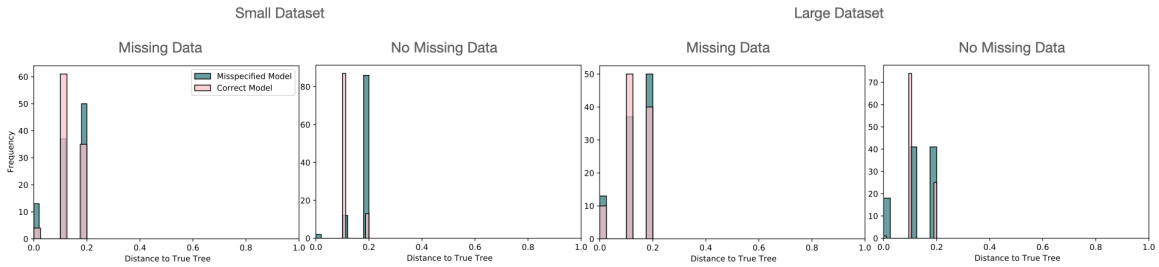


Figure 2.6: For simulations conducted under long-branch attraction conditions (LBA), topological error is more pronounced than in other simulation conditions. This suggests that when the topological problem is especially difficult, performing the analysis correctly is more important.

Long Branch Tree

For the simulations in long-branch attraction conditions, an effect of correct state partitioning can be seen on the estimated tree (Fig. 2.6). It appears that when long-branch conditions are occurring, the effect of partitioning is more important.

LBA simulations show a different pattern for tree lengths. When LBA conditions persist, it is possible for the tree to be several-fold too long (Fig. 2.7). Tree lengths of up to five times the true tree length were observed when the data are not partitioned.

2.4 Discussion

2.4.1 General issue of coding in morphological characters

Morphological characters have always been an important means of estimating phylogenetic trees. This has historically been accomplished via parsimony, and as such many fundamental questions remain about how to model morphological characters appropriately. Since the inception of including morphological characters in likelihood and Bayesian analyses (Lewis, 2001), much work has been contributed on modeling

among-character rate variation (Wagner, 2012; Harrison and Larsson, 2015), about exchangeabilities and character frequencies (Nylander et al., 2004; Wright et al., 2016; Klopfstein et al., 2019), and how to partition a data matrix (Clarke and Middleton, 2008). All these questions rely on knowledge of the phylogenetic characters being modeled.

At a more fundamental level, all of the above applications rely on having a matrix that describes the rate of changes between sites, a Q-matrix. A Q-matrix must be initialized at a given size, and that size is determined by the researcher. However, the true number of states at a character may be obscured from the researcher. As shown on Fig. 2.1, patchy sampling in the fossil record may lead to some character states not being observed, either because the organisms expressing that character state are never sampled, or the fossils themselves are incomplete and lack the character (and therefore state). Some character states may not be observable by a human observer, or observer bias or error may lead to incorrect coding of states. While nucleotide polymorphisms and sequencing error are a problem for molecular data, the Q-matrix always remains the same size: 4, the number of nucleotides. Morphologists cannot rely on this default assumption.

In my set of experiments, I examined two sources of Q-matrix error: one in which the correct number of character states cannot be known due to missing data, and the Q-matrix is therefore too small for some characters. The other treatment is declining to partition by character state space, in effect using a Q-matrix that is too large for most characters. Both of these treatments introduced phylogenetic error, though not always enough to mislead a conclusion from the analysis. In the inclusive sampling experiments, there is little effect on topology from oversizing the Q-matrix. However,

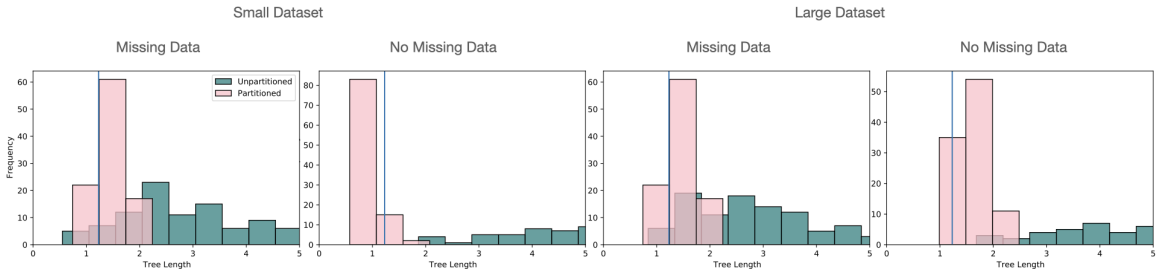


Figure 2.7: For simulations conducted under long-branch attraction conditions (LBA), tree length inflation is quite large when datasets are not partitioned correctly.

in the rejection sampling experiments, when all the larger state space characters have exactly the same state space, and are inappropriately parameterized in the exact same way, we observe a stronger signal of phylogenetic error (Fig. 2.4), which would be expected given the bias in branch lengths under theoretical model misspecification conditions in Fig. 2.3. Thus, we may conclude that the magnitude of the mispecification error matters greatly to the final conclusions. When the underlying tree has long-branch attraction, we additionally find the tree search being mislead by model misspecification (Fig. 2.6). Under LBA conditions, there is a clear tendency for unpartitioned analyses to estimate more nodes of the tree incorrectly. This implies that for difficult problems, such as LBA, it is becomes more important to parameterize models appropriately.

The effect does not stop with the topology alone. As shown on Fig. 2.4 and 2.7, failure to correctly account for character state space produces erroneous branch lengths. In particular, when the tree is prone to LBA, the tree may be up to five-fold too long when the data are treated as a single partition. In the case of the LBA simulations, the effect of model misspecification is large enough to mislead the analysis, leading to approximately double phylogenetic error in these treatment

conditions (Fig. 2.6). In these results, I am seeing the effects of the synergy of topology and branch length error.

The effect of model misspecification on branch lengths has been known since the first inclusion of morphology with likelihood and Bayesian models (Lewis, 2001). When describing the Mk model, Lewis noted that failing to account for the fact that morphologists typically do not collect invariant characters would lead to an inflation of branch lengths. Further, morphologists often do not collect characters that differ at a single taxon in the focal clade. This leads to a further reduction in the number of low evolutionary rate characters, causing more inflation of branch lengths. In the non-LBA tree simulations, we often observe branch lengths being too short (Fig. 2.4). When the maximum state in the Q-matrix is misspecified, we observed different patterns. As seen on Fig. 2.4, tree lengths of simulation replicates analysed under the correctly-specified model of evolution typically center on the true tree length. Misspecified models result in trees that are too short. When there is an incorrect maximum state (too-small Q-matrix), this means that, in the model, there are fewer possible transitions that a character can make than in reality (Fig. 2.2). With too few changes possible, fewer changes are observed. Therefore, the underestimation in this set of simulations is expected (Fig. 2.3). In unpartitioned models, in which the Q-matrix is too large for some characters, I still observe this effect. This is due to a larger proportion of characters not displaying changes into larger character state spaces, causing more characters to fall into the low rate-of-evolution categories of the Gamma-distributed among character rate variation, lowering the overall rate of changes observed across the tree. In effect, the model conflates the lack of transitions to the 4 and 5 character states in binary and trinary characters to a low rate of

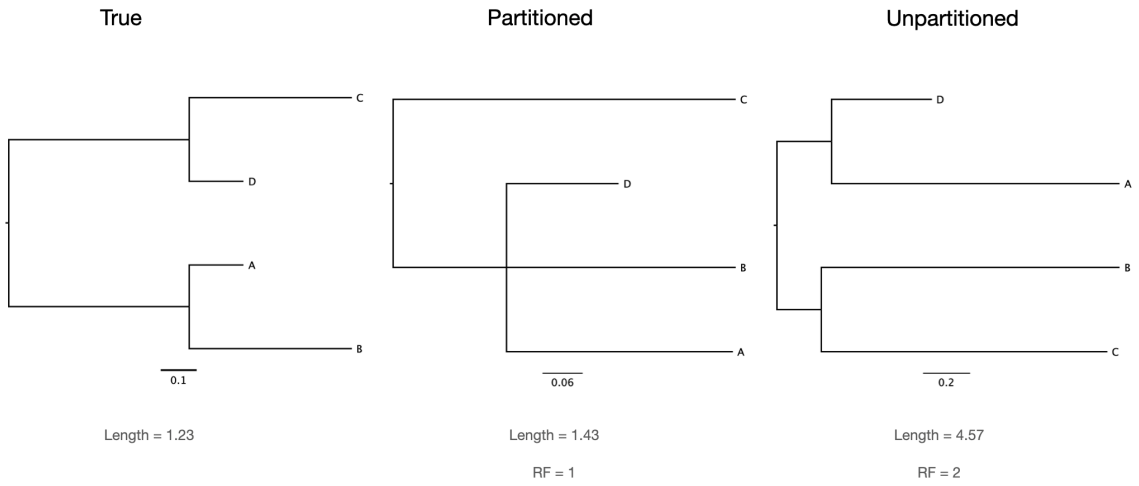


Figure 2.8: A figure showing the true LBA tree and two sampled estimated partitioned and unpartitioned trees. As can be seen, the unpartitioned tree has higher error in both topology and branch length.

evolution, and this is consistent with the relatively short branch lengths.

On the LBA trees, however, the tree topology itself tends to be more actively mislead. As seen on Fig. 2.8, the partitioned model estimates several relationships as occurring in a polytomy. In the unpartitioned model, these relationships are incorrectly resolved. The two highest-rate branches are in a clade together, with a greatly-inflated rate of evolution in the entire clade. For difficult problems, such as LBA, therefore, it appears to be very important to use an appropriate model of evolution to ensure correctness in topology. But the effect of branch lengths cannot be ignored: while likelihood-based models are less prone to LBA artifact (Felsenstein, 1978), the likelihood of a tree is still dependent on the likelihood of the topology and the branch lengths. Strong LBA can still pose problems for Bayesian analyses.

Given the overall importance of partitioning correctly, performing this task should

be relatively easy in phylogenetic software. In previous generation software, such as MrBayes (Huelsenbeck and Rannala, 2003; Ronquist and Huelsenbeck, 2003), character matrices are automatically split up by character state number. In the software RevBayes, a data matrix is not partitioned by state number by default. The reason for this is to give the researcher more control over how data are partitioned and modeled, but this comes at a higher cognitive burden to model correctly. I have implemented a method in RevBayes, *setNumStatesVector()*, to automate splitting up a phylogenetic matrix by state number to reduce the researcher burden to implement partitioned models. An example analysis with this method can be found in the online supplemental material.

In this study, I have examined how partitioning by character state space impacts phylogenetic estimation. As interest in genuine inclusion of morphological data continues to grow, spurred by methods such as the Fossilized Birth-Death process (Heath et al., 2014) and growing acknowledgment that fossils are crucial for comparative methods, we must ask fundamental questions about morphological character coding. I have demonstrated a consistent effect of incorrect character state partitioning on phylogenetic estimation. In particular, as the topological question becomes more difficult, such as when LBA conditions persist, the effect of choosing a correctly-partitioned model is more important. However, this study is not the end. Many more questions about how morphological data are modeled in a phylogenetic context and the general applicability of molecular methods for estimation remain, and I encourage researchers to think carefully and thoroughly about the choices they make when modeling morphological characters.

Bibliography

Bapst, D. W., H. A. Schreiber, and S. J. Carlson. 2018. Combined analysis of extant rhynchonellida (brachiopoda) using morphological and molecular data. Systematic Biology 67:32–48.

Barden, P. and D. A. Grimaldi. 2016. Adaptive radiation in socially advanced stem-group ants from the cretaceous. Current Biology 26:515–521.

Barido-Sottani, J., G. Aguirre-Fernández, M. J. Hopkins, T. Stadler, and R. Warnock. 2019. Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth–death process. Proceedings of the Royal Society B: Biological Sciences 286:20190685.

Brazeau, M. D. 2011. Problematic character coding methods in morphology and their effects. Biological Journal of the Linnean Society 104:489–498.

Ciampaglio, C. N., M. Kemp, and D. W. McShea. 2001. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. Paleobiology 27:695–715.

Clarke, J. A. and K. M. Middleton. 2008. Mosaicism, modules, and the evolution of birds: results from a bayesian approach to the study of morphological evolution using discrete character data. Systematic biology 57:185–201.

Fabreti, L. G. and S. Höhna. 2022. Convergence assessment for bayesian phylogenetic analysis using mcmc simulation. Methods in Ecology and Evolution 13:77–90.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic zoology 27:401–410.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

Flannery Sutherland, J. T., B. C. Moon, T. L. Stubbs, and M. J. Benton. 2019. Does exceptional preservation distort our view of disparity in the fossil record? Proceedings of the Royal Society B 286:20190091.

Goldberg, E. E. and B. Igić. 2008. On phylogenetic tests of irreversible evolution. Evolution 62:2727–2741.

Gould, S. J. 1970. Dollo on dollo's law: irreversibility and the status of evolutionary laws. Journal of the History of Biology 3:189–212.

Harrison, L. B. and H. C. Larsson. 2015. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. Systematic Biology 64:307–324.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:E2957–E2966.

Höhna, S., T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P. Huelsenbeck. 2014. Probabilistic graphical model representation in phylogenetics. Systematic Biology 63:753–771.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology 65:726–736.

Huelsenbeck, J. P. and B. Rannala. 2003. Detecting correlation between characters in a comparative analysis with uncertain phylogeny. Evolution 57:1237–1247.

Jukes, T. and C. Cantor. 1969. Evolution of protein molecules. Mammalian Protein Metabolism 3:21–132.

Keating, C. F. 1985. Human dominance signals: The primate in us. Pages 89–108 *in* Power, dominance, and nonverbal behavior. Springer.

Klopfstein, S., R. Ryer, M. Coiro, and T. Spasojevic. 2019. Mismatch of the morphology model is mostly unproblematic in total-evidence dating: insights from an extensive simulation study. BioRxiv Page 679084.

Lartillot, N. and H. Philippe. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution 21:1095–1109.

Lewis, P. O. 2001. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. Systematic Biology 50:913–925.

McKinney, W. et al. 2011. pandas: a foundational python library for data analysis and statistics. Python for high performance and scientific computing 14:1–9.

Nylander, J. A., F. Ronquist, J. P. Huelsenbeck, and J. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47–67.

Rambaut, A. and A. J. Drummond. 2011. Tracer v1.5. http://tree.bio.ed.ac.uk/software/tracer/.

Robinson, D. F. and L. R. Foulds. 1979. Comparison of weighted labelled trees. Pages 119–126 *in* Combinatorial mathematics VI. Springer.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical biosciences 53:131–147.

Ronquist, F. and J. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Stavenga, D. G. and K. Arikawa. 2006. Evolution of color and vision of butterflies. Arthropod structure & development 35:307–318.

Stevens, P. F. 1980. Evolutionary polarity of character states. Annual Review of Ecology and Systematics 11:333–358.

Sukumaran, J. and M. T. Holder. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Some Mathematical Questions in Biology: DNA Sequence Analysis 17:57–86.

Wagner, P. J. 2000. The quality of the fossil record and the accuracy of phylogenetic inferences about sampling and diversity. Systematic Biology 49:65–86.

Wagner, P. J. 2012. Modelling rate distributions using character compatibility: implications for morphological evolution among fossil invertebrates. Biology Letters 8:143–146.

Waskom, M. L. 2021. Seaborn: statistical data visualization. Journal of Open Source Software 6:3021.

Wright, A. M. and D. M. Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS One 9:e109210.

Wright, A. M., G. T. Lloyd, and D. M. Hillis. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. Systematic Biology 65:602–611.

Wright, A. M., K. M. Lyons, M. C. Brandley, and D. M. Hillis. 2015. Which came first: the lizard or the egg? robustness in phylogenetic reconstruction of ancestral states. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 324:504–516.

Chapter 3

Site-Heterogeneous Character Change Models for Morphology

Abstract

Modeling morphological character evolution for the estimation phylogenetic trees is challenging, in part due to challenges with assuming a single, common mechanism across a dataset. Unlike molecular data, in which a nucleotide or amino acid may be assumed to have the same properties everywhere it occurs in an alignment, morphological character state may confer different meanings in different columns in a character matrix. This complexity has inhibited the implementation of additional models of character evolution. The Mk model is the most commonly used model for incorporating morphology in maximum likelihood and Bayesian phylogenetic estimation. This model is a generalization of the Jukes-Cantor model which assumes that there is an equal rate of gains and losses in morphological traits. In this study, I

explore the use of models that allow asymmetric transition rates. I use an expanded combined set of two ant (Formicidae) data matrices, composed of one matrix of extinct and extant ants and one matrix of extant ants to estimate a phylogeny for the group. There is extensive sampling in the stem ants, which leads to classes of gains and loss within the data, as ant apomorphies are gained, but characters more similar to the outgroup are lost. This violates the assumption of equal transition rates among character states. In particular, I examine the use of two models to allow the equilibrium character frequency to vary. One model, assumes a Beta distribution as a prior on character frequencies. This model is similar to a model implemented in MrBayes and can be used to allow asymmetrical transition rates in binary characters. I also describe a Dirichlet model to allow asymmetrical transitions in multistate characters. These two models can be used to explicitly model different character frequencies within one dataset, as opposed to assuming a single common mechanism with respect to among character frequency variation. I have also compared the models using stepping-stone analyses and model averaging using reversible jump Markov chain Monte Carlo. Additionally, I test for model adequacy using posterior predictive simulations to determine the adequacy of different models for morphological evolution.

3.1 Introduction

3.1.1 Bayesian Modeling of Morphology

Phylogenetic trees provide the backbone for a wide range of studies in taxonomy and macroevolution (Hennig and Davis, 1966; de Queiroz and Gauthier, 1992; Felsenstein,

1985; Harvey and Pagel, 1991; O'Meara, 2012; Uyeda et al., 2018; Dunn et al., 2018). Including extinct species in a phylogenetic analysis is important for detecting patterns of diversification across many taxonomic scales (Benson and Choiniere, 2013; Ezard et al., 2013; Near et al., 2014; Price et al., 2014; Betancur-R et al., 2015; Slater et al., 2012; Heath et al., 2014). Beyond informing macroevolutionary dynamics, fossils can also be used to break up long branches in clades that are sparse in the present (Wiens, 2005; Cobbett et al., 2007; Magallón, 2010). While the reconstruction of phylogenetic relationships among extant species has been greatly facilitated by the accumulation of large quantities molecular sequence data, the placement of extinct species is more challenging. The oldest DNA samples recovered to date are under one million years old (Orlando et al., 2013), and recovering DNA much older is not likely (Kirkpatrick et al., 2016). Therefore, the phylogenetic placement of most extinct species can only be accurately estimated using fossil evidence, with the main source of phylogenetically informative characters being fossil morphology. These morphological traits are most commonly coded as a matrix of discrete characters, which are then used to infer a phylogeny by various reconstruction algorithms.

While maximum parsimony methods have dominated classical approaches for reconstructing fossil species relationships from discrete morphological characters, probabilistic approaches have become increasingly popular because they allow for the use of mechanistic models describing the processes underlying the evolution of extinct and extant morphological and molecular characters in a unified statistical framework (Nylander et al., 2004), enabling researchers to rigorously test evolutionary hypotheses taking as much data into account as possible (Sullivan and Joyce, 2005; Xie et al., 2010; Lewis et al., 2014; Brown, 2014a). In particular, Bayesian methods have enabled

advances in model design and flexibility, allowing for the use of complex hierarchical models unifying a wide range of data types in the analysis of a single underlying tree model or diversification process (Höhna et al., 2016).

The most commonly used probabilistic model of discrete morphological character evolution is the Mk model Lewis (2001). The Mk model assumes an equal rate of transition between any two character states, and can therefore be seen as a generalization of the Jukes-Cantor model for nucleotide evolution (Jukes and Cantor, 1969) extended to an arbitrary number of k character states. Due to the simplicity of its assumptions, some researchers have raised concerns about the realism of the Mk model (Goloboff et al., 2018b,a), and much of the work on methods for estimating phylogenies from discrete traits has focused on how well this method performs, particularly in comparison to parsimony (Wright and Hillis, 2014; Puttick et al., 2017; Brown et al., 2017; Schrago et al., 2018).

In particular, a common criticism of the Mk model in its most basic formulation is that it fails to account for heterogeneity in the meaning of character states across characters. Unlike a matrix of nucleotide characters, in which a state carries common meaning across sites, a state at one morphological character may represent a derived state requiring many underlying genetic changes, while at another character, it could represent a state requiring only one underlying genetic change (Wagner, 1989). In this case, the rate of change is expected to vary across characters. Characters can also differ in meaning with respect to whether they are coded as ancestral or derived (Stevens, 1991). For example, if a matrix was coded with respect to a specific outgroup, plesiomorphic characters in the outgroup taxa will most commonly be coded with a '0' character state, while ingroup taxa will typically be coded as '1' or higher.

However, this may not be the case for composite matrices or matrices coded without specific reference to an outgroup or other polarizing information. This lack of common meaning among character states has been used to argue that a single common mechanism model, such as the Mk model, cannot be applied to discrete morphological data (Goloboff et al., 2018a). This is in contrast to molecular data, where base pairs or amino acid residues are generally accepted to have similar properties across an alignment, and exceptions are often defined with respect to known features, such as stem and loop domains Pagel and Meade (2004); Letsch et al. (2010), gene, or codon position (Brown and Lemmon, 2007).

In order to alleviate some of these problems, several extension of the Mk have been developed. For example, the Mk model has been extended to accommodate patterns of character evolution commonly associated in parsimony analysis (Wiley and Lieberman, 2011; Swofford, 1985) such as ordered characters (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003), or Dollo-like irreversible losses (Nicholls and Gray, 2006). The Mk model has also been extended to accommodate Gamma-distributed among-site rate variation (ASRV), as is commonly employed in the analysis of molecular data (Yang, 1994).

Some extensions to the Mk model have also been developed to relax the assumption of an equal rate of transition between character states. In particular, two models have been implemented in MrBayes (Huelsenbeck and Ronquist, 2001) that account for heterogeneity in state frequencies across characters, but a formal mathematical description of these models was never published. As currently implemented in MrBayes, the state frequencies across characters are assumed to be distributed according to a symmetric Dirichlet distribution. For binary characters, this corresponds to a

Beta distribution, which is then discretized into a user-specified number of categories. The character data likelihood is then computed by integrating over these discrete categories at each site.

There are clear ways to improve this concept by borrowing from the molecular literature. For example, the CAT model, implemented in PhyloBayes (Lartillot and Philippe, 2004; Lartillot et al., 2009), assumes that site-specific state frequencies are distributed into an arbitrary number of categories according to a Dirichlet process, a non-parametric model of flexible complexity. In this way, the CAT model helps to automatically adapt to the heterogeneity present in large phylogenomic datasets of thousands of loci, but may not be appropriate for the analysis of small morphological datasets.

Here, I present a number of additional extensions to the Mk model, building on the models implemented in MrBayes, and a finite mixture model similar to the CAT model Lartillot et al. (2009). Specifically, for binary character data, I have extended the symmetric Dirichlet model of MrBayes to accommodate an asymmetric Beta prior distribution. For multistate characters, I have developed a site-heterogeneous model for discrete morphology (SDHM) that assumes character state frequencies are distributed according to a finite mixture model with an arbitrary number of discrete categories, each drawn from a Dirichlet distribution that may be symmetric or asymmetric. All models were implemented in the Bayesian statistical phylogenetics software package RevBayes. These extensions to the Mk model help to better accommodate the heterogeneity present in discrete morphological data, and help broaden the usefulness of fossil morphological data in the probabilistic analysis of evolutionary processes and species diversification.

3.1.2 Morphological Phylogeny of the Formicidae

I tested the efficacy of these modeling techniques using discrete morphological character data from ants (Family: Formicidae). Ants are ecologically crucial organisms as both interacting partners for a variety of plants and animals, and as shapers of the ecosystem via soil cycling and nest building. As such, efforts towards reconstructing the Ant Tree of Life (Ward et al., 2005) are invaluable for understanding the origin and evolution of biodiversity in a wide variety of ecological systems. Ants also have a rich fossil record, with most subfamilies being represented in systematic work going back into the 1800s (Schweigger, 1819), as well in a variety of other ecological and evolutionary studies (Morayma and Kraemer, 2007; Perrichot and Girard, 2009; Pie and Tschá, 2009; Moreau and Bell, 2013). Beginning with the discovery of stem ant *Sphecomyrma freyi* in 1967 (Wilson et al., 1967), fossilized representatives from outside the crown of ants have also helped to inform our understanding of ant paleodiversity, evolution and ecology.

Phylogenetic resolution in ants has historically been well-supported by morphological data at the subfamily level, with considerable disagreement at any finer scale (Urbani et al., 1992; Grimaldi et al., 1997; Ward et al., 2005). For example, IX current subfamilies had been previously considered to be a single subfamily, the Ponerinae, but recent evidence from analysis of morphological data (Bolton, 2003) and molecular sequences (Saux et al., 2004; Brady et al., 2006; Moreau et al., 2006; Ouellette et al., 2006) indicated this family should be broken up into six distinct subfamilies, one of which is now called Ponerinae, which together with the rest is often referred to as the ponerimorph group. Recent morphological work has specifically aimed at expanding the sampling of ants from in the ponerimorph grouping, which had been previously

undersampled Keller (2011).

Sampling of fossil ants was also expanded to include more specimens from the Cretaceous. In recent years, sampling of stem ants from the Cretaceous has been greatly expanded. The genera *Gerontofromica* (Nel et al., 2004) (Albian amber in France; 100 mya) and *Camelomecia* (Burmese amber; 92 mya) have been recently described and these taxa have been placed in a polytomy on the stem of ants with *Sphecomyrma freyi* by morphological evidence (Barden and Grimaldi, 2016). Even with expanded taxon sampling in the Ponerinae, there is still conflict between the molecular and morphological estimates of phylogenetic relationships. Absent inputs of additional information, such as clade constraints, morphological trees are still substantially unresolved.

In this study, I combine the extant matrices of Keller, and the extinct-extant matrix of Barden and Grimaldi to expand the taxon and character sampling. Because there are stem ant lineages represented in Barden and Grimaldi's dataset, there are taxa that have characteristics that are lost after the divergence of the stem lineages. This extremely one-sided loss structure violates the Mk model assumption of equal transition rates. We would expect for these characters to be better described by a model that can accommodate asymmetrical transition rates. Some apomorphies of the ant group are also gained after the divergence of the stem lineages and not lost frequently, which also violates the assumption of equal change probabilities. Because of these model violations, this dataset is an excellent test case for models that relax assumptions of the Mk model.

I use Bayes Factor model selection to assess the fit of these relaxed models to the data. Using this dataset, I strongly support that the use of models that relax key

assumptions of the Mk model can greatly improve the fit of the model to the data. I also perform simulations to demonstrate that the true number of transition rate asymmetry categories is detectable from the data, and that using an appropriately-specified model of morphological evolution is less likely to lead to incorrect phylogenetic inference than a correctly specified model (i.e., lead to polyomties rather than incorrectly-resolved nodes).

However, the morphological tree of ants remains poorly-resolved, consistent with prior morphological work in the group. I do, however, support the monophyly of what has been considered the crown group of ants, and support a monophyletic Haidomyrmicine ants, as in Barden and Grimaldi (2016). Despite the lack of resolution, I think that this model is more reflective of the evolutionary processes operating within this dataset, and provide a discussion for the continued ambiguity in the group.

3.2 Methods

3.2.1 Modeling site-heterogeneous state frequencies

I model morphological character evolution using the Mk model (Lewis, 2001), and additionally account for site-specific substitution rates among character states by allowing the equilibrium state frequencies π to vary among sites, such that each site i has its own state frequency vector π_i.

Discretized Beta model

I model site-specific binary state frequencies by assuming the state frequencies π_i are drawn from a 2-dimensional Dirichlet (Beta) prior distribution with hyperparameters

α, β. Then, I compute the likelihood of the data D by marginalizing over the value for π_i at each site.

$$f(D \mid \alpha, \beta) = \prod_{i=1}^{n} \int_0^1 f(D_i \mid \pi_i) f(\pi_i \mid \alpha, \beta) d\pi_i \tag{3.1}$$

To simplify the computation, I approximate this integrated likelihood by assuming π_i is drawn from a mixture distribution, where the mixture categories are defined deterministically by discretizing a Beta distribution into k bins whose boundaries are defined using $k-1$ quantiles. The value of mixture category j is indicated by ϕ_j and is computed as the interquantile median of the j-th bin. In other words, $\phi_j = I^{-1}_{(2j-1)/2k}(\alpha, \beta)$, where $I_x(\alpha, \beta)$ is the regularized incomplete Beta function. The mixing proportion is $1/k$ for each category. Then the likelihood is computed by summing over the k discrete mixture categories for π_i at each site.

$$f(D \mid \alpha, \beta) = \prod_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{k} f(D_i \mid \pi_i = \phi_j) \tag{3.2}$$

Note that in the limit as the number of mixture categories approaches infinity, equation 3.2 is identical to equation 3.1. This model is shown as a graphical model in figure 3.1A.

Site-heterogeneous multistate model

When analyzing multistate character data with S character states, I assume the state frequencies π_i at each site i are drawn from a Dirichlet distribution with concentration parameters $\alpha_1, \ldots, \alpha_S$. Again, I use a mixture distribution to simplify the computation, but instead of defining the mixture categories using a discretized Dirichlet

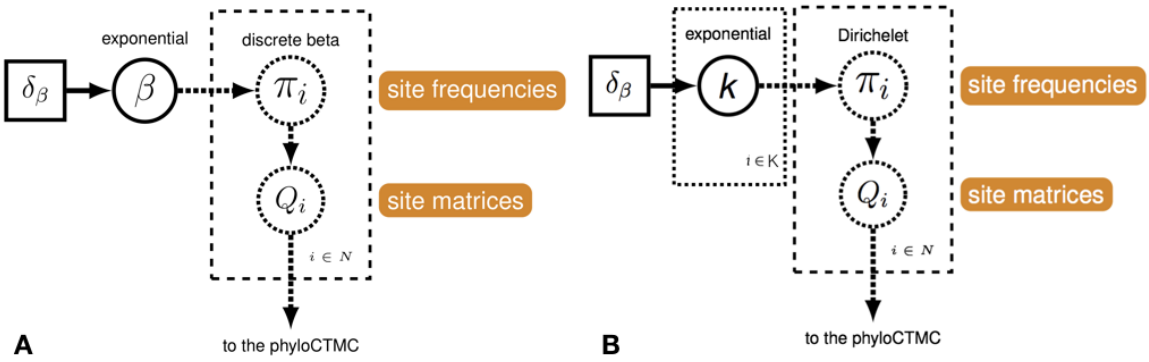


Figure 3.1: A. A graphical model demonstrating the discrete Beta model. B. A graphical model displaying the SHDM. Model realization following **?**

distribution, the value δ_j of mixture category j is assumed to be drawn from the same underlying Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_S$. The mixing proportions are θ_j, which are themselves drawn from a uniform Dirichlet prior distribution.

$$f(D \mid \delta_1, \ldots, \delta_k, \theta_1, \ldots, \theta_k) = \prod_{i=1}^{n} \sum_{j=1}^{k} f(D_i \mid \pi_i = \delta_j)\theta_j$$

A graphical model of this model can be seen in figure 3.1B.

3.2.2 Data Matrices

Empirical Matrices

Several large and well-documented ant matrices were used in this study. The first was that of Keller (2011). This matrix is of extant ant groups. Keller's matrix was collected with special attention to the poneromorph subgroups (Amblyoponinae, Ectatomminae,Heteroponerinae, Paraponerinae, Ponerinae, and Proceratiinae) and has 139 characters and 105 taxa. The matrix is fairly complete for a morphological

dataset, with nearly all cells being filled. Of the total character set, 100 characters were binary and 39 characters were multistate. Because of the scope of Keller's study, the taxon sampling is biased towards the poneromorphs, and away from the other ant subfamilies, including large subfamilies such as the Dolichoderinae and the Formicinae.

I also used data from Barden and Grimaldi (2016). This matrix contained both extant and extinct ants, and was collected in order to place stem ants from the Cretaceous period, and was compiled for compatibility with Grimaldi et al. (1997). One crown ant amber fossil, *Kyromyrma neffi*, was included in this matrix, but the key feature of this matrix is the sampling of the stem group, including multiple samples from the *Gerontoformica* genus. This matrix also expands sampling in the non-ponerine groups that are underrepresented in the Keller matrix. The Barden-Grimaldi matrix contained 42 characters and 41 taxa. Of these characters, 26 were binary.

In order to achieve maximum coverage, I merged these character matrices. The taxonomic overlap between the two was 11 taxa, and all subfamilies of the Formicidae except the extinct Formiciinae were represented in the matrix. Eleven characters were represented in both matrices without any recoding, and both matrices had many characters that were inapplicable to the other matrices. Characters that are inapplicable (i.e., scoring the morphology of a body part that has been lost in part of the tree) were retained in the matrix due to their ability to resolve bi-partitions in the the authors' respective groups of interest.

Six characters were recoded to make the character states uniform between the two matrices. Most of these changes simply involved changing the terminology used

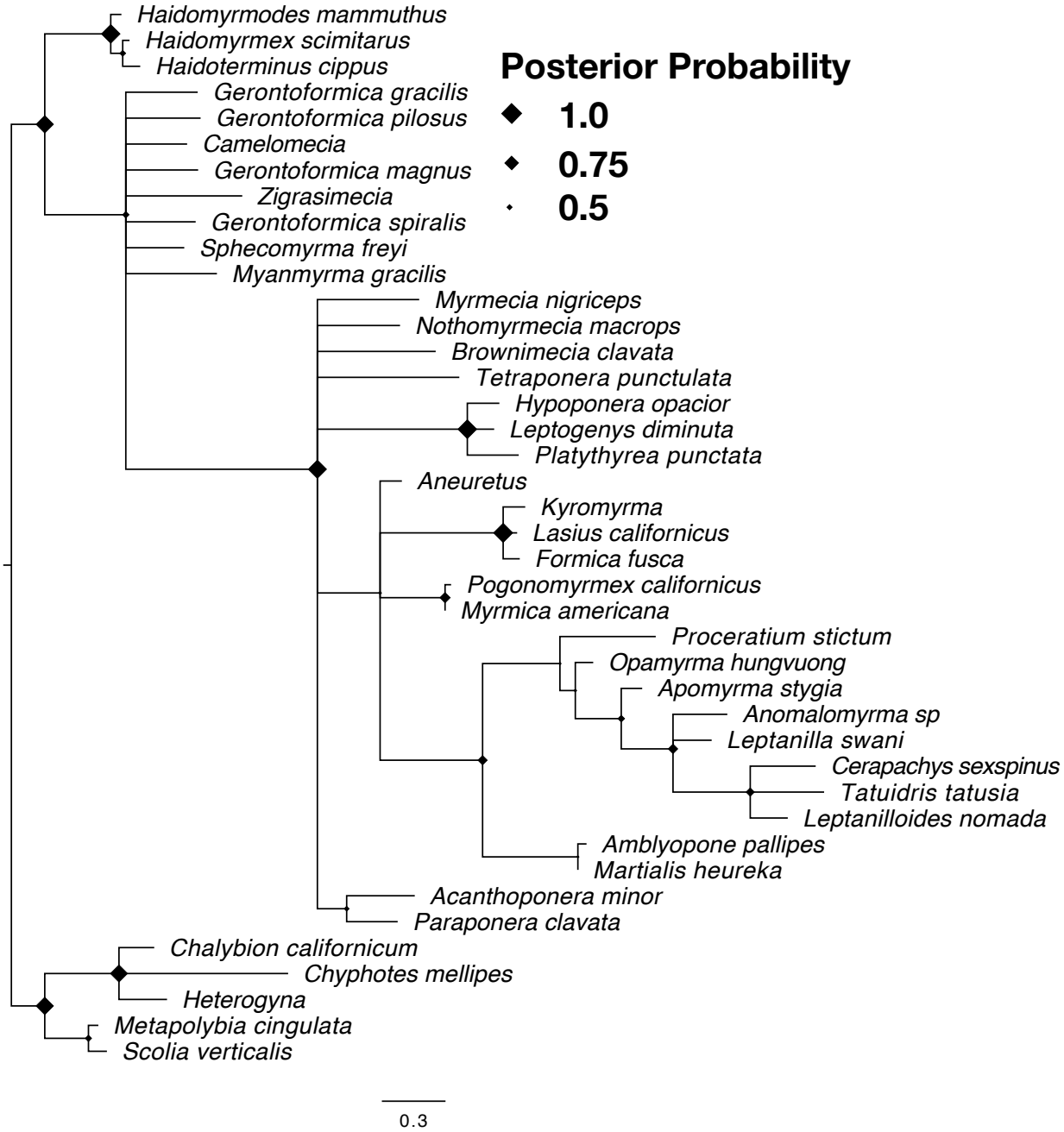


Figure 3.2: Strict consensus tree estimated from the dataset of Barden and Grimaldi (2015) under the SHDM.

in reference to the character states. For example, Barden and Grimaldi's character 15 is the same as Keller's character 47, but the two matrices had inverse character codings relative to one another. The remainder of the changes were changing binary characters to a multistate in cases where one author had more states than the other. I will refer to this dataset as the combined dataset.

I modeled all three datasets according to both models. First, to test the Beta model on binary-only datasets, we used Barden and Grimaldi's dataset with the multistate characters removed. This dataset was chosen because the dataset size and completeness is very typical to a morphological dataset (Wright et al., 2016; Barido-Sottani et al., 2019), particularly one involving fossils. Because of the presence of stem ants, we also expect to see strong violations of the assumptions of the Mk model in this dataset. I also used the combined matrix, stripped of multistate characters to test the model to look at the effect of estimating a larger tree using more characters. In both cases, I performed Bayes Factor model fitting to determine how many discrete Beta categories best model the data.

To test the full SHDM model, I used both the Barden-Grimaldi dataset and the combined dataset. In testing the SHDM model, I modeled the binary and multistate characters together. This allowed me to both examine the impact of allowing character frequencies to be drawn from a Dirichlet, and to examine the effect of adding more data (both more taxa and more characters) to the phylogenetic question.

All estimations were performed in the software RevBayes.

3.2.3 Model Testing

Stepping-Stone Analyses

The appropriate number of categories k for each of the two empirical datasets was selected by comparing estimates of the marginal likelihood under different values for k. Marginal likelihoods were estimated using the steppingstone sampling method (Xie et al., 2010) implemented in RevBayes.

Simulated Matrices To test the performance of the SHDM model on idealized datasets, I used RevBayes to simulate four sets of 100 replicated of morphological data. The first two sets of simulations were based on Barden and Grimaldi's dataset size and tree. Under this set of conditions, datasets of 41 taxa and 42 characters were simulated in RevBayes. In one set, only binary characters were evolved under the Beta model ($k = 4$). In the other hand, both binary and multistate characters were evolved under the Dirichlet prior ($k = 4$).

The other set of two simulations were based on the full datasets. One set of 139 binary characters was generated for the full 105 taxa under the binary model; a second set including multistate characters was generated under the SHDM model, with the distribution discretized into four categories ($k = 4$). The tree used to generate the data was the tree estimated from the combined empirical dataset using the SHDM model.

Empirical Data Trees were estimated from each empirical dataset using multiple models of evolution. The Mk model was used for each dataset, with an appropriately-sized substitution rate matrix. For the binary-only datasets, the binary model was

used. The datasets containing both binary and multistate characters were analyzed using the SHDM model, with the dimensionality of the Dirichlet process being equal to the largest number of character states. Each dataset was run under several different numbers of discrete state frequency variation categories, from two to six categories. All datasets were corrected for not observing invariant characters (Lewis, 2001).

For the binary datasets, Bayes Factor model fitting was used to assess what the optimal number of discrete categories were. I used stepping-stone model fitting to calculate the marginal likelihood. I first ran an MCMC analysis to determine how many generations are required to attain convergence (about one million). Each of the 74 stepping stones was then run for one million generations. Due to computational limits, for the multistate datasets, we used the mean likelihood of the MCMC distribution to compare across different discretizations of the model.

Simulated Data Because the true tree cannot be known from empirical data, we used simulated data to observe the effect this model has on accuracy of estimation, and to see if it is possible to detect the true number of discrete categories of state frequency variation. To this end, I estimated trees for each replicate under the true model ($k = 4$). We also estimated trees for each replicate under 5 misspecified models: the Mk model, 2 and 3 category models (under-parameterized models, $k = 2$ and 3), and 5 and 6 category models (over-parameterized models, $k = 5$ and 6). I performed the estimations in RevBayes.

Model Averaging using Reversible Jump Markov-Chain Monte Carlo

Model selection approaches such as steppingstone sampling can pose challenges as different model components can be informed by different types of data. To tackle with issues as such, reversible jump Markov chain Monte Carlo (rjMCMC) has been proposed as a more reliable approach (May and Rothfels, 2023). Reversible jump MCMC (Green and Hastie, 2009) allows researchers to vary the number of parameters present in a Bayesian model. In addition to the traditional MCMC analysis, in rjMCMC the number of parameters itself can be varied. This way the best model can be selected by the data itself. In this case, how often model is sampled by MCMC is itself informative.

In the case of comparing between Mk model and SHDM model using rjMCMC, I first chose Mk model and SHDM model with 3 rate categories. Additionally, I varied the parameters of the SHDM model such that rate categories of 2 to 8 were included. This way, the rjMCMC algorithm could sample whichever model is informed by the data.

Posterior Predictive Simulation

To test for model adequacy, I performed posterior predictive simulations (PPS) (Brown, 2014b) under different models. I tried three different categories of PPS to evaluate the adequacy of the models — Data PPS, Inference PPS, and Mixed PPS. The dataset I used for these analyses are the combined dataset, as it accounts for both binary and multistate characters. PPS determines the absolute fit of a model to the data. A schematic of the PPS workflow can be seen in figure 3.3. I have a brief explanation of how PPS works but for a more in-depth explanation see Höhna et al. (2018). The

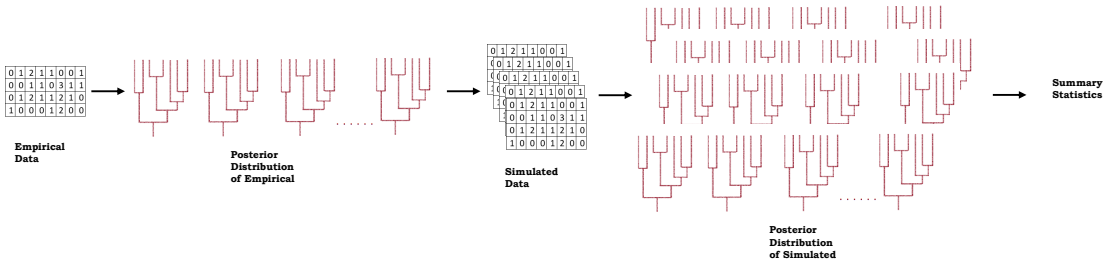


Figure 3.3: A generalized workflow for Posterior Predictive Simulation. The empirical data is used for estimating the posterior distribution which is then used in simulating some matrices. The simulated matrices can be used for comparing with empirical matrix and can also be used to estimate the posterior which can then be compared with the posterior distribution of the empirical dataset.

first step is to analyze the empirical data under certain model under test. It generally involves performing an MCMC inference and sampling the posterior. After we obtain the posterior distribution of the trees and the parameters, it can be used to simulate datasets. The newly simulated dataset is then used to infer posterior distribution which is then used to calculate some essential summary statistics and compared with the empirical posterior distribution. Some summary statistics can also be used to compare just the simulated datasets with the empirical datasets. The idea over this method is that if the simulated datasets and the posterior distribution is similar to the empirical dataset, then the model is adequate. This would indicate if the model is fit for further analyses. In this portion, I investigate the fit of SHDM model with 3 category, 6 category and 8 category with the fit of Mk model.

Data PPS In data posterior predictive simulation, I compare the empirical data matrix with the data simulated from the posterior distribution of the empirical dataset. In principle, if the model is adequate the summary statistics used should

show the alignment of the simulated summary statistics with the empirical summary statistics.

This is one of first times, posterior predictive simulation has been applied to morphological data. Phylodynamic models, models for nucleotide substitution, and even molecular clock models have been tested for model adequacy in various studies (Duchêne et al., 2015; Duchene et al., 2019; Slater and Pennell, 2014). I have tried to use existing summary statistics that are applied for molecular data as explained in Höhna et al. (2018). Some of the summary statistics that I have applied in this analysis is listed below:

- Number of Invariant Sites - This summary statistic captures the characteristic of an alignment where sites have different rates of evolution.
- Maximum Pairwise Differences - This test statistic intends to capture the rate-variation among site and/or among branches.
- Maximum Variable Block Length - This test statistic shows the maximum number of blocks that have varying sites.
- Minimum Pairwise Differences - This statistic finds the pair of sequences that has the smallest pairwise distance.
- Gower's Coefficient - Gower's coefficient (GC) (Gower, 1971) is commonly used to calculate the disparity metric generally used in invertebrate studies (Hopkins and Smith, 2015). This metric is similar to GED but it deals with missing

characters differently. GC can be written as

$$S_{ij} = \frac{\sum_{k=1}^{v} S_{ijk}^2 W_{ijk}}{\sum_{k=1}^{v} \delta_{ijk}^2 W_{ijk}} \qquad (Lloyd,\ 2016) \qquad (3.3)$$

where S_{ij} is the total distance between taxa i and j, v is the total number of characters in the data matrix, W_{ijk} is the weight of the k^{th} character and δ_{ijk} is coded as 1 if both taxa i and j can be coded for k (character states observed for both taxa) and coded as 0 if the character state cannot be coded.

- Generalized Euclidean Distances - Generalized Euclidean Distances (GED) (Wills, 1998) is a popular disparity metric used with vertebrate research. This metric is similar to the raw euclidean distances with minor modifications accounting for missing data. GED can be written as

$$S_{ij} = \sqrt{\sum_{k=1}^{v} S_{ijk}^2 W_{ijk}} \qquad (Lloyd,\ 2016) \qquad (3.4)$$

The metrics GED and GC were calculated using the R package *Claddis* (Lloyd, 2016) and the effect sizes were calculated by taking the mean disparity calculated for each dataset.

There were some predefined summary statistics in this pipeline like Minimum GC content and variance of GC content which do not apply for my dataset as it contains morphological character states only. The full set of summary statistics used in this pipeline can be found in the appendix A. I simulated 5001 datasets from the posterior of the empirical data under the Mk model, SHDM with 3 categories, SHDM with 6 categories and SHDM with 8 categories to test the absolute model fit.

Inference PPS In this category, I have compared the posterior distribution of the empirical dataset with the posterior distribution of the simulated dataset. I have compared 100 randomly selected simulated posterior distribution as this method is computationally intensive. The summary statistics that were used are as follows:

- Mean Robinson Foulds - The Robinson-Foulds metric (Robinson and Foulds, 1981) shows the topological distance scaled by the branch lengths on the trees. RF is calculated from each tree in the posterior distribution and a mean is taken to obtain this metric.

- Quantiles of Robinson Foulds Distance - This measure is similar to the mean RF. Different quantile positions can be used to probe various parts of the distribution (Brown, 2014a)

- Mean Tree Length - This test statistic is the sum of all branch lengths which is averaged across the posterior distribution of the trees. This will indicate the number of evolutionary changes in the inferred tree.

- Variance in Tree Length - This test statistic is designed to capture the uncertainty in the posterior distribution of the branch lengths.

- Entropy - The entropy of the tree captures the information gain between the marginal prior and the posterior distribution of the tree topologies.

All the inference PPS summary statistics were calculated in RevBayes.

Mixed-PPS This method is referred to as mixed PPS as it utilized both the datasets and the posterior distribution to calculate the summary statistics. I used the

datasets that I have used for **Inference PPS**. The first step is to obtain maximum clade credibility (MCC) trees (Helfrich et al., 2018) from both the empirical and simulated posterior and then use the data to calculate the summary statistic along with the MCC tree. The summary statistics I used for the Mixed PPS are listed below:

- Consistency Index - Consistency Index (CI) (Kluge and Farris, 1969) is a measure of homoplasy within the data set and can be calculated as

$$CI = \frac{m}{s} \tag{3.5}$$

  where m is the minimum possible number of changes in a tree and s is the reconstructed number. This is a common metric in paleontology and has be applied in model adequacy studies in molecular data (Duchêne et al., 2018)

- Retention Index - Retention Index (RI) (Farris, 1989) is similar to CI and it calculates the potential synapomorphy observed along the tree. This can be calculated as

$$RI = \frac{g - s}{g - m} \tag{3.6}$$

  where g is the maximum number of possible steps in a given tree.

Both of these test statistics were calculated in the R package *phangorn* (Schliep, 2011).

All the analyses were performed in RevBayes. The posterior prediction plots for the summary statistics were plotted using *RevGadgets* R package (Tribble et al., 2022). I also calculated the effect sizes for all the summary statics. Effect sizes can be used to compare the summary statistics' ability to distinguish the fit between the

competing models. The effect sizes were calculated using

$$ES = \frac{|empTS - simTS|}{SDsimTS} \tag{3.7}$$

where empTS is the empirical test statistic, simTS is the test statistic of a simulated replicate and SDsimTS is the standard deviation across all the simulated replicates.

3.3 Results

3.3.1 Empirical Phylogenetic Analyses

The marginal likelihood comparisons for the combined dataset with the multistate characters removed can be seen in Fig. 2A, and including the multistate characters can be seen in 2B. The marginal likelihood comparisons for the Barden and Grimaldi dataset can be seen in Fig. 2C and 2D.

In the Barden datasets, the data support 4 character state frequency variation categories for both the binary and multistate data ($k = 4$). In the combined dataset, 6 categories are supported for binary data ($k = 6$) and 7 are supported for multistate ($k = 7$).

When a phylogeny is estimated from the Barden dataset under the supported model, many nodes are still unresolved, or resolved, but poorly-supported. The tree (Fig. 3) supports the monophyly of crown ants. Haidomyrmicine ants are monophyletic, and less closely related to the crown ants than the Sphecomyrmine ants are, a relationship also supported by Barden and Grimaldi. Many relationships in

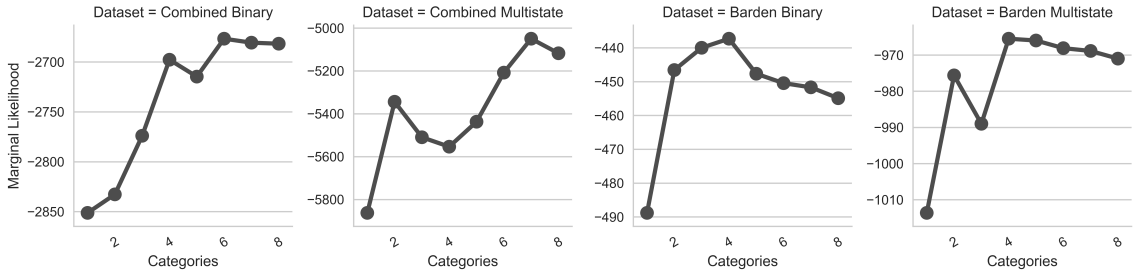


Figure 3.4: Marginal likelihood calculations for the eight models tested for teach of the four empirical datasets. The number of categories into which the Beta (binary data) or Dirichlet (multistate data) is discretized is indicated on the x-axis. Note that due to differing dataset sizes, each plot has a unique y-axis.

the crown ants supported by Barden and Grimaldi, such as monophyletic Ponerine, Myrmicinae, and Formicinae subfamilies are also supported in my analysis. I also find support for an assemblage of Doryline, Leptilline and other ant families found by previous workers (Barden and Grimaldi, 2016; Wilson et al., 1967), though the exact relationships are different, and poorly-supported.

Much as in Barden and Grimaldi's unconstrained analysis, many subfamilies are monophyletic, as expected from molecular analysis, but their relationships to one another remain vague. Several relationships are notably different than Barden and Grimaldi's analysis. Heteroponera and Paraponerinae form a clade, as do Amblyoponinae and Martialinae. Both of these clades are poorly-supported.

3.3.2 Simulated Phylogenetic Analyses

Binary Data

In the large datasets, underparameterized models have a deleterious impact on the accuracy of phylogenetic analyses. However, overparameterized models did not show

a similar decrease in accuracy. In the smaller datasets, accuracy does not seem to be tied to the parameterization of the model. Overall, phylogenetic error is lower in larger datasets, even with appropriately parameterized models.

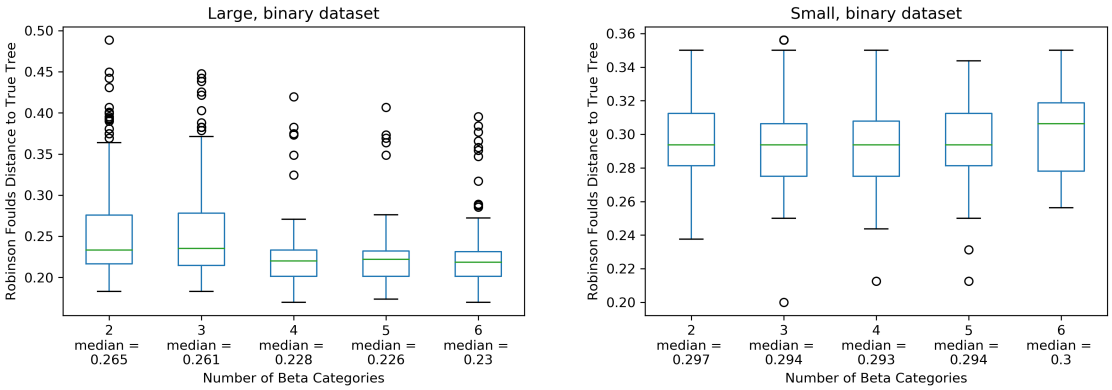
Multistate Data

In the large, multistate data sets, accuracy is improved using a correctly-specified model over both over- and underparameterized models. In the small datasets, a pattern similar to that of the large binary datasets is observed, in which underparameterization negatively affected accuracy, while overparameterization did not. As with binary data, dataset size has a strong relationship with phylogenetic accuracy. However, the multistate datasets show a stronger relationship between appropriate parameterization and accuracy. This is especially obvious in the small multistate datasets, in which choosing a correct model improves accuracy by approximately 40%.

3.3.3 Model Testing

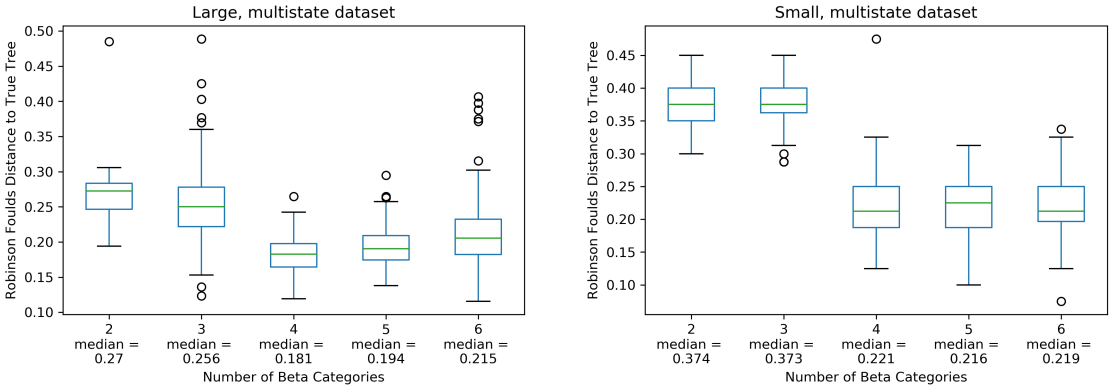
Model Averaging using reversible jump Markov chain Monte Carlo

In using rjMCMC for model selection, mixed results were obtained. While including all the rate categories $k = 2, 3, 4, 5, 6, 7, 8$ for the SHDM and the Mk model, the MCMC algorithm strongly preferred rate category $k = 8$. In addition to this, when including rate categories $k = 6$ and $k = 8$ with Mk model, it strongly chose rate category $k = 8$. One thing to keep in mind here is that the data used was the combined multistate data. The stepping stone sampling preferred $k = 7$ for the same dataset.



(a) (b)

Figure 3.5: (a) Accuracy of phylogenetic estimation in large (135 taxa, 163 character) simulated binary datasets under the Beta model, with 4 discrete categories. Underparameterized models have a wider spread of accuracy values than appropriately or overparamterized models. (b) Accuracy of phylogenetic estimation in small (41 taxa, 26 characters) simulated binary datasets under the Beta model. The correct number of simulation categories is 4. In datasets this size, there does not seem to be a relationship between parameterization and accuracy.



(a) (b)

Figure 3.6: (a) Accuracy of phylogenetic estimation in large (135 taxa, 163 character) simulated multistate datasets under the SHDM with four transition rate asymmetry categories. Both over- and underparameterization appear to be detrimental to accuracy, though underparamterization is worse. (b) Accuracy of phylogenetic estimation in small (41 taxa, 26 characters) simulated multistate datasets under the SHDM. The correct number of simulation categories is 4. Much as in the large, binary dataset, underparameterization appears to be more detrimental than overparameterization.

Mk model was not selected for in any of the analyses.

Posterior Predictive Simulation

Posterior predictive simulation gave a distribution of chosen summary statistics which I then compared to the empirical summary statistic. I also calculated the p-values for each summary statistic based on comparing empirical and posterior predictive test statistic values. Using these test statistics allows us to convert the empirical data and the posterior predictive output to numerical summary values that can be compared. The effect sizes calculated for each summary statistic is an indication of the ability of model to explain the data. It is a measure of difference between the observed data and the data that would be expected under the model. It is calculated as the difference between the empirical test statistic value and the median of the posterior distribution of the simulated value, divided by the standard deviation of the posterior predictive distribution. A large effect size would indicate that the observed data would be significantly different from the data that would be expected under the model indicating a bad fit of the model. On the other hand, a small effect size indicates that the observed data is not significantly different from the data simulated under the model which would indicate the model is a good fit. In simple terms, if the value of effect size is closer to zero, the model can be considered more adequate to the data.

Data PPS To compare the models, I calculated p-values for all the test statistics listed in **Data PPS** under 3.2.3. A p-value of 0 or 1 indicates that the summary statistic is not able to explain the data well under the model of choice.

There was no consistent pattern between the models or the summary statistics to test the combined dataset.

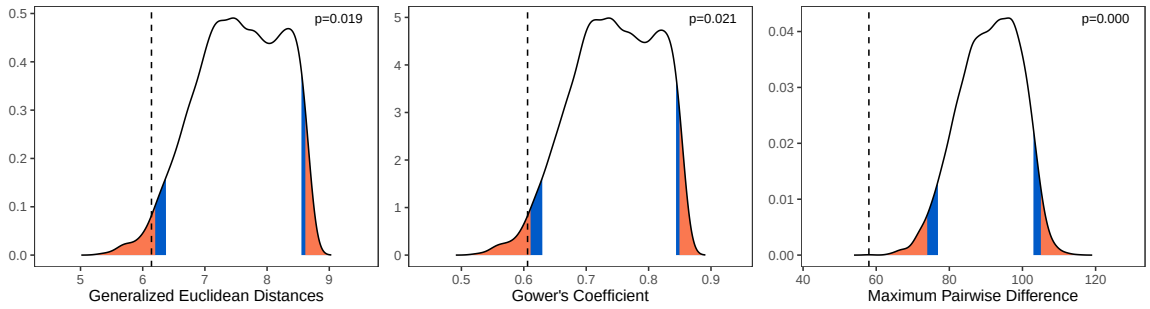


Figure 3.7: Density graphs for GED, GC, and Maximum Pairwise Difference under the Mk model.

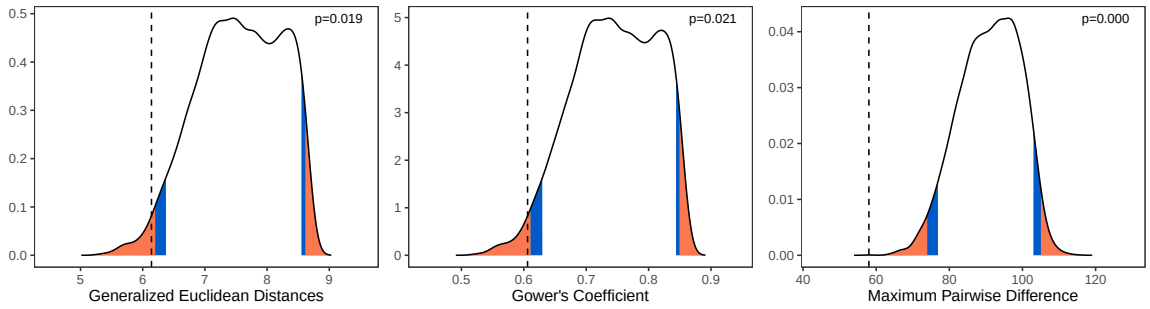


Figure 3.8: Density graphs for GED, GC, and Maximum Pairwise Difference under the SHDM model with 3 categories.

A table of the summary statistics and the p-values for all the models can be seen in table 3.1. For data PPS, the effect size of the summary statistics varied widely. Most of the effect sizes were in the higher range (> 20) indicating an inadequate model fit according to the test statistic. The only test statistic that produced a suitable value of effect size was Maximum Variable Block Length for which the effect size was 0.25.

Some summary statistics that define the morphological data well can be seen in figures 3.7, 3.8, 3.9, and 3.10. The full list of figures for the summary statistic used

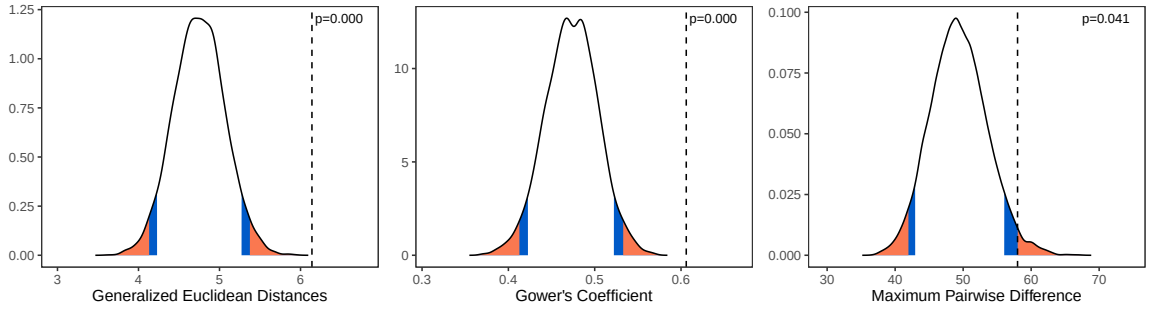


Figure 3.9: Density graphs for GED, GC, and Maximum Pairwise Difference under the SHDM model with 6 categories.

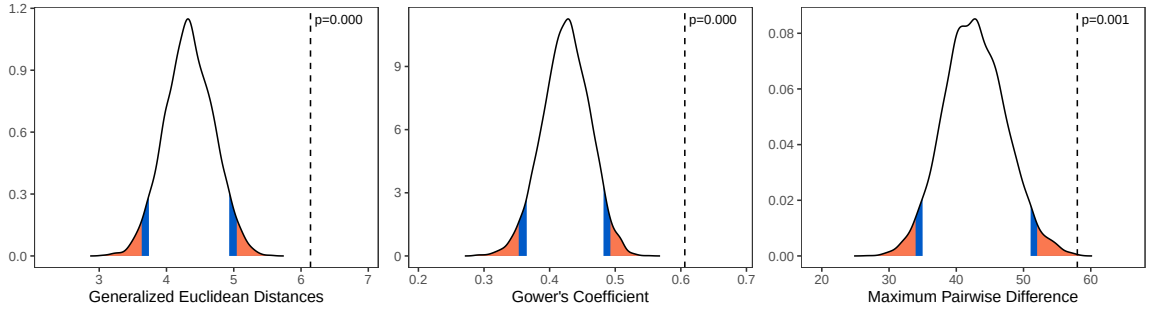


Figure 3.10: Density graphs for GED, GC, and Maximum Pairwise Difference under the SHDM model with 8 categories.

can be found in appendix A.

Inference PPS In inference PPS, randomly selected simulated datasets were used to infer tree under the model in question and it was compared using summary statistics listed in section 3.2.3. The SHDM model that was used contained 3 rate categories. Amongst the nine summary statistics used, some of them (quantiles of RF) were redundant and have been excluded from the figure here.

Robinson Foulds distance is a metric that shows the topological distance scaled by the branch lengths on the trees. Comparing this between the list of inferred trees from the simulated and empirical data shows that trees inferred using SHDM model

Summary Stats	**Mk**	**3-category SHDM**	**6-category SHDM**	**8-category SHDM**
Number of Invariant Sites	1.000	1.000	1.000	0.370
Maximum Pairwise Difference	0.000	0.736	0.041	0.001
Maximum Variable Block Length	0.049	0.255	0.459	0.375
Gower's Coefficient	0.021	0.006	0.000	0.000
Generalized Euclidean Distances	0.019	0.006	0.000	0.000

Table 3.1: P-Values for selected summary statistics used in Data PPS. Values colored in grey indicate the best model according to the summary statistic.

($p = 0.250$) are closer to the empirical tree than the trees inferred with the Mk model ($p = 0.020$). On the other hand, the mean tree length for the trees inferred using the Mk model is more closer to the empirical tree than the trees inferred using the SHDM model. The p-value for the Mk mean tree length is 0.850 whereas the p-value for the SHDM mean tree length is just 0.230. Although the tree length do not match closely while using the SHDM model, the variance in tree length is smaller when using it to infer trees using the simulated data. The tree lengths varied more amongst the inferred tree while using Mk model. The p-value for variance in tree length for Mk model is 0 while the p-value for variance in tree length for SHDM model is 0.900.

The effect sizes for the inference summary statistics show a similar pattern. The Mean Robinson Foulds and Variance in Tree Length indicate that the SHDM model is more adequate with effect sizes of 1.104 and 0.216 respectively. On the other hand, Mean Tree Length indicates that the Mk model is more adequate with a effect size of 1.064 which is similar to that for SHDM (1.075).

Different inference summary statistics show different result in model adequacy

using inference PPS. This can be seen in figures 3.11, and 3.12.

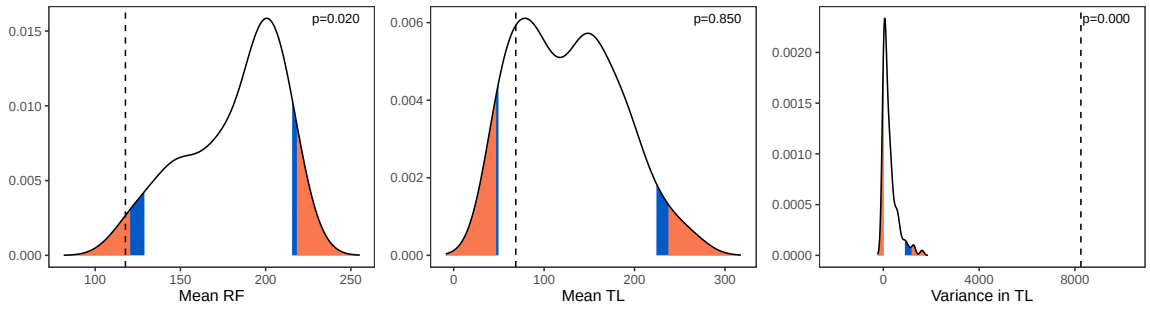


Figure 3.11: Density graphs for Mean Robinson Foulds, Mean Tree Length, and Variance in Tree Length under the Mk model.

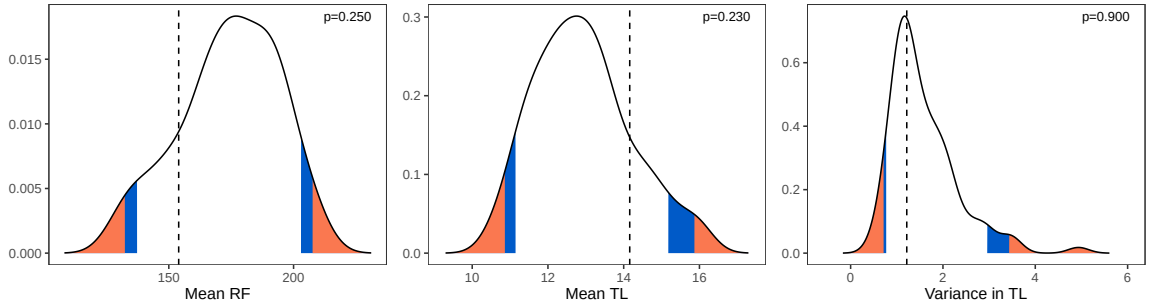


Figure 3.12: Density graphs for Mean Robinson Foulds, Mean Tree Length, and Variance in Tree Length under the SHDM model.

Mixed PPS This category of PPS uses both the datasets and inferred trees in the calculation of summary statistics. Given the lack of well-defined summary statistics for morphological data, I used consistency and retention indices to explore if it could be used to explain the data under a model of choice.

The figures with consistency and retention indices for the Mk and SHDM models can be seen in figure 3.13.

As shown in the figure, both CI and RI have p-value of 0.00 for both the models indicating either the models are not adequate for the data or the summary statistics

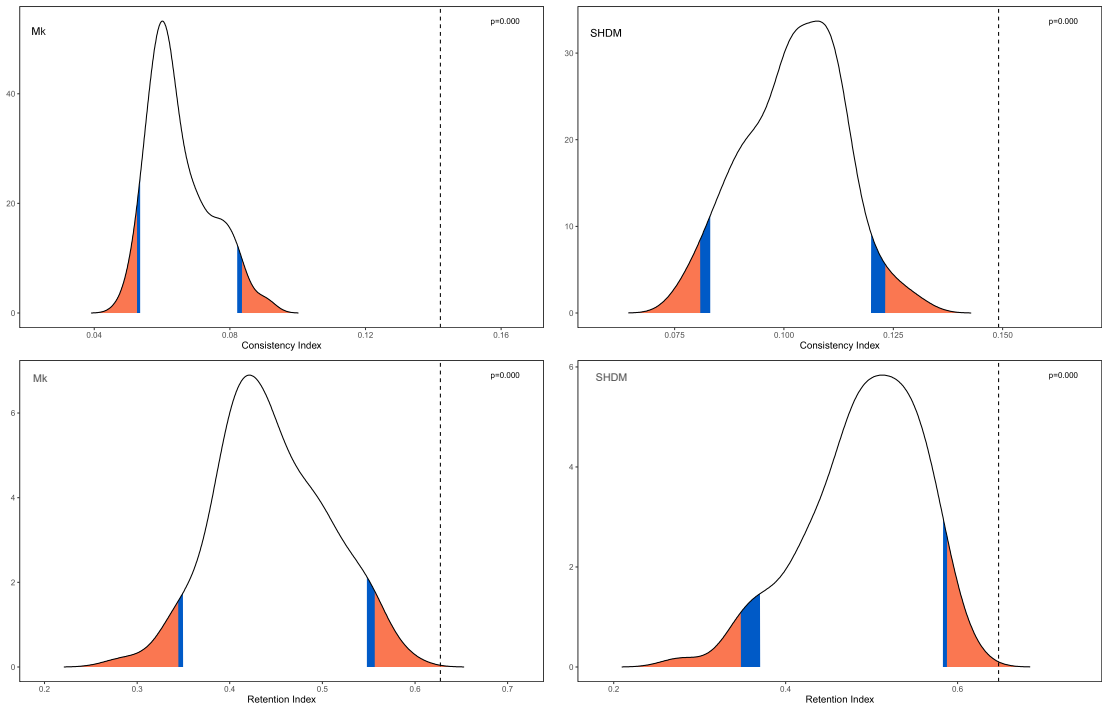


Figure 3.13: Density graphs for Consistency and Retention Indices for both Mk and SHDM models.

cannot explain the data that well. The effect sizes for both CI and RI for SHDM model were 4.17 and 2.34 respectively and for the Mk model were 8.41 and 3.08 respectively. Although the effect sizes were smaller for the SHDM model, it does not indicate that the model is adequate for the data.

3.4 Discussion

The discussion surrounding the incorporation of morphological data into phylogenetic analyses has largely focused on two endpoints on a spectrum: Parsimony and the Mk model (Lewis, 2001). The idea behind parsimony is simple: The tree with the least changes implied should be favored. The simplicity of this idea belies the complexity of the implications about character evolution. Each character can have its own length (number of steps), with one common tree for the whole dataset. When written out as a likelihood model, parsimony is referred to as the No Common Mechanisms model (Tuffley and Steel, 1997), referring to the fact that a different evolutionary process is modeled per character. However, though the model is statistically consistent under many circumstances (Steel, 2010), it has been demonstrated that this model is so complex, and features so many parameters that it is never chosen by even a liberal information criterion (Holder et al., 2010). Subsequent work involving small empirical and simulated datasets supported the use of biologically-inspired models over the No Common Mechanisms model (Huelsenbeck et al., 2011).

In many ways, the Mk model is similar to an unweighted parsimony model, assuming that the transitions between any two states are equally likely. But the Mk model assumes a common mechanism across sites, a generalized Jukes-Cantor model.

Application of among site rate variation (ASRV) allows characters in the same dataset to have differing rates of evolution, accommodating natural variation in phenotypic characters (Yang, 1994). Allowing multiple substitutions on a branch is known to be a benefit in phylogenetic analysis (Felsenstein, 1978), and incorporating rate variation can potentially be more important than an incorrectly-specified model of evolution (Lemmon and Moriarty, 2004). Previous work in morphological phylogenetics has also highlighted the importance of appropriate assumptions about rate heterogeneity (Harrison and Larsson, 2015).

The model discussed here is an effort to apply a similar framework to ASRV, but for the actual model of evolution. We may know little about the probability of a particular transition between two character states *a priori*, in the same way that we may know little about the rate of evolution of any particular site or character in a matrix. Under the discrete Beta model and the SHDM, state frequencies are drawn from either a Beta (binary data) or Dirichlet (multistate data) prior distribution. The site likelihoods are then computed by marginalizing over the value of the vector of state frequencies, π_i. Mechanistically, this is very similar to the way in which ASRV is calculated.

In practice, these model allows us to assume multiple possible mechanisms across a single dataset. Previous work on similar models (Nylander et al., 2004) in the software MrBayes (Wright et al., 2016) has found that the Mk model is supported using steppingstone model selection in about half of phylogenetic datasets tested. However, for modeling the other half, a single common mechanism is not adequate. In this chapter, I have implemented the Beta model from MrBayes, and clarified a model for multistate characters to allow modeling of evolution when a single, common

mechanism cannot be assumed.

Relaxing the assumption of a single mechanism with equal transitions between all states is particularly important for this dataset. Cretaceous ants are known to have features that are both similar to ants and to wasps (Wilson et al., 1967; Barden and Grimaldi, 2016). The wasp-like characters generally become lost, and are not regained in crown ants. By contrast, ant-like characters are often gained but not lost. This implies that for the datasets used in this paper, multiple evolutionary mechanisms will apply among characters. Some should have a bias towards losses, some towards gains. These biases violate the assumptions of the Mk model, but do not argue for there being no common mechanisms at all across sites. Together, these character classes argue strongly for a symmetric Beta distribution, in which if there is a set of characters with a certain bias (for example, more '0' to '1' transitions), it is expected that there would be another set of characters with an opposite bias. These data, therefore, make an excellent test case for a model that lies somewhere on a continuum of models between the Mk model and parsimony.

In simulation, for larger datasets, underparameterization of the morphological model appears to have more serious deleterious effects on accuracy than does overparameterization. This is expected, as statistical literature indicates that underparameterization often leads to error (Revell et al., 2005), but overparameterization often leads to less certainty (Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Brown and Lemmon, 2007). In the small datasets, we observe different dynamics. The binary datasets do not show a relationship between the appropriateness of the model and accuracy. These datasets are quite small - 41 taxa and 26 characters, modeled on Barden and Grimaldi (2015). This dataset is on the extreme small end

of the datasets explored by Wright et al. (2016). For the multistate data, the choice of model greatly affects accuracy. Under an overly-simple model of evolution, phylogenetic error is nearly 40%. But under the correctly-specified model, the error is cut by about 40%. Why should these two datasets show such differences? It's possible that there is simply not enough information to support a more complex model for the binary data. The multistate dataset has nearly twice as many characters for the same amount of taxa. Secondly, with more character states, and more state frequency parameters, the model violations associated with under- and overparameterization are more severe in the multistate datasets.

Each empirical dataset strongly supports the use of a model that relaxes the assumption of equal transition rates. How many different asymmetrical transition categories is supported differs among datasets (Fig. 2). Even with strongly supported models, the tree is still poorly resolved. The lack of resolution in our recovered topologies is consistent with the lack of resolution in the papers from which the matrices were sampled (Barden and Grimaldi, 2016). Lineages in Barden and Grimaldi (2016) were sampled to maximize representation among the recognized subfamilies, and to include as many stem lineages as possible. The Keller (2011) matrix was assembled to achieve maximum representation in the previously undersampled ponerimorph groups. Ants are known to have strong convergence within characters. The group is large and diverse, but fairly morphologically constrained. The morphological matrices of Barden & Grimaldi and Keller were assembled to avoid homoplasious characters, such as sculpturing and body shape. However, in such a large and diverse group, not all homoplasy can be avoided, and previous phylogenetic and taxonomic work in the group has also reported a lack of support for well-supported molecu-

lar clades due to a dearth of morphological synapomorphies among these groups. While likelihood-based methods are expected to fare better than parsimony when convergence is an issue (Felsenstein, 1978), they are not perfect and can be mislead, particularly in the presence of model violations such as structured missing data. Even in these challenging conditions, our methods do reasonably well at teasing apart homoplasy from phylogenetic signal. For example, *Martialis heureka*, an ant that is the only known member of the subfamily Martialinae (Rabeling et al., 2008), is weakly supported as sister to *Amblyopone pallipes*. This is not a placement supported by molecular phylogenetics (Rabeling et al., 2008), but does indicate the ability of the method to disentangle signal in the presence of convergent characters, such as elongate mandibles strongly reminiscent of stem ants, particularly the Haidonmyrmecine ants.

If, for small matrices, a model that is better fit does not yield a more resolved tree, or stronger support for the bifurcations that are discovered, what does this mean for morphologists? This model is more computationally intense, meaning waiting longer for a result that may not represent a substantial improvement. Pragmatically, prior work on phylogenetic estimation from discrete morphological data has supported that model choice is more important when the problem is challenging (Wright and Hillis, 2014; Wright et al., 2016). While a tree may not be more resolved, it is less likely to be incorrect if the model of evolution is not incorrect. Even if more resolution cannot be recovered, either due to homoplasy or lack of information, the correct model can help us avoid selecting an incorrect tree.

Philosophically, the choice of model should reflect the researcher's knowledge of the data. Models inherently simplify the process of evolution; it is not possible to

include every biological process in any model, particularly given the biased sample of the past that we have. However, we can hope to increase the availability of models that capture relevant axes of variation. ASRV allows researchers to model different rates of character state change within a dataset. Likewise, the discrete Beta model and the SHDM model allow researchers to model asymmetrical forward and backward transition rates among characters. In the case of ants, the existence of characters with a strong signature of loss (wasp-like characters) and characters with a stronger signature of gain (ant apomorphies) in the data suggest that it should be expected to find strong support for a model that allows this variation. If this is our expectation, then this more complex model should remain in the set of models tested for fit. Indeed, this assumption is borne out as accurate in these data (Fig. 3.4). Therefore, even if measurably more resolution is not forthcoming, the philosophical underpinnings of the model remain important.

3.4.1 Model Adequacy for Morphological Data

Morphological data, even though very essential for time dating phylogenetic studies, have not been explored much as molecular data. With inclusion for fossils, models that explain the data well are very essential for phylogenetic studies using morphology. Model adequacy is one of the powerful methods that help researchers asses the absolute fitness of a model to the data being used. While assessing for adequacy, researcher needs to have a deep knowledge of the data/organism being studied.

In the case of comparison between Mk and SHDM models, neither of the models seem particularly adequate. There might be two reasons that causes the results we have seen. The first might be that the model is not adequate for analyses of

morphological data. Another reason might be that the summary statistics that I have used might not explain the models and its exploration of data well. There is a lack of well-defined summary statistics for morphological data and additional simulation studies to understand the behavior of summary statistics would be useful in morphological phylogenetic studies. Some studies have mentioned that the summary statistics like Generalized Euclidean Distances (GED) is influenced by the amount of missing data in the matrix (Lehmann et al., 2019). To account for this I used Gower's coefficient (GC) which is supposed to account for missing data as well. But in the results that I have obtained, both GC and GED show similar results for all the models in consideration. Summary statistics (Consistency and Retention Indices) that I have used in mixed PPS while used widely in paleobiology, are based on parsimony (Farris, 1989, 1970; Goloboff, 1991). This study was based on probabilistic models which do not conform to parsimony expectations which might be one of the reasons that the summary statistics were unable to explain the data under the models in consideration.

In the case of Inference PPS and Mixed PPS, I had analysed the SHDM model with 3 rate categories which could be causing erroneous results in the analyses. Due to computational limitations, analyses with 6 and 8 rate categories for SHDM were not accomplished. This could be a further addition to this study for a more robust result. Per this study, the models that were under consideration did not appear adequate for the dataset being used. Further work is needed for a better working mechanistic model of evolution for morphological data. This might be difficult as morphological data cannot be generalized unlike molecular data. With more mathematical advancements, inclusion of more biologically realistic parameters in phylogenetic studies is possible and must be analysed according to the nature of data under study.

The modeling of discrete morphological characters has long been constrained to the end points on a spectrum of possible models. Parsimony represents one extreme in terms of the elaboration of the mechanism, allowing each character to have its own tree and length. The Mk model represents the other, in its simplicity. Researchers may have reasons to believe that either one of these models is a good descriptor of their data, or that assumptions of one model fit better than the other, but that the framework (frequentist or Bayesian) makes more suitable assumptions about the process of evolution. Use of new combinations of priors, and borrowing from the molecular toolkit enables researchers to explore more widely the continuum between these two points in a Bayesian context. Bayesian modeling allows researchers to make use of well-described model fit procedures to test the how well a model describes the data. RevBayes enables these sorts of explorations through use of modular design, allowing almost any parameter to be combined with most priors, and including the functions to instantly simulate under the model. The framework enables assumptions about morphological evolution to be advanced, and tested rigorously.

Bibliography

Barden, P. and D. A. Grimaldi. 2016. Adaptive radiation in socially advanced stem-group ants from the cretaceous. Current Biology 26:515–521.

Barido-Sottani, J., G. Aguirre-Fernández, M. J. Hopkins, T. Stadler, and R. Warnock. 2019. Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth–death process 286:20190685.

Benson, R. B. and J. N. Choiniere. 2013. Rates of dinosaur limb evolution provide evidence for exceptional radiation in mesozoic birds. Proc. R. Soc. B 280:20131780.

Betancur-R, R., G. Ortí, and R. A. Pyron. 2015. Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes. Ecology Letters 18:441–450.

Bolton, B. 2003. Synopsis and classification of Formicidae. American Entomological Institute.

Brady, S. G., T. R. Schultz, B. L. Fisher, and P. S. Ward. 2006. Evaluating alternative hypotheses for the early evolution and diversification of ants. Proceedings of the National Academy of Sciences 103:18172–18177.

Brown, J. M. 2014a. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit 63:334–348.

Brown, J. M. 2014b. Predictive approaches to assessing the fit of evolutionary models 63:289–292.

Brown, J. M. and A. R. Lemmon. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics 56:643–655.

Brown, J. W., C. Parins-Fukuchi, G. W. Stull, O. M. Vargas, and S. A. Smith. 2017. Bayesian and likelihood phylogenetic reconstructions of morphological traits are not discordant when taking uncertainty into consideration: a comment on puttick et al. Proc. R. Soc. B 284:20170986.

Cobbett, A., M. Wilkinson, and M. A. Wills. 2007. Fossils impact as hard as living taxa in parsimony analyses of morphology. Systematic Biology 56:753–766.

de Queiroz, K. and J. Gauthier. 1992. Phylogenetic taxonomy. Annual Review of Ecology and Systematics 23:449–480.

Duchêne, D. A., S. Duchêne, and S. Y. Ho. 2018. Differences in performance among test statistics for assessing phylogenomic model adequacy. Genome Biology and Evolution 10:1375–1388.

Duchêne, D. A., S. Duchêne, E. C. Holmes, and S. Y. Ho. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. Molecular Biology and Evolution 32:2986–2995.

Duchene, S., R. Bouckaert, D. A. Duchene, T. Stadler, and A. J. Drummond. 2019. Phylodynamic model adequacy using posterior predictive simulations. Systematic Biology 68:358–364.

Dunn, C. W., F. Zapata, C. Munro, S. Siebert, and A. Hejnol. 2018. Pairwise comparisons across species are problematic when analyzing functional genomic data. Proceedings of the National Academy of Sciences 115:E409–E417.

Ezard, T. H. G., G. H. Thomas, and A. Purvis. 2013. Inclusion of a near-complete fossil record reveals speciation-related molecular evolution. Methods in Ecology and Evolution 4:745–753.

Farris, J. S. 1970. Methods for computing wagner trees. Systematic Biology 19:83–92.

Farris, J. S. 1989. The retention index and the rescaled consistency index. Cladistics: the international journal of the Willi Hennig Society 5:417–419.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology 27:401–411.

Felsenstein, J. 1985. Phylogenies and the comparative method Pages 1–15.

Goloboff, P. A. 1991. Homoplasy and the choice among cladograms. Cladistics 7:215–232.

Goloboff, P. A., M. Pittman, D. Pol, and X. Xu. 2018a. Morphological data sets fit a common mechanism much more poorly than dna sequences and call into question the mkv model. Systematic Biology Page syy077.

Goloboff, P. A., A. Torres, and J. S. Arias. 2018b. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. Cladistics 34:407–437.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. Biometrics Pages 857–871.

Green, P. J. and D. I. Hastie. 2009. Reversible jump mcmc. Genetics 155:1391–1403.

Grimaldi, D. A., D. Agosti, and J. M. Carpenter. 1997. New and rediscovered primitive ants (hymenoptera, formicidae) in cretaceous amber from new jersey, and their phylogenetic relationships. american museum novitates; no. 3208 .

Harrison, L. B. and H. C. E. Larsson. 2015. Among-character rate variation distributions in phylogenetic analysis of discrete morphological characters. Systematic Biology 64:307–324.

Harvey, P. H. and M. D. Pagel. 1991. The comparative method in evolutionary biology vol. 239. Oxford university press Oxford.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences, USA 111:E2957–E2966.

Helfrich, P., E. Rieb, G. Abrami, A. Lücking, and A. Mehler. 2018. TreeAnnotator: Versatile visual annotation of hierarchical text relations. *in* Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) European Language Resources Association (ELRA), Miyazaki, Japan.

Hennig, W. and D. D. Davis. 1966. Phylogenetic systematics. University of Illinois Press.

Höhna, S., L. M. Coghill, G. G. Mount, R. C. Thomson, and J. M. Brown. 2018. P3: Phylogenetic posterior prediction in revbayes. Molecular biology and evolution 35:1028–1034.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference

using graphical models and an interactive model-specification language. Systematic Biology 65:726–736.

Holder, M. T., P. O. Lewis, and D. L. Swofford. 2010. The Akaike Information Criterion Will Not Choose the No Common Mechanism Model. Systematic Biology 59:477–485.

Hopkins, M. J. and A. B. Smith. 2015. Dynamic evolutionary change in post-paleozoic echinoids and the importance of scale when interpreting changes in rates of evolution. Proceedings of the National Academy of Sciences 112:3758–3763.

Huelsenbeck, J. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Huelsenbeck, J. P. and B. Rannala. 2004. Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. Systematic Biology 53:904–913.

Huelsenbeck, J. P., M. A. Suchard, and M. E. Alfaro. 2011. Biologically Inspired Phylogenetic Models Strongly Outperform the No Common Mechanism Model. Systematic Biology 60:225–232.

Jukes, T. and C. Cantor. 1969. Evolution of protein molecules. Mammalian Protein Metabolism 3:21–132.

Keller, R. A. 2011. A phylogenetic analysis of ant morphology (hymenoptera: Formicidae) with special reference to the poneromorph subfamilies. Bulletin of the american museum of natural history Pages 1–90.

Kirkpatrick, J. B., E. A. Walsh, and S. D'Hondt. 2016. Fossil dna persistence and decay in marine sediment over hundred-thousand-year to million-year time scales. Geology 44:615.

Kluge, A. G. and J. S. Farris. 1969. Quantitative phyletics and the evolution of anurans. Systematic Biology 18:1–32.

Lartillot, N., T. Lepage, and S. Blanquart. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot, N. and H. Philippe. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular biology and evolution 21:1095–1109.

Lehmann, O. E., M. D. Ezcurra, R. J. Butler, and G. T. Lloyd. 2019. Biases with the generalized euclidean distance measure in disparity analyses with high levels of missing data. Palaeontology 62:837–849.

Lemmon, A. R. and E. C. Moriarty. 2004. The importance of proper model assumption in bayesian phylogenetics. Systematic Biology 53:265–277.

Letsch, H. O., P. Kück, R. R. Stocsits, and B. Misof. 2010. The impact of rrna secondary structure consideration in alignment and tree reconstruction: Simulated data and a case study on the phylogeny of hexapods. Molecular Biology and Evolution 27:2507–2521.

Lewis, P. O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic biology 50:913–925.

Lewis, P. O., W. Xie, M.-H. Chen, Y. Fan, and L. Kuo. 2014. Posterior predictive Bayesian phylogenetic model selection 63:309–321.

Lloyd, G. T. 2016. Estimating morphological diversity and tempo with discrete character-taxon matrices: implementation, challenges, progress, and future directions. Biological Journal of the Linnean Society 118:131–151.

Magallón, S. 2010. Using fossils to break long branches in molecular dating: A comparison of relaxed clocks applied to the origin of angiosperms. Systematic Biology 59:384–399.

May, M. R. and C. J. Rothfels. 2023. Diversification models conflate likelihood and prior, and cannot be compared using conventional model-comparison tools. Systematic Biology Page syad010.

Morayma, S. and M. Kraemer. 2007. Systematic, palaeoecology, and palaeobiogeography of the insect fauna from mexican amber. Palaeontographica Abteilung A Pages 1–133.

Moreau, C. S. and C. D. Bell. 2013. Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. Evolution 67:2240–2257.

Moreau, C. S., C. D. Bell, R. Vila, S. B. Archibald, and N. E. Pierce. 2006. Phylogeny of the ants: diversification in the age of angiosperms. Science 312:101–104.

Near, T. J., A. Dornburg, M. Tokita, D. Suzuki, M. C. Brandley, and M. Friedman. 2014. Boom and bust: Ancient and recent diversification in bichirs (polypteridae: Actinopterygii), a relictual lineage of ray-finned fishes. Evolution 68:1014–1026.

Nel, A., G. Perrault, and D. Néraudeau. 2004. The oldest ant in the lower cretaceous amber of charente-maritime (sw france)(insecta: Hymenoptera: Formicidae). Geologica Acta: an international earth science journal 2:23–30.

Nicholls, G. K. and R. D. Gray. 2006. Quantifying uncertainty in a stochastic model of vocabulary evolution. Phylogenetic methods and the prehistory of languages Pages 161–171.

Nylander, J. A., F. Ronquist, J. P. Huelsenbeck, and J. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data 53:47–67.

O'Meara, B. 2012. Evolutionary inferences from phylogenies: A review of methods 43.

Orlando, L., A. Ginolhac, G. Zhang, D. Froese, A. Albrechtsen, M. Stiller, M. Schubert, E. Cappellini, B. Petersen, I. Moltke, et al. 2013. Recalibrating equus evolution using the genome sequence of an early middle pleistocene horse. Nature 499:74.

Ouellette, G. D., B. L. Fisher, and D. J. Girman. 2006. Molecular systematics of basal subfamilies of ants using 28s rrna (hymenoptera: Formicidae). Molecular phylogenetics and evolution 40:359–369.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Systematic Biology 53:571–581.

Perrichot, V. and V. Girard. 2009. A unique piece of amber and the complexity of ancient forest ecosystems. Palaios 24:137–139.

Pie, M. R. and M. K. Tschá. 2009. The macroevolutionary dynamics of ant diversification. Evolution: International Journal of Organic Evolution 63:3023–3030.

Price, S. L., S. Powell, D. J. C. Kronauer, L. A. P. Tran, N. E. Pierce, and R. K. Wayne. 2014. Renewed diversification is associated with new ecological opportunity in the neotropical turtle ants. Journal of Evolutionary Biology 27:242–258.

Puttick, M. N., J. E. O'Reilly, A. R. Tanner, J. F. Fleming, J. Clark, L. Holloway, J. Lozano-Fernandez, L. A. Parry, J. E. Tarver, D. Pisani, et al. 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. Proc. R. Soc. B 284:20162290.

Rabeling, C., J. M. Brown, and M. Verhaagh. 2008. Newly discovered sister lineage sheds light on early ant evolution. Proceedings of the National Academy of Sciences 105:14913–14917.

Revell, L. J., L. J. Harmon, and R. E. Glor. 2005. Under-parameterized Model of Sequence Evolution Leads to Bias in the Estimation of Diversification Rates from Molecular Phylogenies. Systematic Biology 54:973–983.

Robinson, D. F. and L. R. Foulds. 1981. Comparison of phylogenetic trees. Mathematical biosciences 53:131–147.

Ronquist, F. and J. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Saux, C., B. L. Fisher, and G. S. Spicer. 2004. Dracula ant phylogeny as inferred by nuclear 28s rdna sequences and implications for ant systematics (hymenoptera: Formicidae: Amblyoponinae). Molecular phylogenetics and evolution 33:457–468.

Schliep, K. P. 2011. phangorn: phylogenetic analysis in r. Bioinformatics 27:592–593.

Schrago, C. G., B. O. Aguiar, and B. Mello. 2018. Comparative evaluation of maximum parsimony and bayesian phylogenetic reconstruction using empirical morphological data. Journal of evolutionary biology 31:1477–1484.

Schweigger, A. F. 1819. Beobachtungen auf naturhistorischen Reisen: anatomisch-physiologische Untersuchungen über Corallen; nebst einem Anhange, Bemerkungen über den Bernstein enthaltend. Walter de Gruyter.

Slater, G. J., L. J. Harmon, and M. E. Alfaro. 2012. Integrating fossils with molecular phylogenies improves inference of trait evolution. Evolution 66:3931–3944.

Slater, G. J. and M. W. Pennell. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Systematic Biology 63:293–308.

Steel, M. 2010. Can We Avoid "SIN" in the House of "No Common Mechanism"? Systematic Biology 60:96–109.

Stevens, P. 1991. Character states, morphological variation, and phylogenetic analysis: a review. Systematic botany Pages 553–583.

Sullivan, J. and P. Joyce. 2005. Model selection in phylogenetics 36:445–466.

Swofford, D. L. 1985. Phylogenetic analysis using parsimony. Illinois Natural History Survey, Champaign, Illinois .

Tribble, C. M., W. A. Freyman, M. J. Landis, J. Y. Lim, J. Barido-Sottani, B. T. Kopperud, S. Hhna, and M. R. May. 2022. Revgadgets: An r package for visualizing

bayesian phylogenetic analyses from revbayes. Methods in Ecology and Evolution 13:314–323.

Tuffley, C. and M. Steel. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bulletin of mathematical biology 59:581–607.

Urbani, C. B., B. Bolton, and P. S. Ward. 1992. The internal phylogeny of ants (hymenoptera: Formicidae). Systematic Entomology 17:301–329.

Uyeda, J. C., R. Zenil-Ferguson, and M. W. Pennell. 2018. Rethinking phylogenetic comparative methods. Systematic Biology 67:1091–1109.

Wagner, G. P. 1989. The origin of morphological characters and the biological basis of homology. Evolution 43:1157–1171.

Ward, P. S., S. G. Brady, B. L. Fisher, and T. R. Schultz. 2005. Assembling the ant "tree of life"(hymenoptera: Formicidae). Myrmecologische Nachrichten 7:87–90.

Wiens, J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Systematic Biology 54:731–742.

Wiley, E. O. and B. S. Lieberman. 2011. Phylogenetics: theory and practice of phylogenetic systematics. John Wiley & Sons.

Wills, M. A. 1998. Crustacean disparity through the phanerozoic: comparing morphological and stratigraphic data. Biological Journal of the Linnean Society 65:455–500.

Wilson, E. O., F. M. Carpenter, and W. L. Brown. 1967. The first mesozoic ants, with the description of a new subfamily. Psyche: A Journal of Entomology 74:1–19.

Wright, A. M. and D. M. Hillis. 2014. Bayesian analysis using a simple likelihood model outperforms parsimony for estimation of phylogeny from discrete morphological data. PLoS One 9:e109210.

Wright, A. M., G. T. Lloyd, and D. M. Hillis. 2016. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors 65:602–611.

Xie, W., P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. 2010. Improving marginal likelihood estimation for bayesian phylogenetic model selection. Systematic biology 60:150–160.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. Journal of Molecular evolution 39:306–314.

Appendix A

Summary Statistics in Data PPS

Listed below are the summary statistics that were used for the Data based PPS. There were a total of 18 summary stats that were applicable to morphological data.

Excluding Ambiguous means that the ambiguous characters (missing characters or gaps) are excluded.

- Number Invariant Sites - This is the column which consists information with characters that have varying rates of evolution.
- Number Invariant Sites Excluding Ambiguous -
- Max Invariant Block Length - This column shows the maximum number of blocks that has no varying sites.
- Max Invariant Block Length Excluding Ambiguous -
- Max Pairwise Difference -This is the column with the statistic that intends to be sensitive to the model of rate-variation among site and/or among branches.

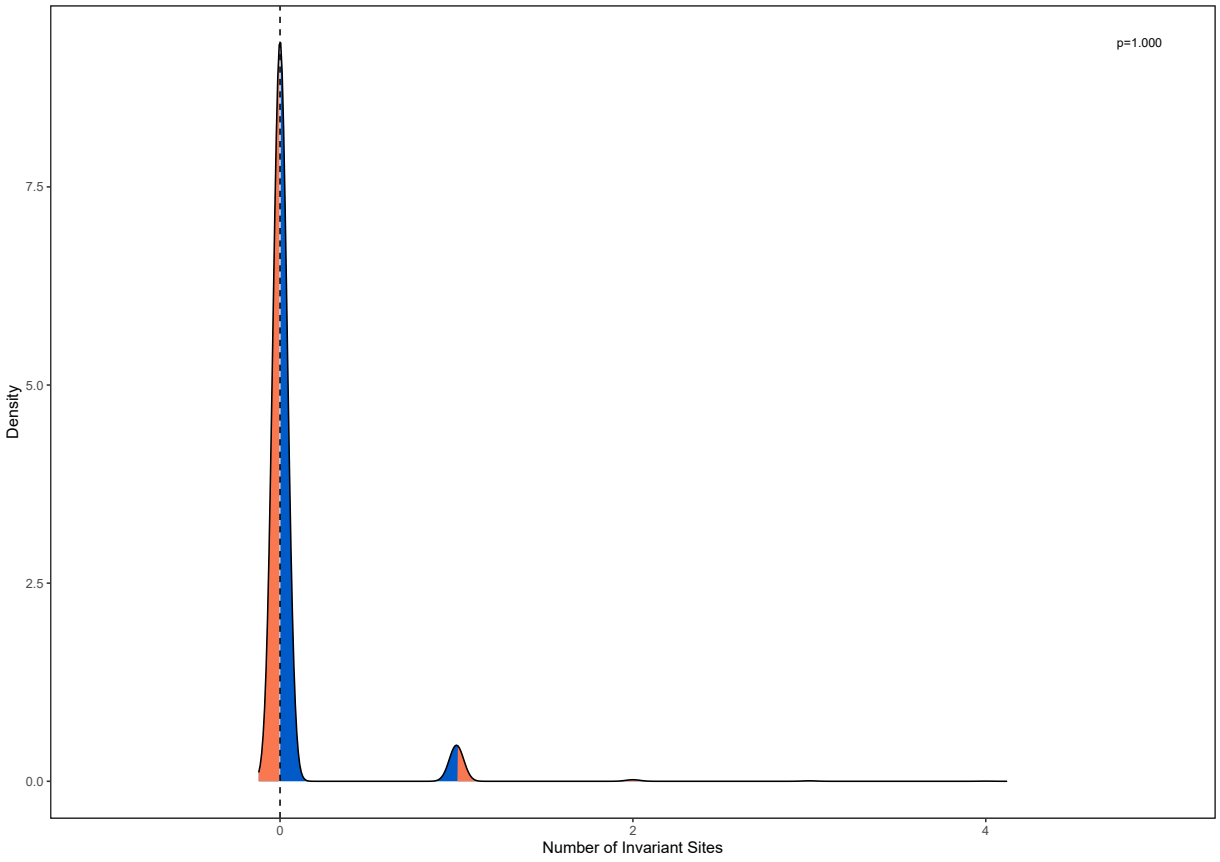
- Max Pairwise Difference Excluding Ambiguous -
- Max Variable Block Length - This column shows the maximum number of blocks that has varying sites.
- Max Variable Block Length Excluding Ambiguous -
- Min Pairwise Difference - This is the column with the statistic that finds the pair of sequences that has the smallest pairwise distance.
- Min Pairwise Difference Excluding Ambiguous -
- Number Invariable Block - This is the number of invariable blocks in the sequence.
- Number Invariable Block Excluding Ambiguous -
- Theta - This test statistic describes the genetic diversity in a population.
- Tajima-D - This test statistic is the difference between two measures of genetic diversity: the mean pairwise differences and the number of segregating sites scaled to be in a neutrally evolving population of constant size.
- Tajima-Pi - This is the sum of the pairwise differences divided by the number of pairs.
- Segregating-Sites - This test statistic show the positions that show polymorphisms between related genes in a sequence.
- Gower's Coefficient - Gower's Coefficient is a test statistic that can be used to measure the similarity between two organisms based on their physical char-

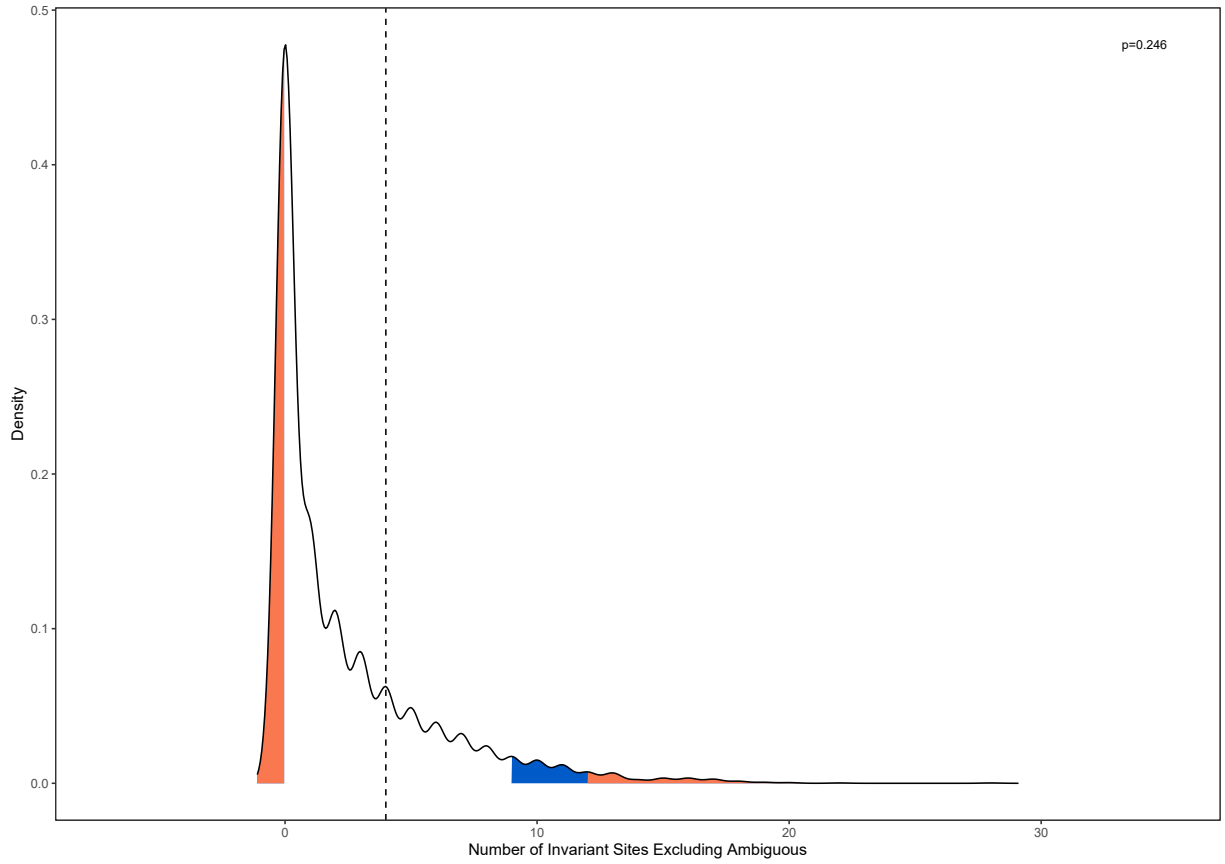
acteristics. It is similar to raw Euclidean distances but accounts for missing characters in the data matrix.

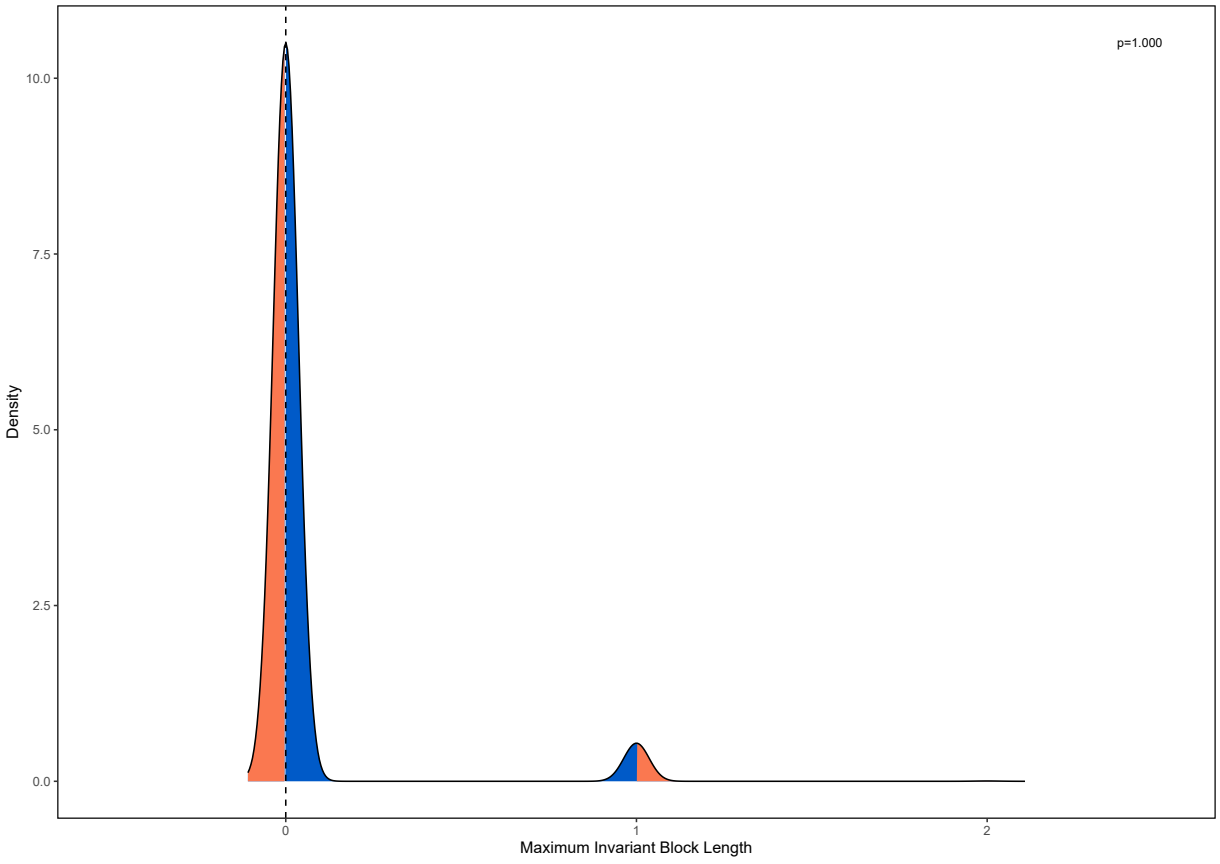
- Generalized Euclidean Distance - This test statistic is similar to Gower's coefficient but has a different way of dealing with missing characters.

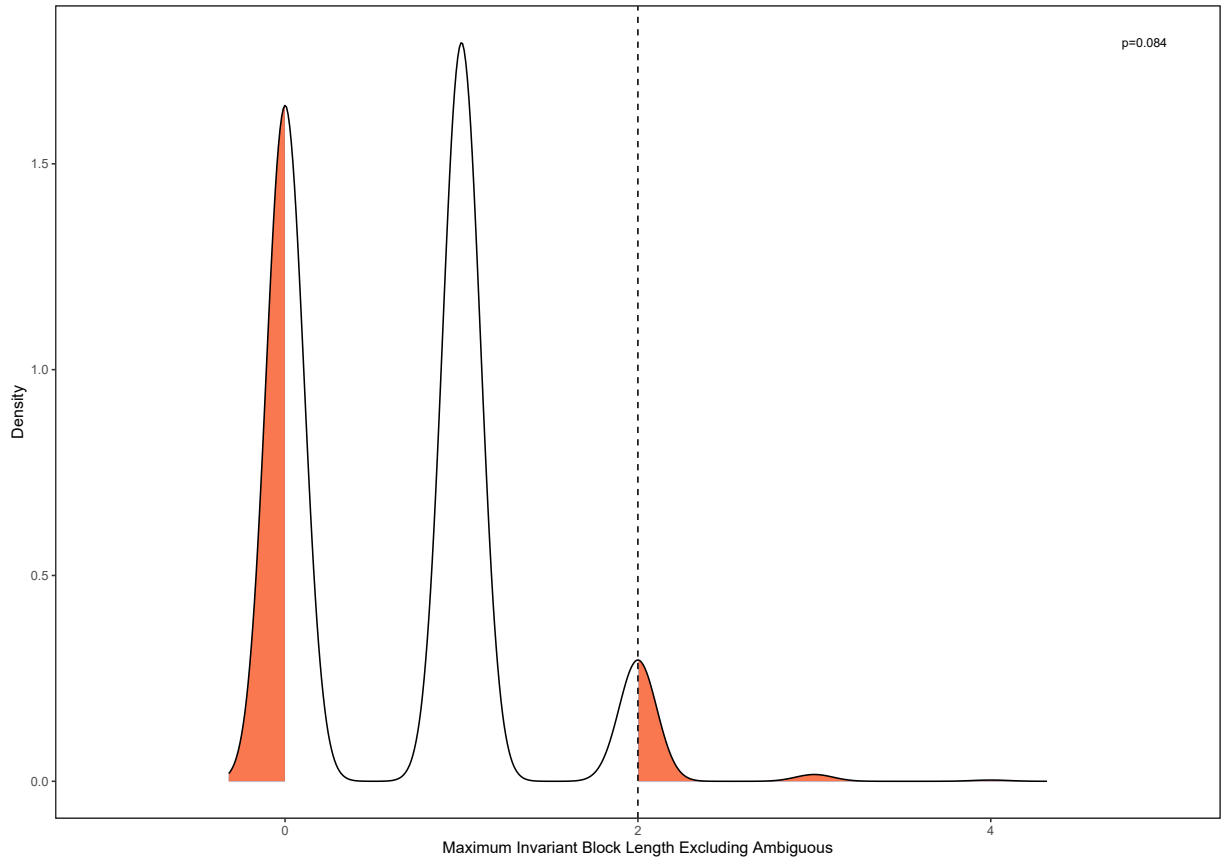
The figures for above listed test statistics are in the following pages.

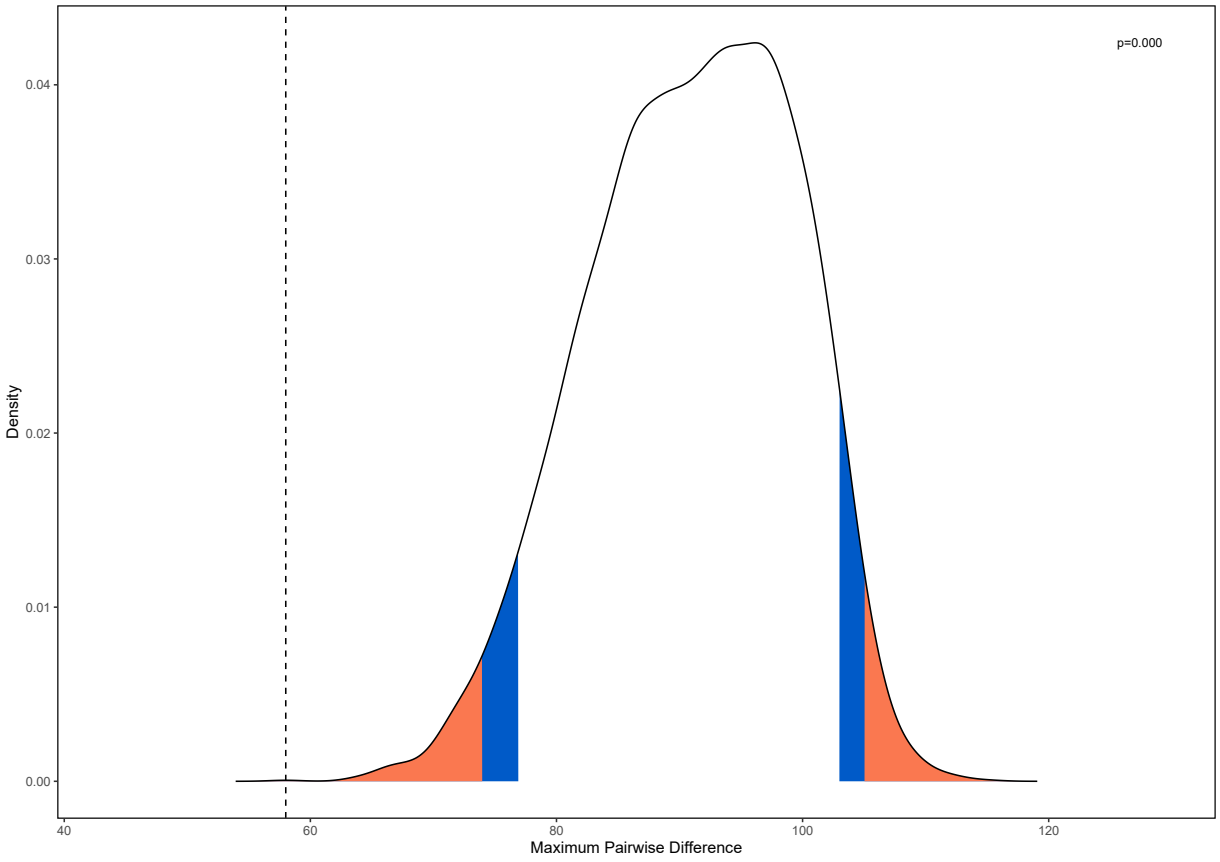
For Mk model, the density graphs for all the summary statistics are as follows.

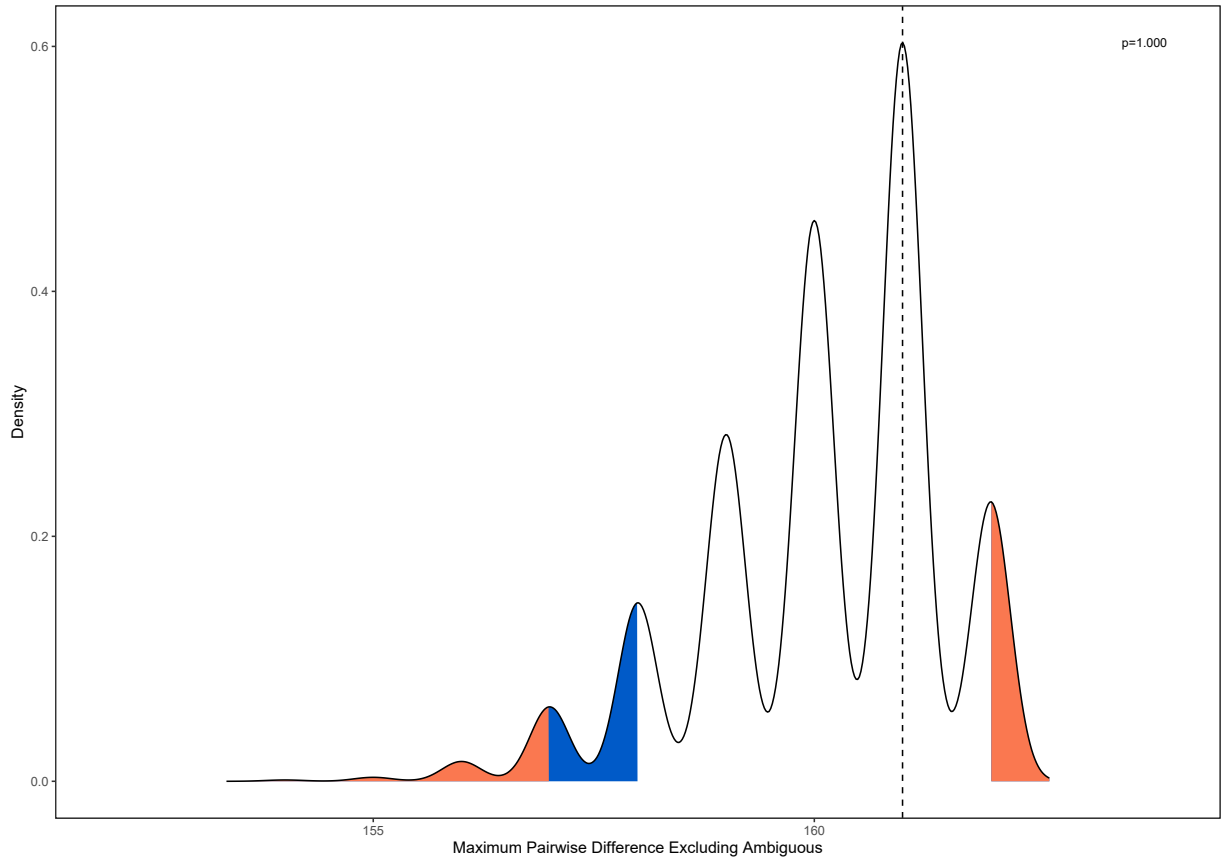

p=1.000
7.5
5.0
2.5
0.0
Density
0
2
4
Number of Invariant Sites

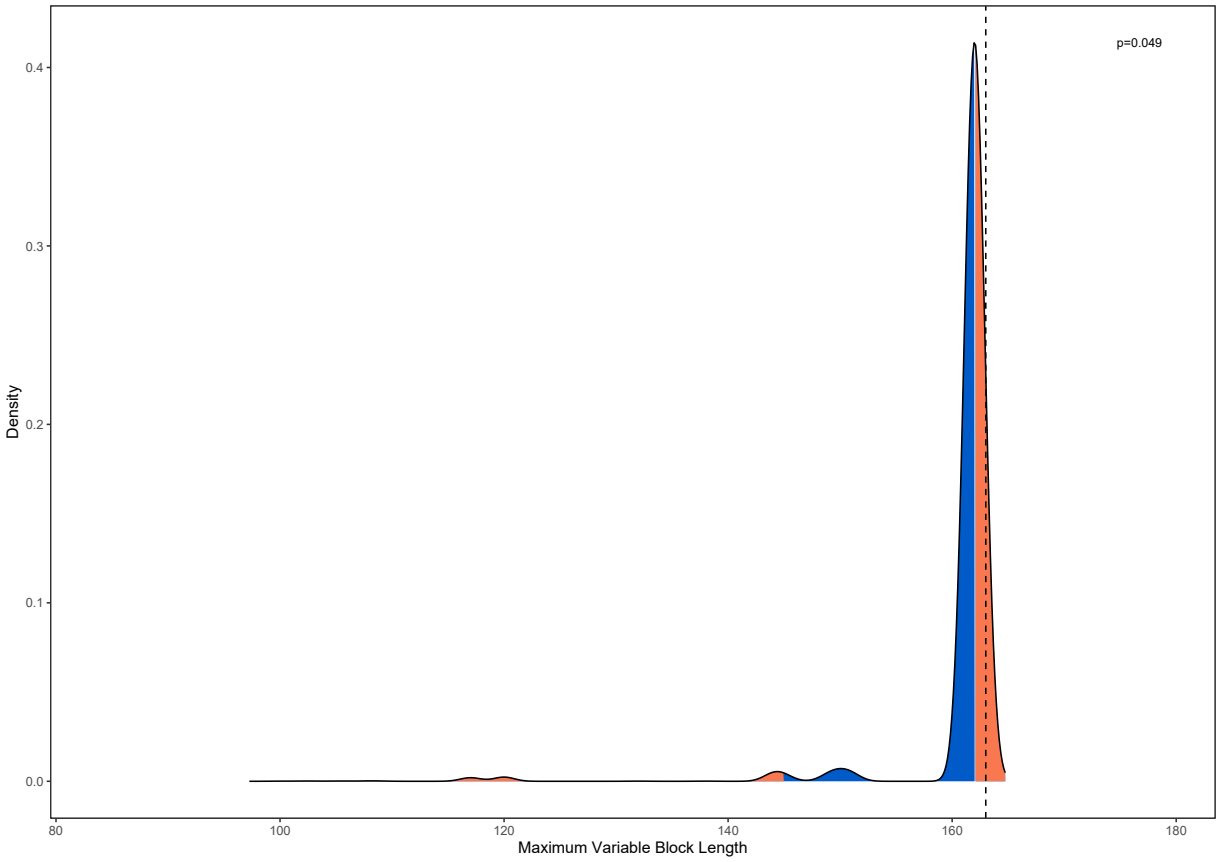

p=0.246
Density
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
Number of Invariant Sites Excluding Ambiguous

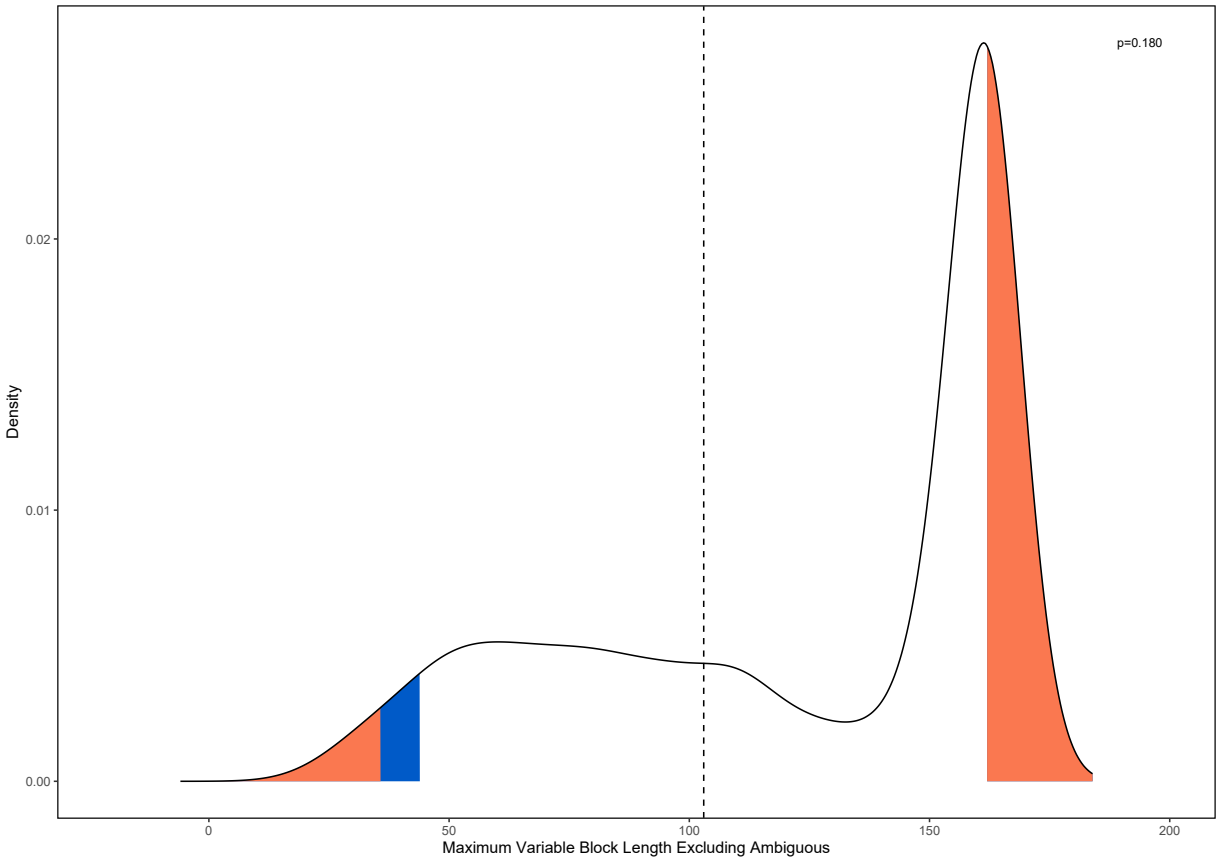

p=1.000
10.0
7.5
5.0
2.5
0.0
Density
0
1
2
Maximum Invariant Block Length

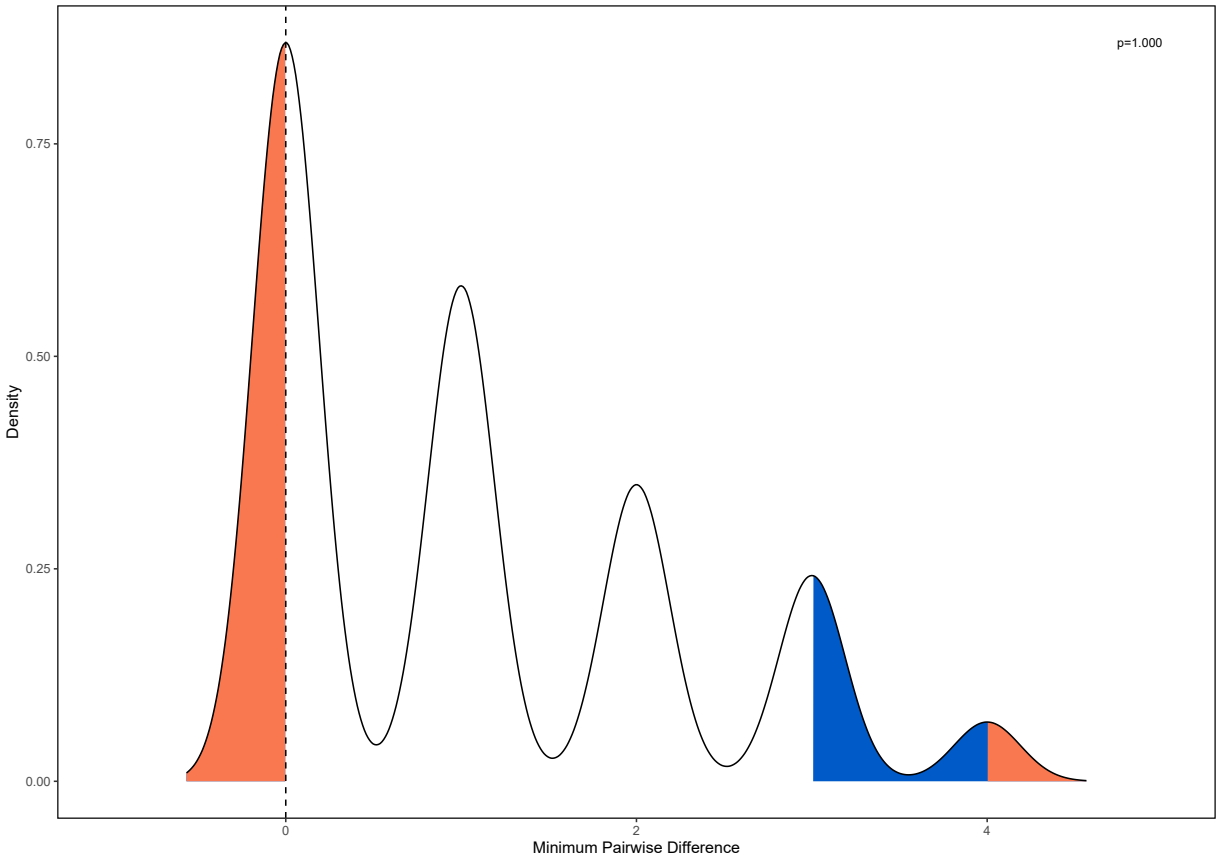

p=0.084
Density
1.5
1.0
0.5
0.0
0
2
4
Maximum Invariant Block Length Excluding Ambiguous

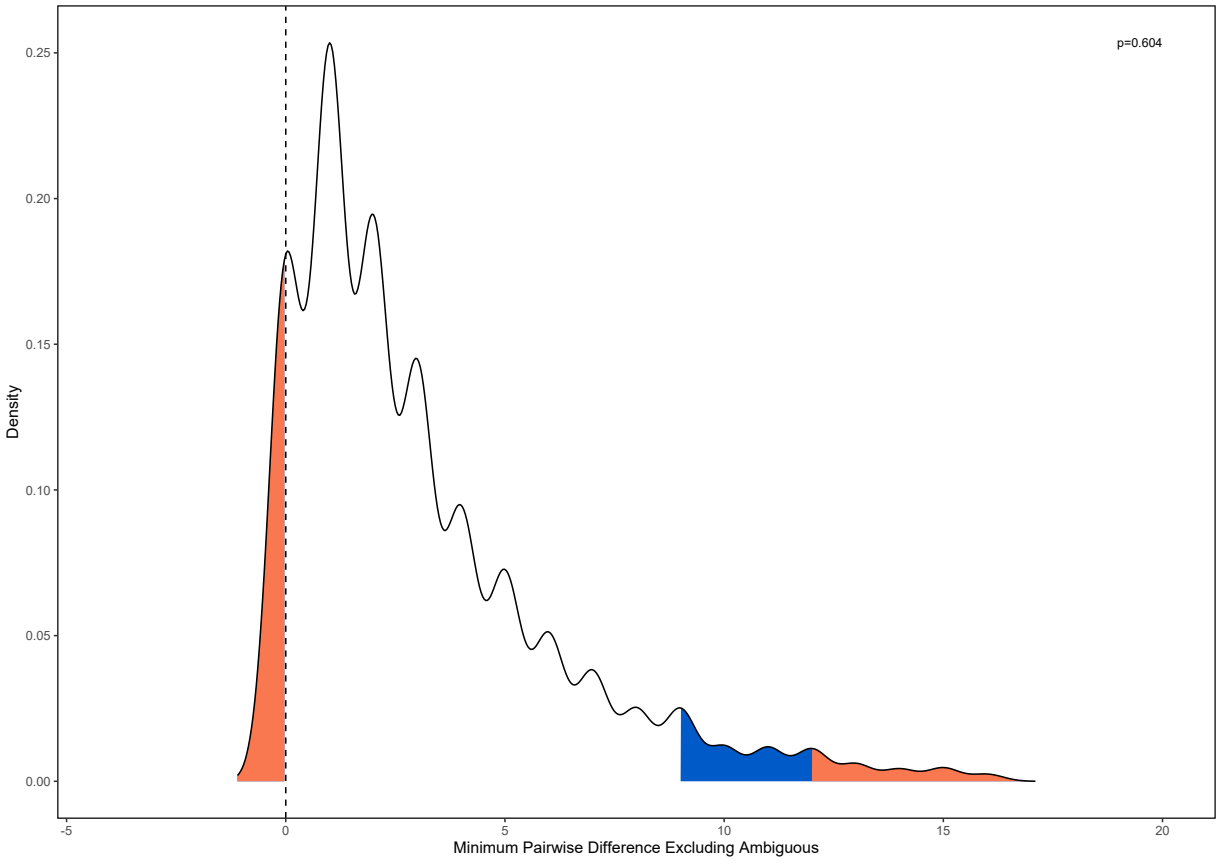

p=0.000
0.04
0.03
0.02
0.01
0.00
Density
40
60
80
100
120
Maximum Pairwise Difference

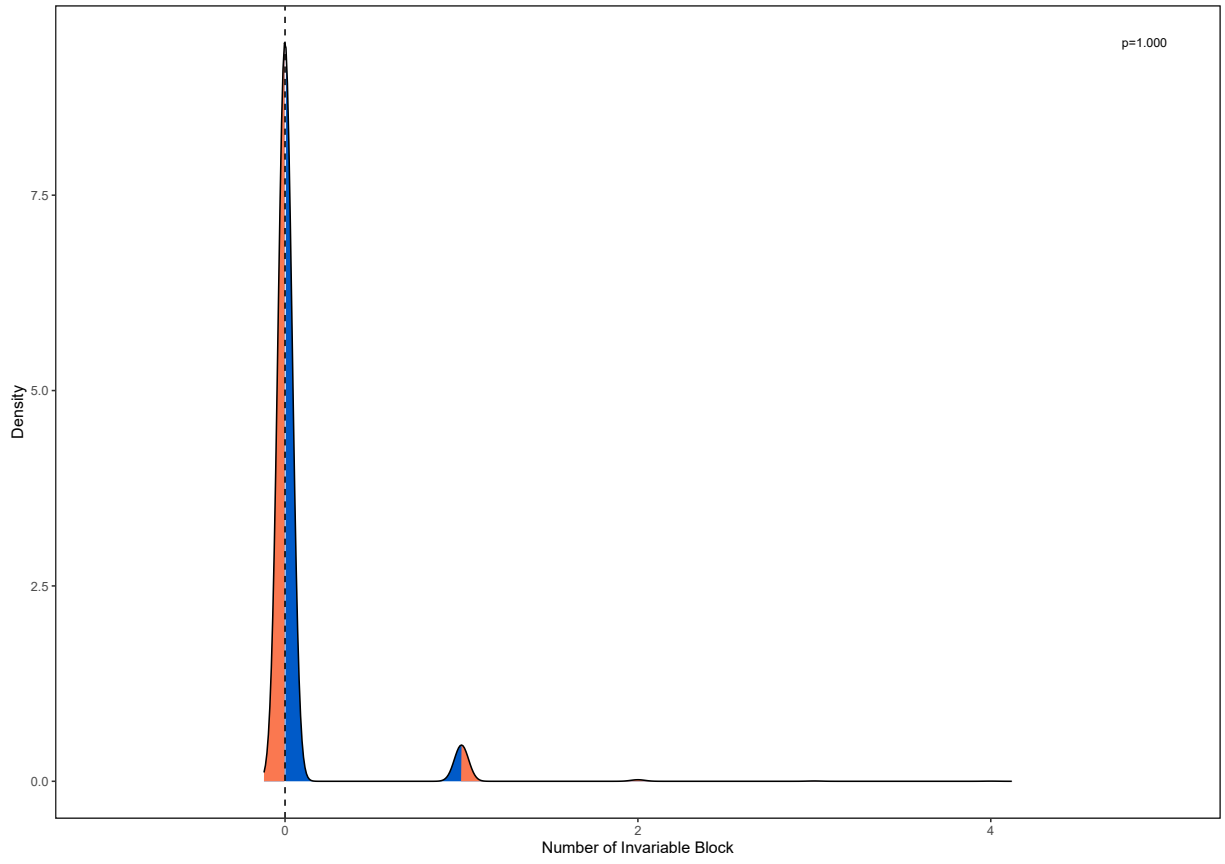

p=1.000
0.6
0.4
0.2
0.0
Density
155
160
Maximum Pairwise Difference Excluding Ambiguous

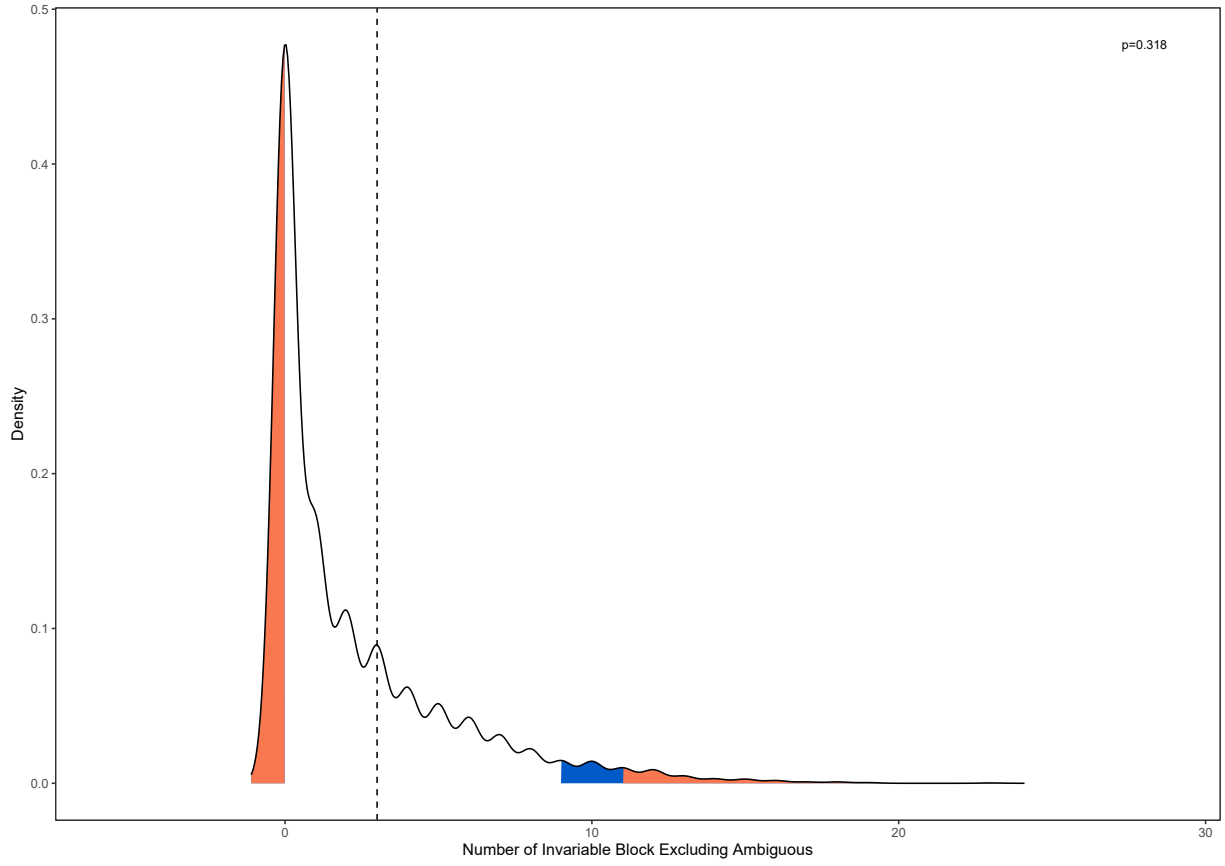

p=0.049
0.4
0.3
0.2
0.1
0.0
Density
80
100
120
140
160
180
Maximum Variable Block Length

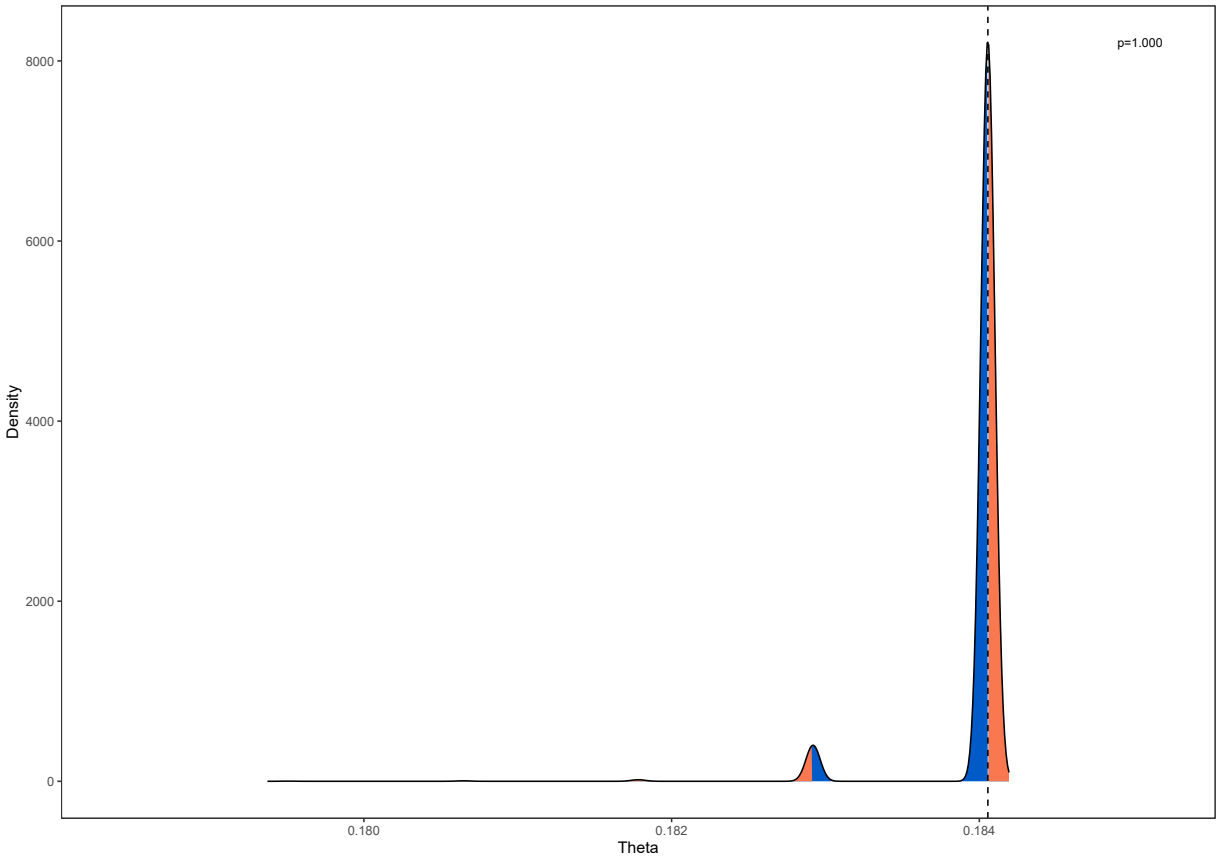

p=0.180
0.02
0.01
0.00
Density
0
50
100
150
200
Maximum Variable Block Length Excluding Ambiguous

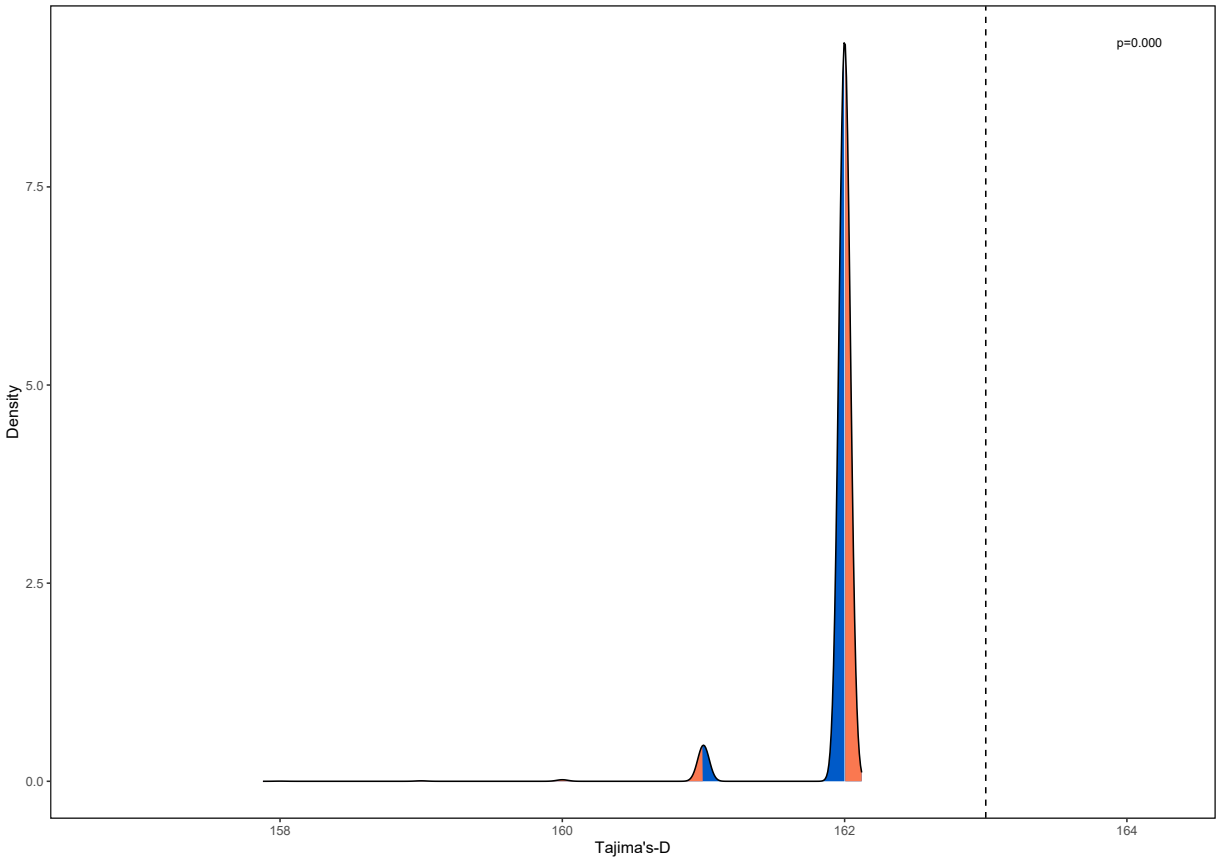

p=1.000
0.75
0.50
0.25
0.00
Density
0
2
4
Minimum Pairwise Difference

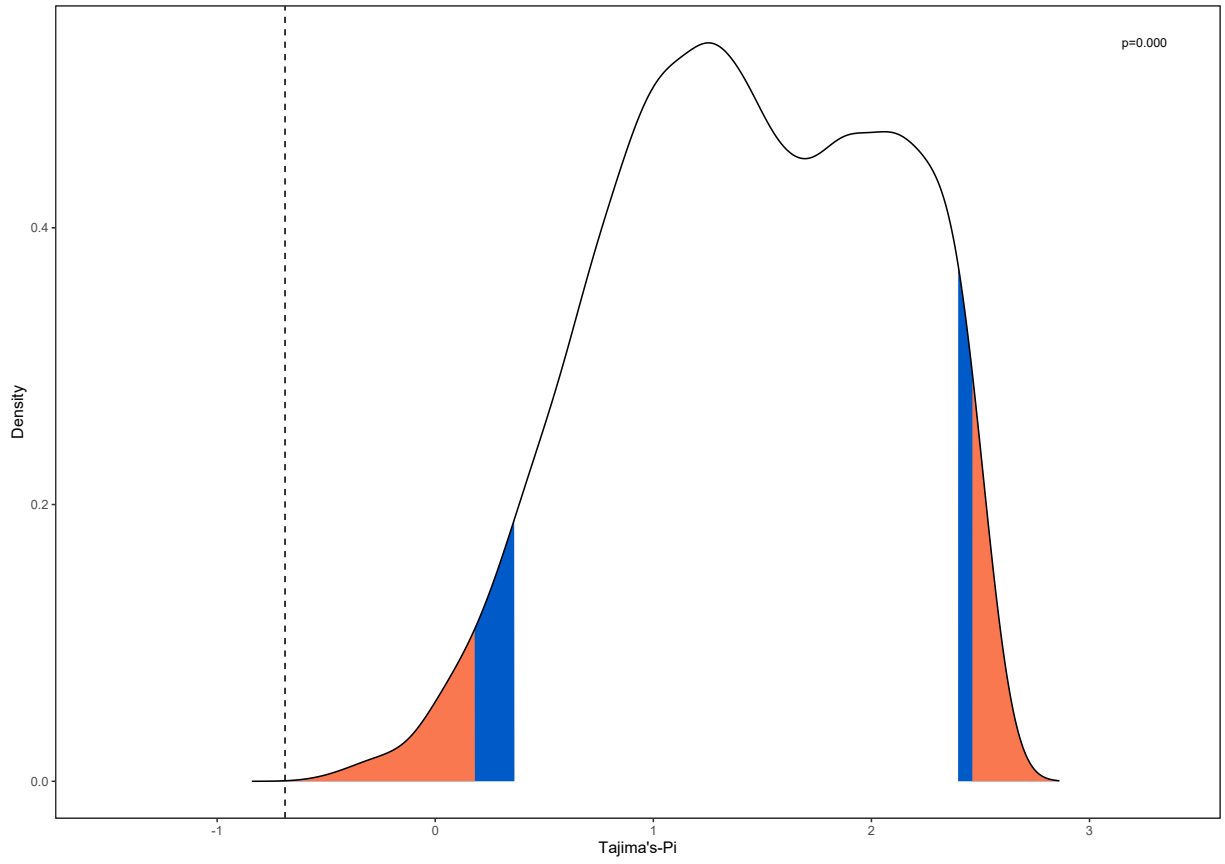

p=0.604
0.25
0.20
0.15
0.10
0.05
0.00
Density
-5
0
5
10
15
20
Minimum Pairwise Difference Excluding Ambiguous

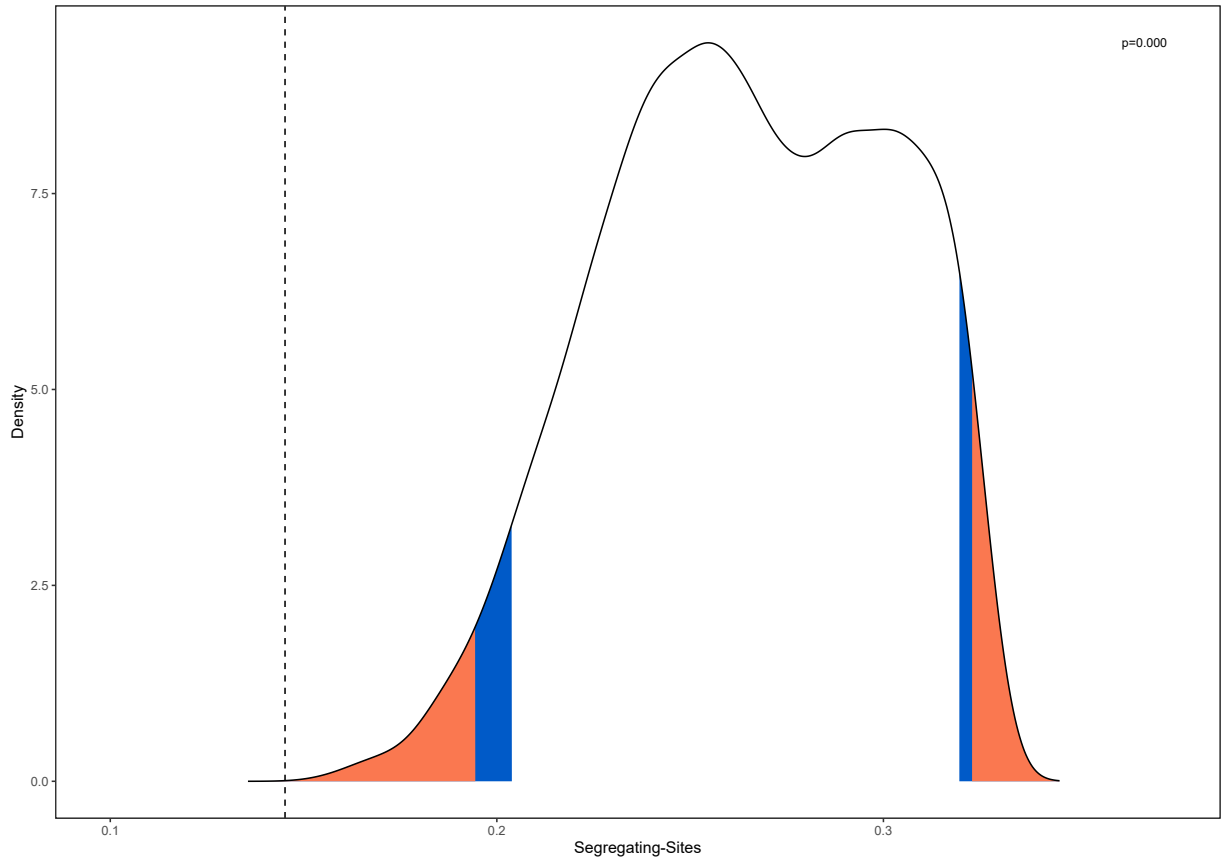

p=1.000
7.5
5.0
2.5
0.0
Density
0
2
4
Number of Invariable Block

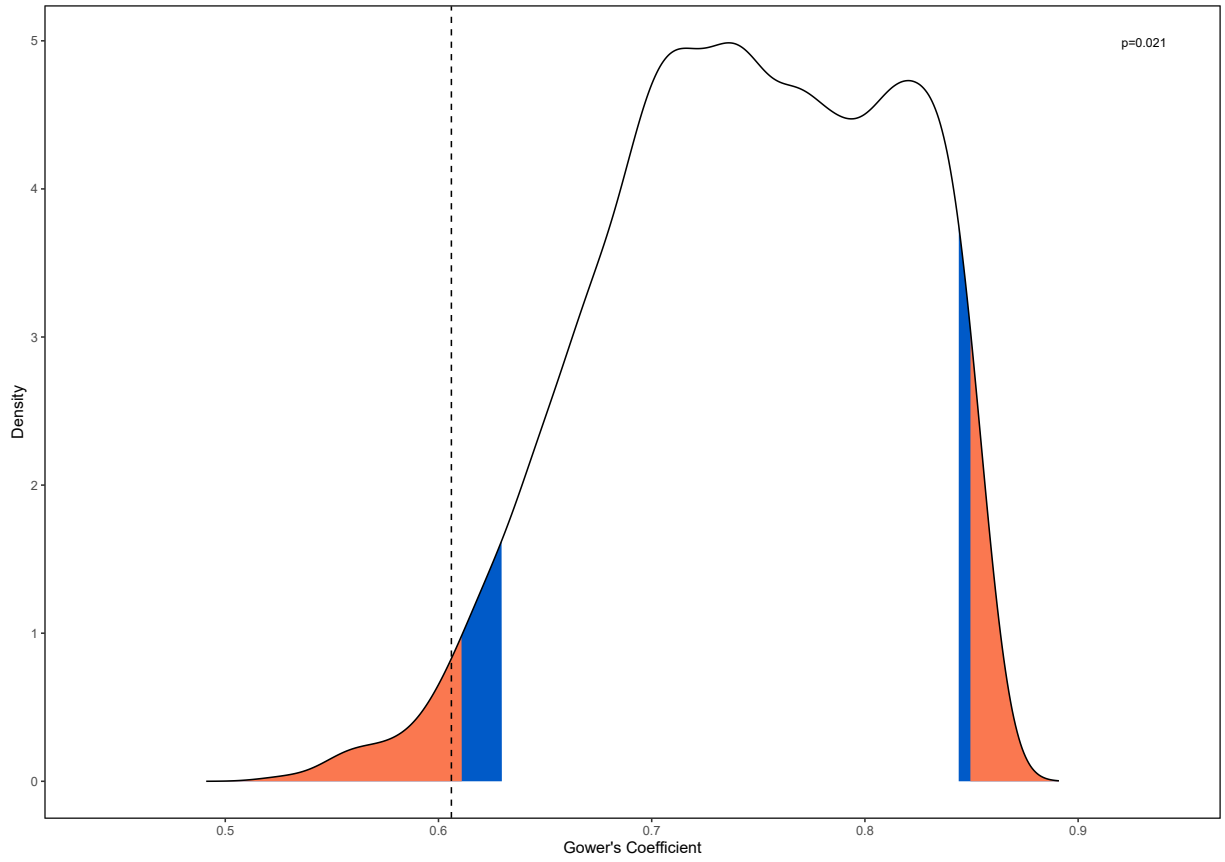

p=0.318
Density
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
Number of Invariable Block Excluding Ambiguous

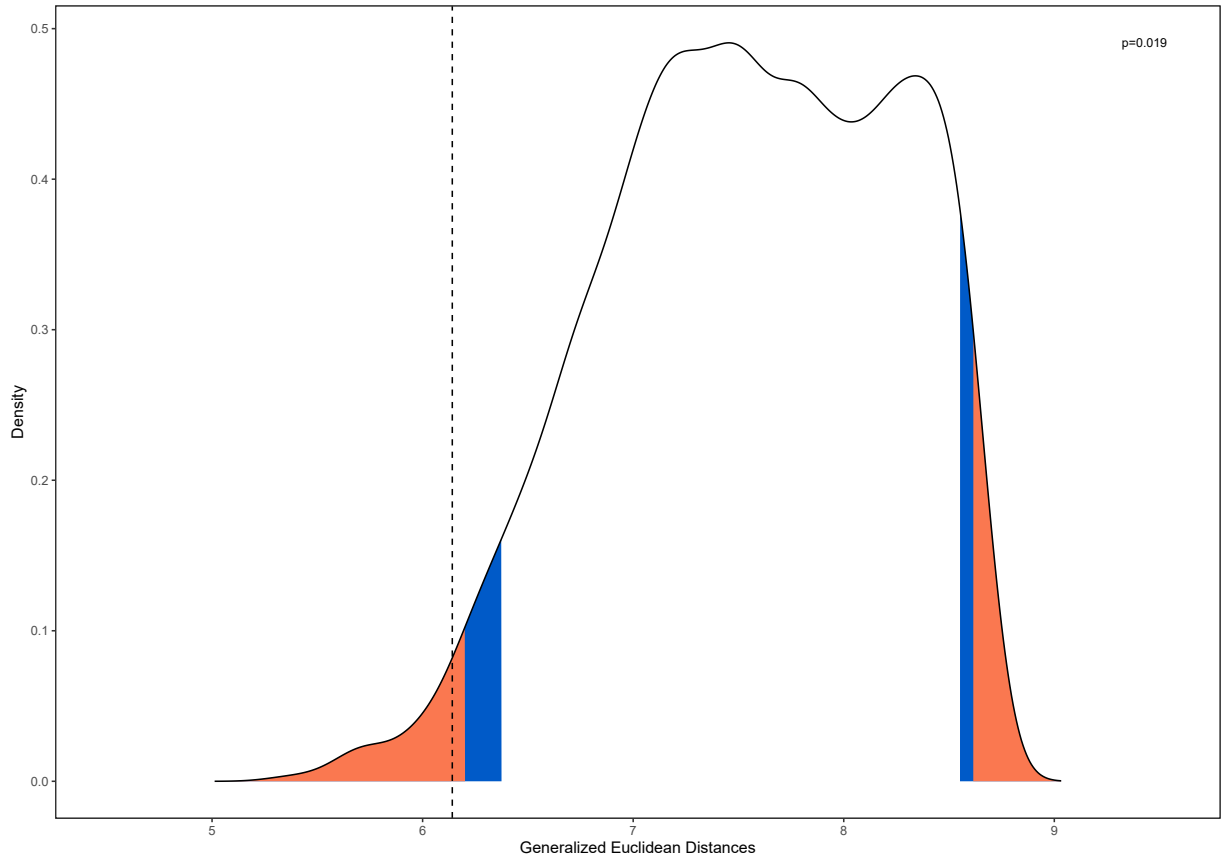

p=1.000
8000
6000
4000
2000
0
Density
0.180
0.182
0.184
Theta


p=0.000
7.5
5.0
2.5
0.0
Density
158
160
162
164
Tajima's-D

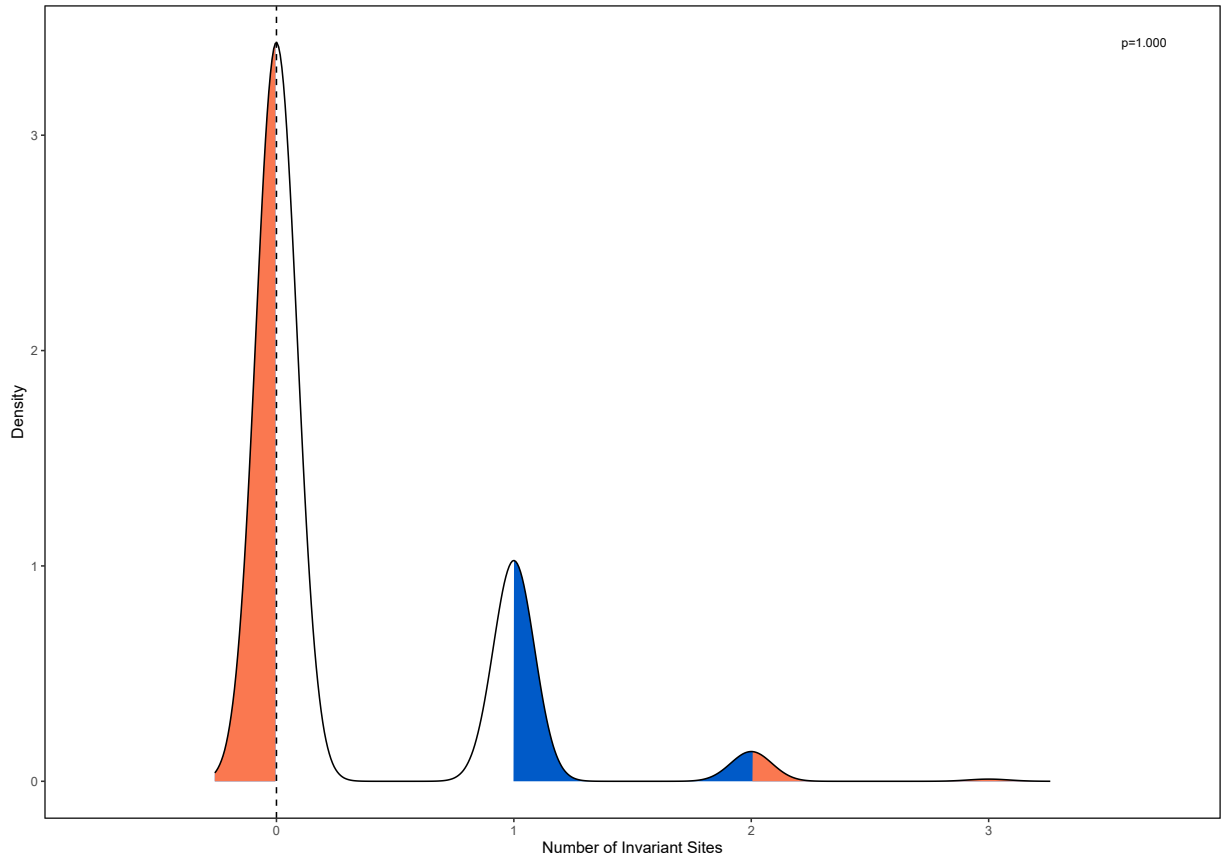

p=0.000
0.4
0.2
0.0
Density
-1
0
1
2
3
Tajima's-Pi

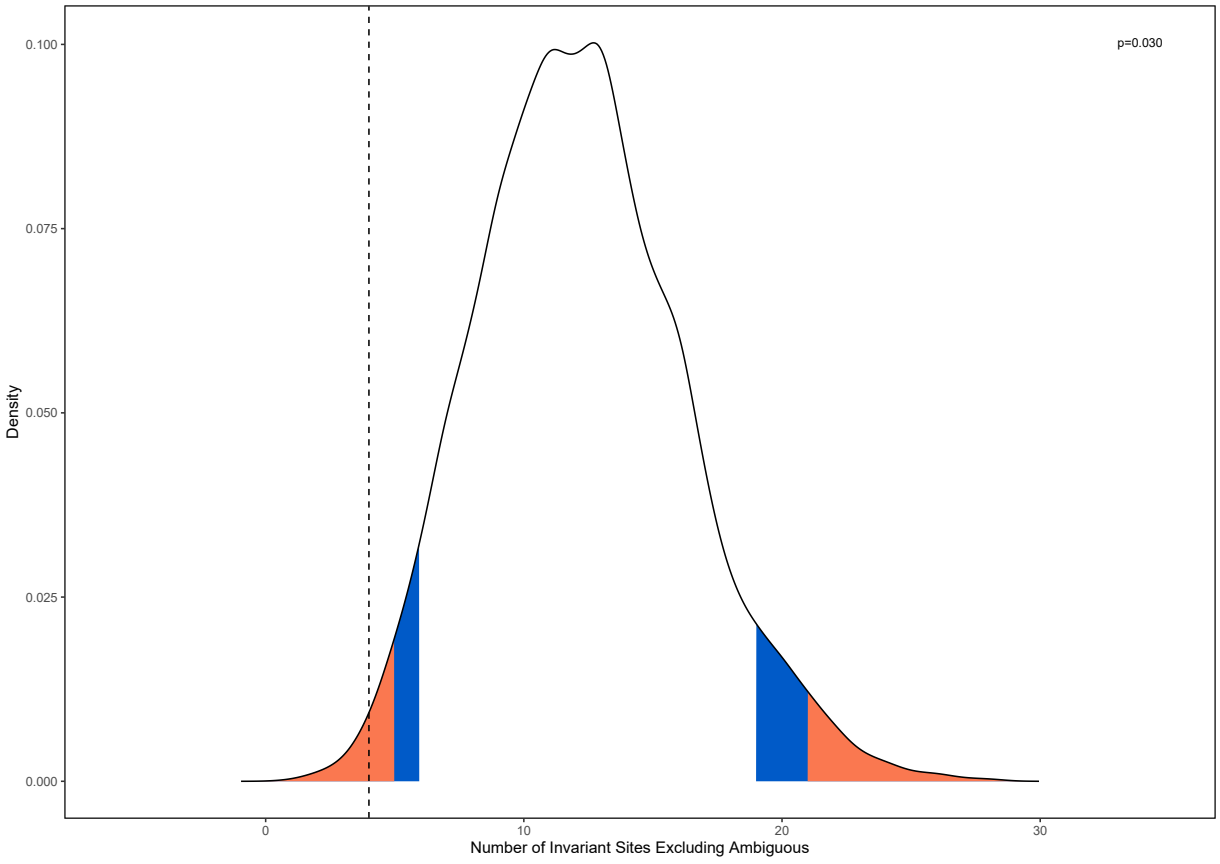

p=0.000
7.5
5.0
2.5
0.0
Density
0.1
0.2
0.3
Segregating-Sites

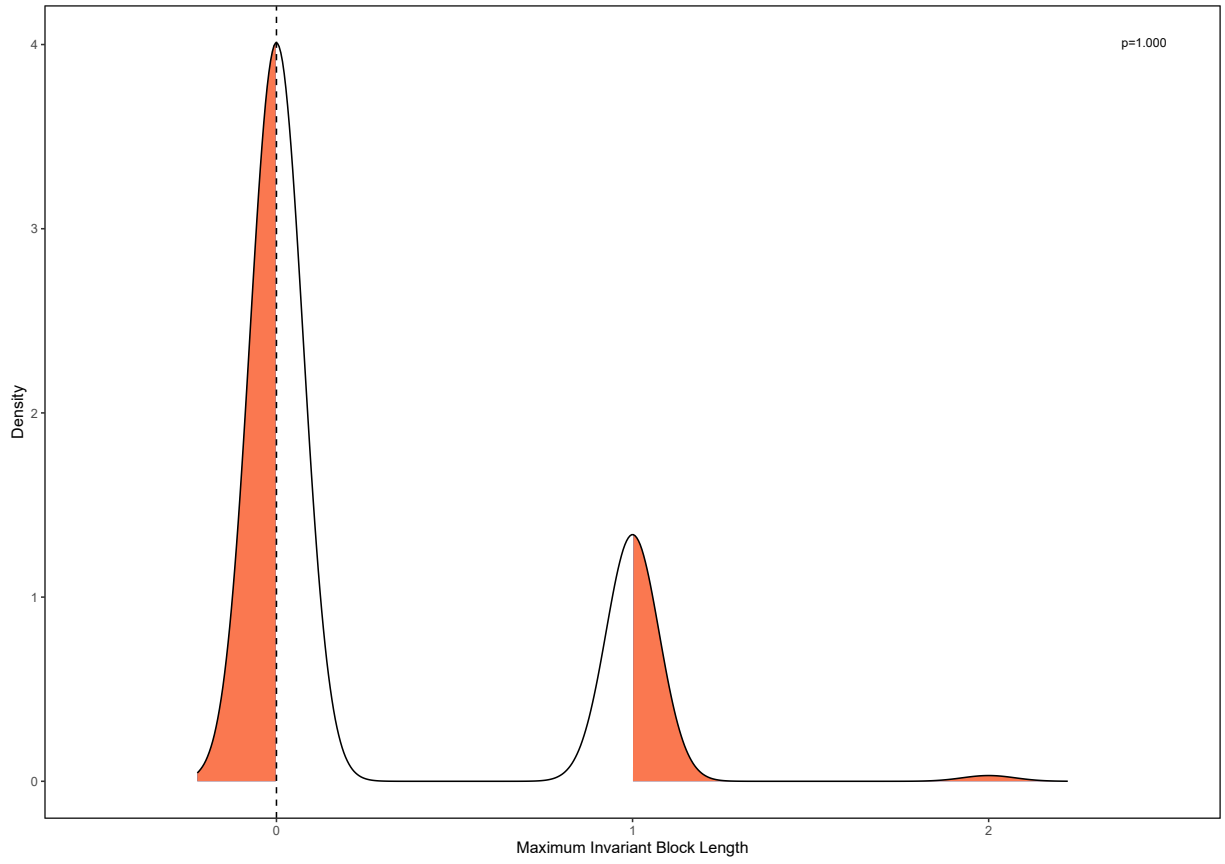

p=0.021
Density
Gower's Coefficient
0
1
2
3
4
5
0.5
0.6
0.7
0.8
0.9

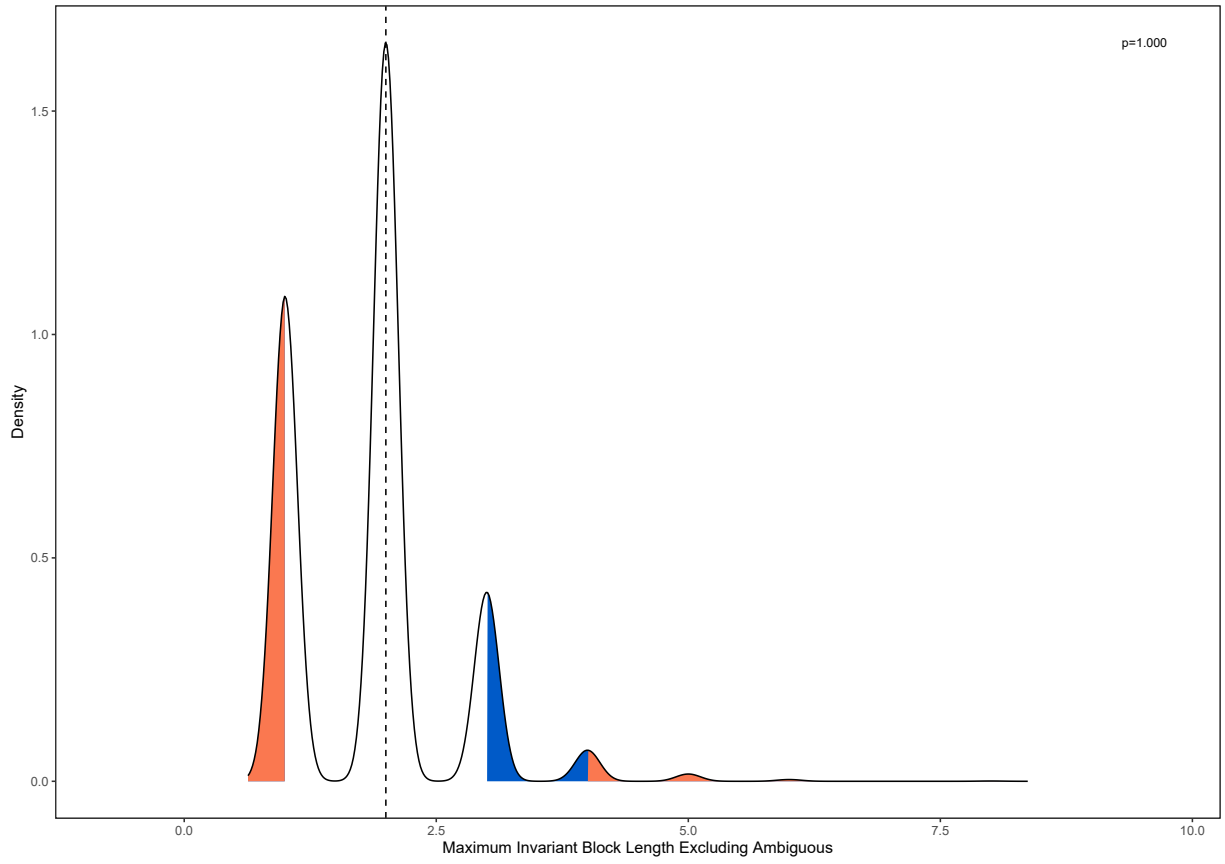

p=0.019
Density
0.0
0.1
0.2
0.3
0.4
0.5
5
6
7
8
9
Generalized Euclidean Distances

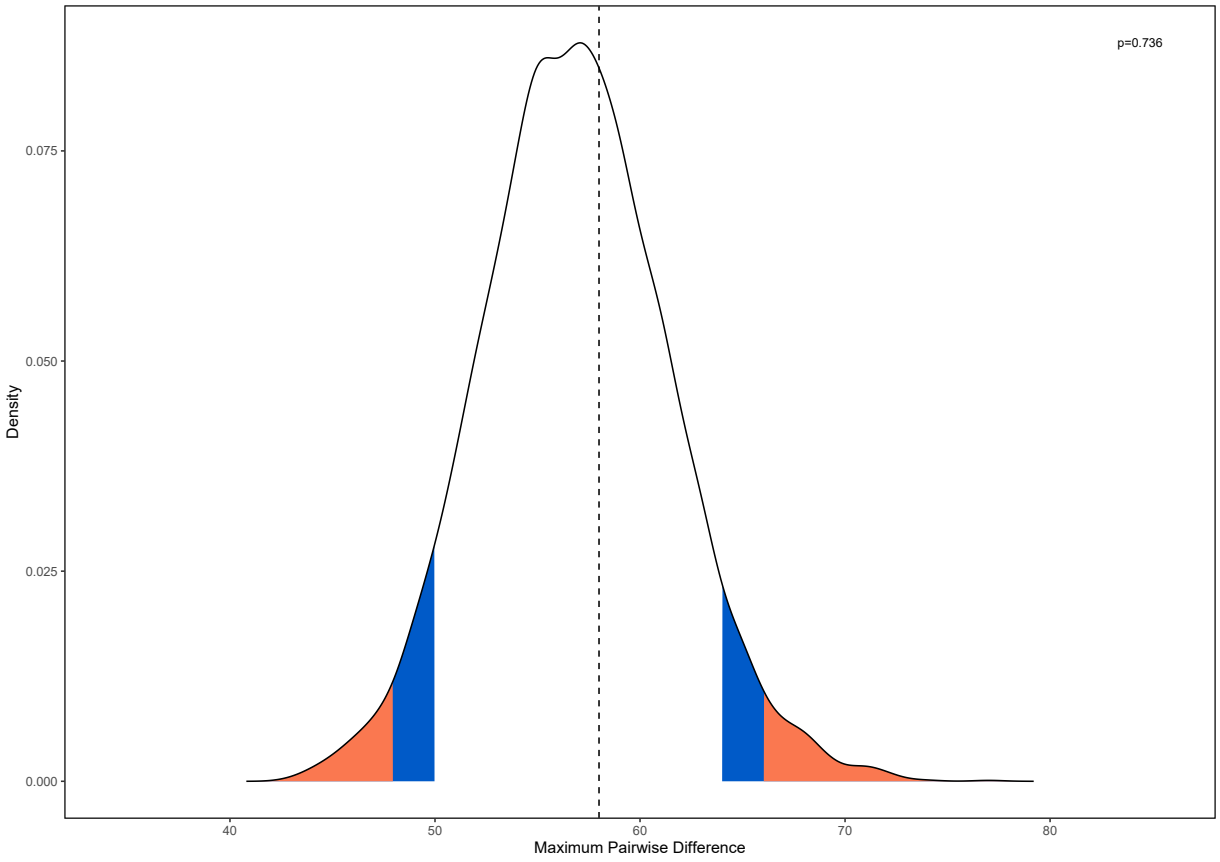
For SHDM model with 3 rate categories, the density graphs for all the summary statistics are as follows.

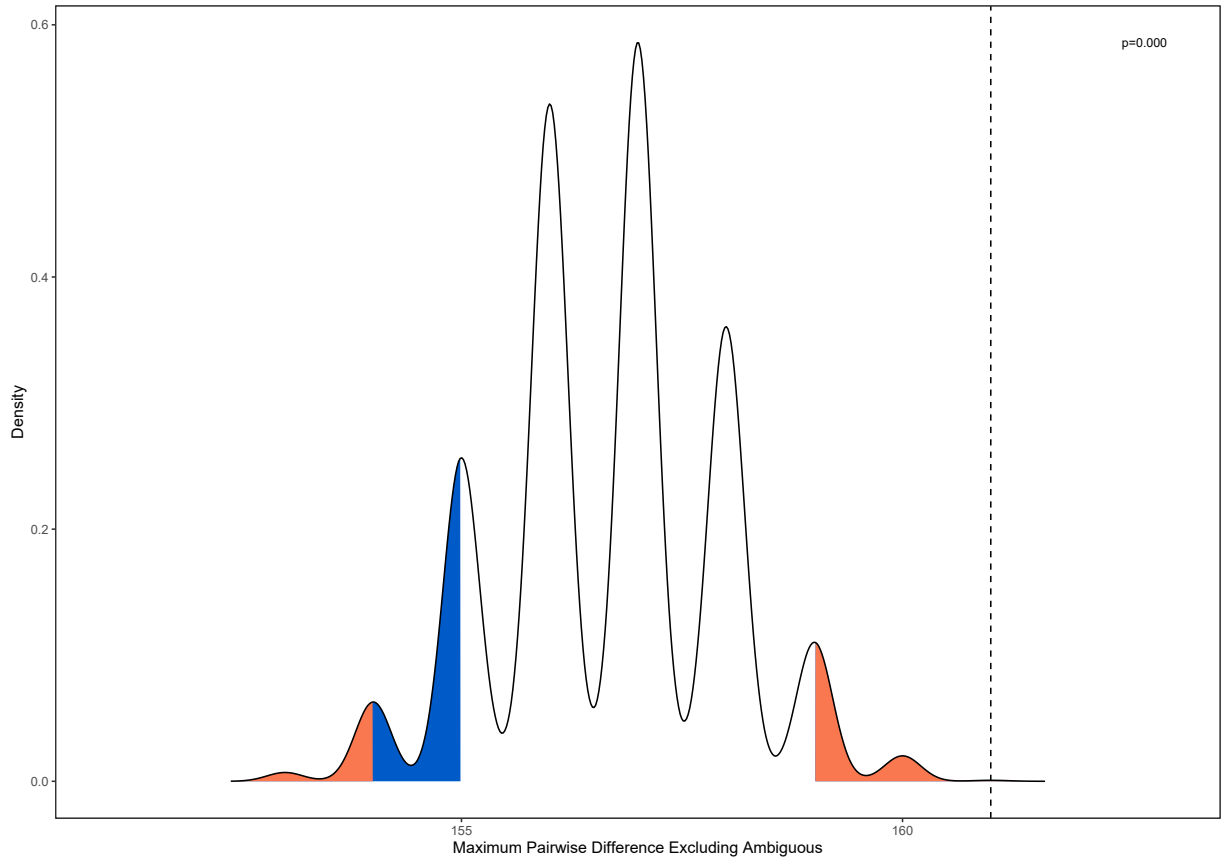

p=1.000
Density
0
1
2
3
0
1
2
3
Number of Invariant Sites

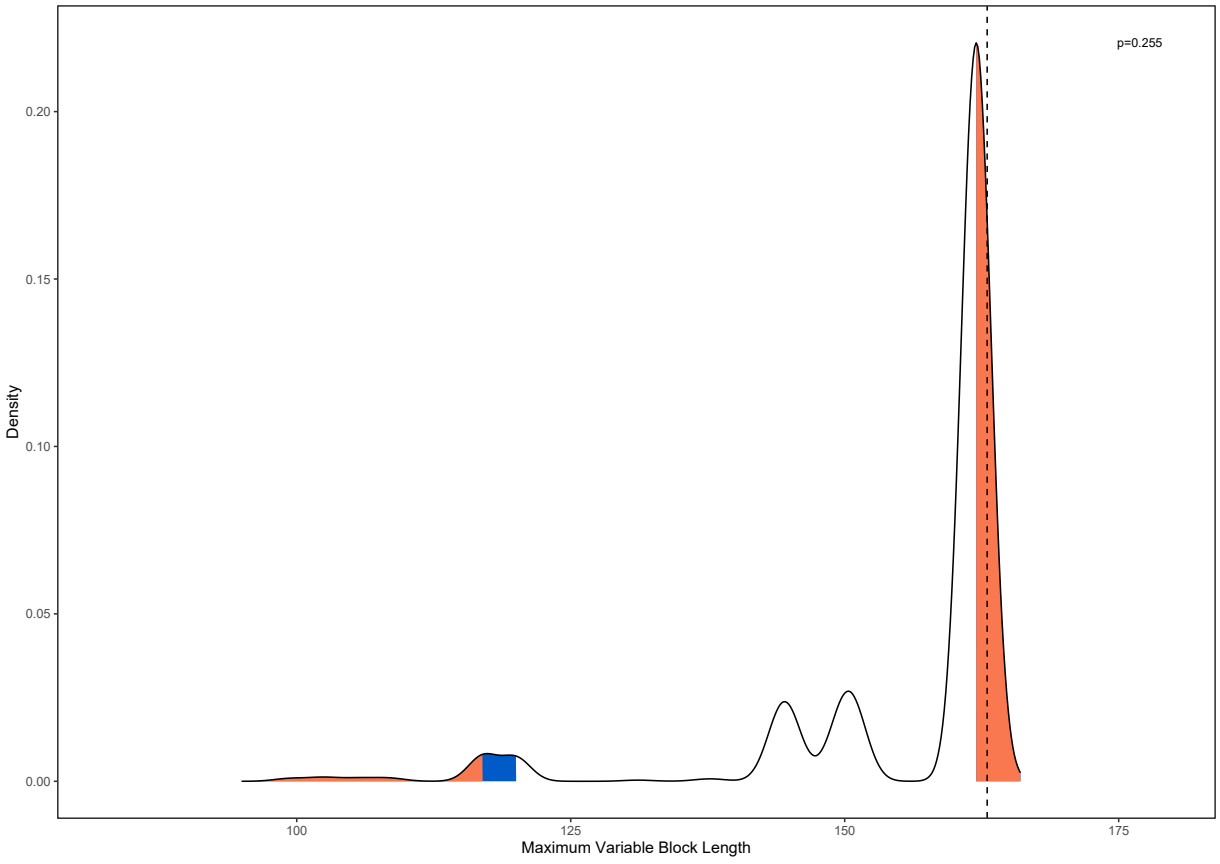

p=0.030
0.100
0.075
0.050
0.025
0.000
Density
0
10
20
30
Number of Invariant Sites Excluding Ambiguous

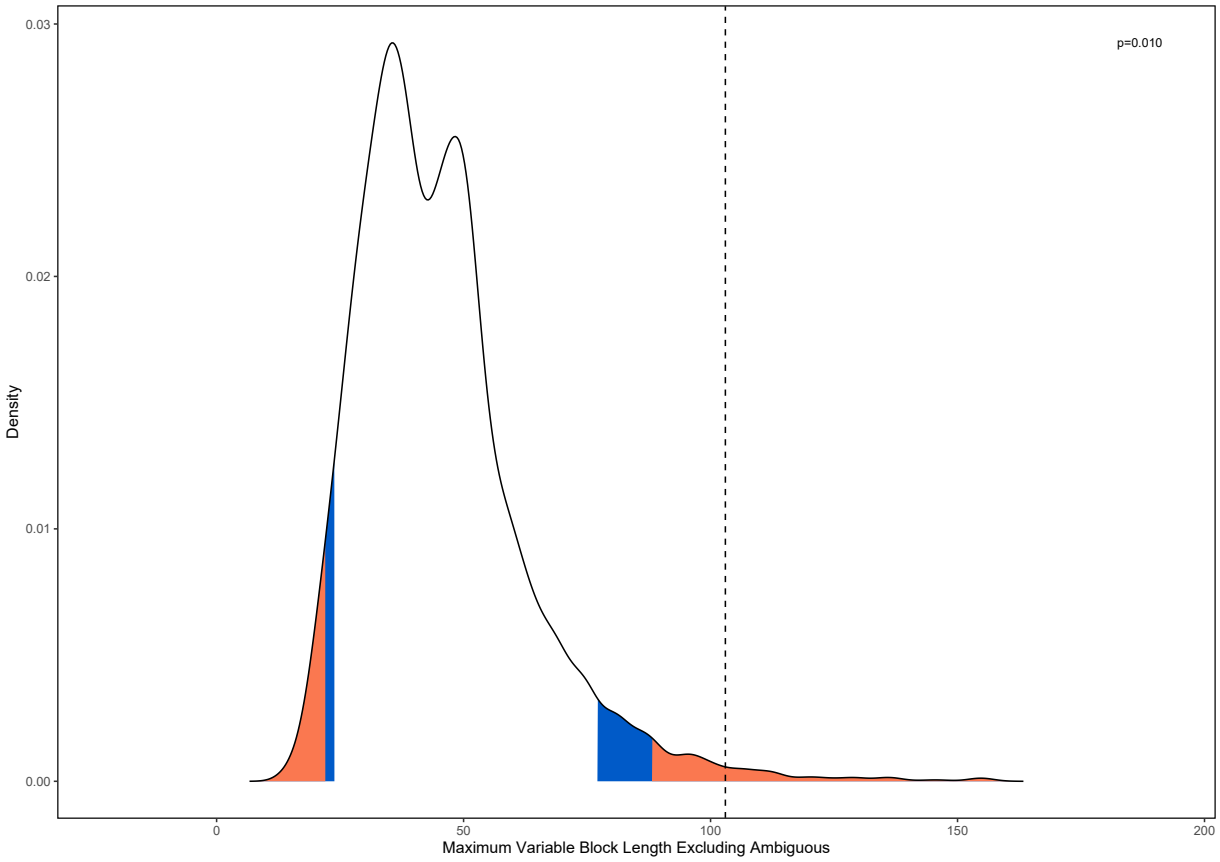

p=1.000
Density
4
3
2
1
0
0
1
2
Maximum Invariant Block Length

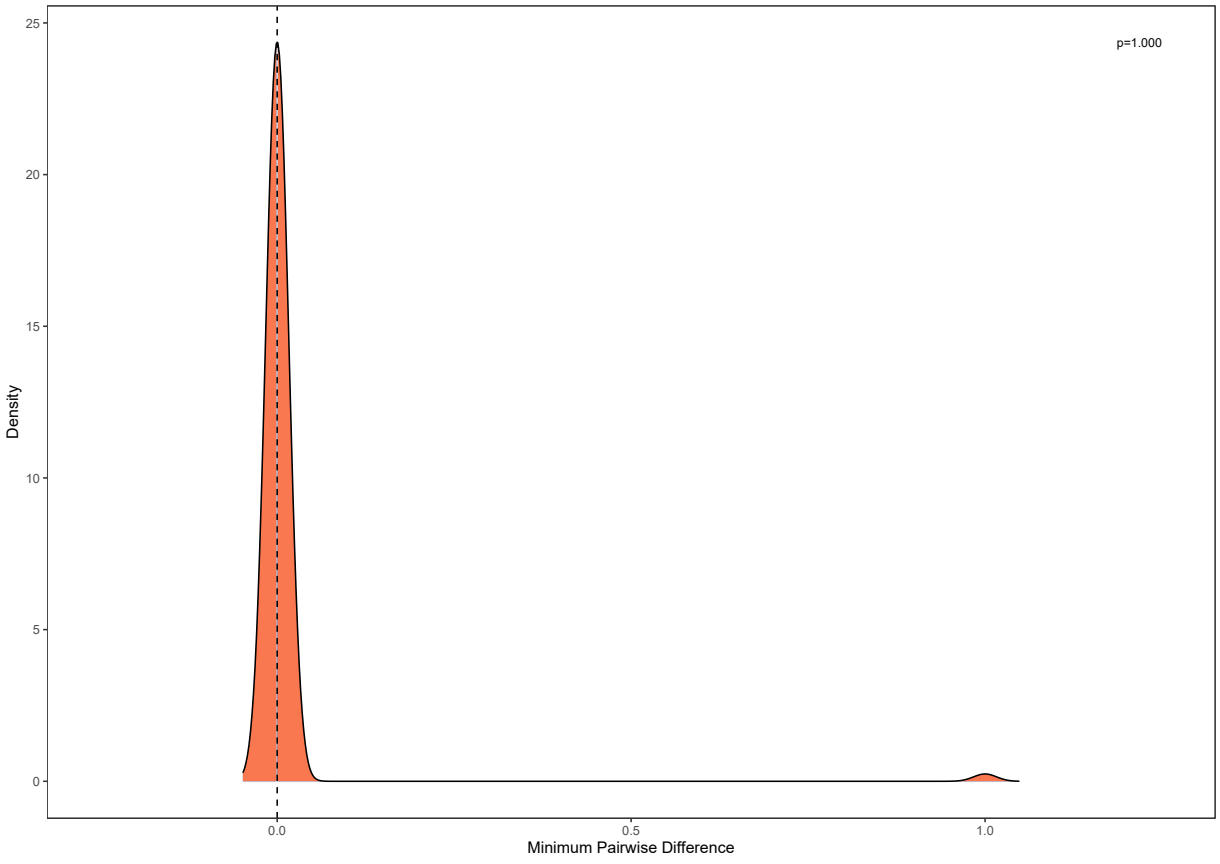

p=1.000
Density
1.5
1.0
0.5
0.0
0.0
2.5
5.0
7.5
10.0
Maximum Invariant Block Length Excluding Ambiguous

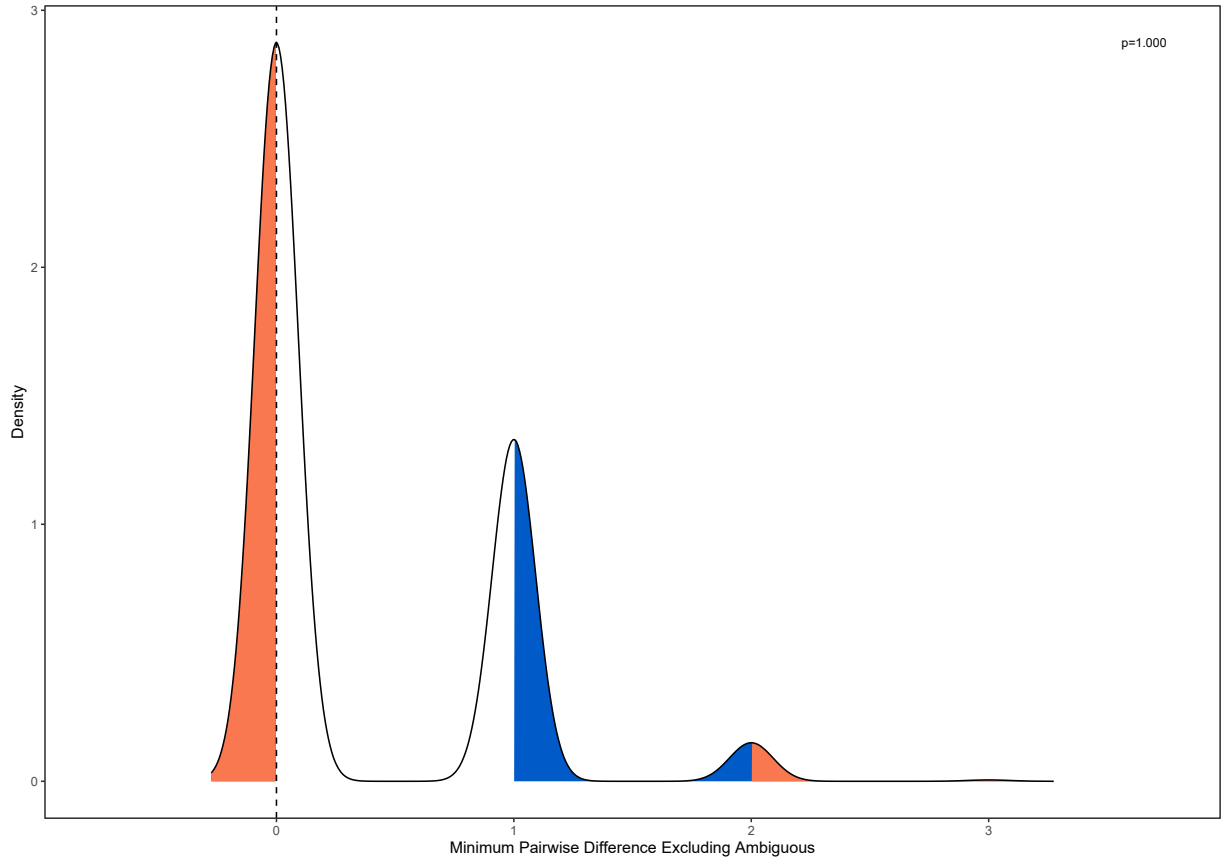

p=0.736
0.075
0.050
0.025
0.000
Density
40
50
60
70
80
Maximum Pairwise Difference

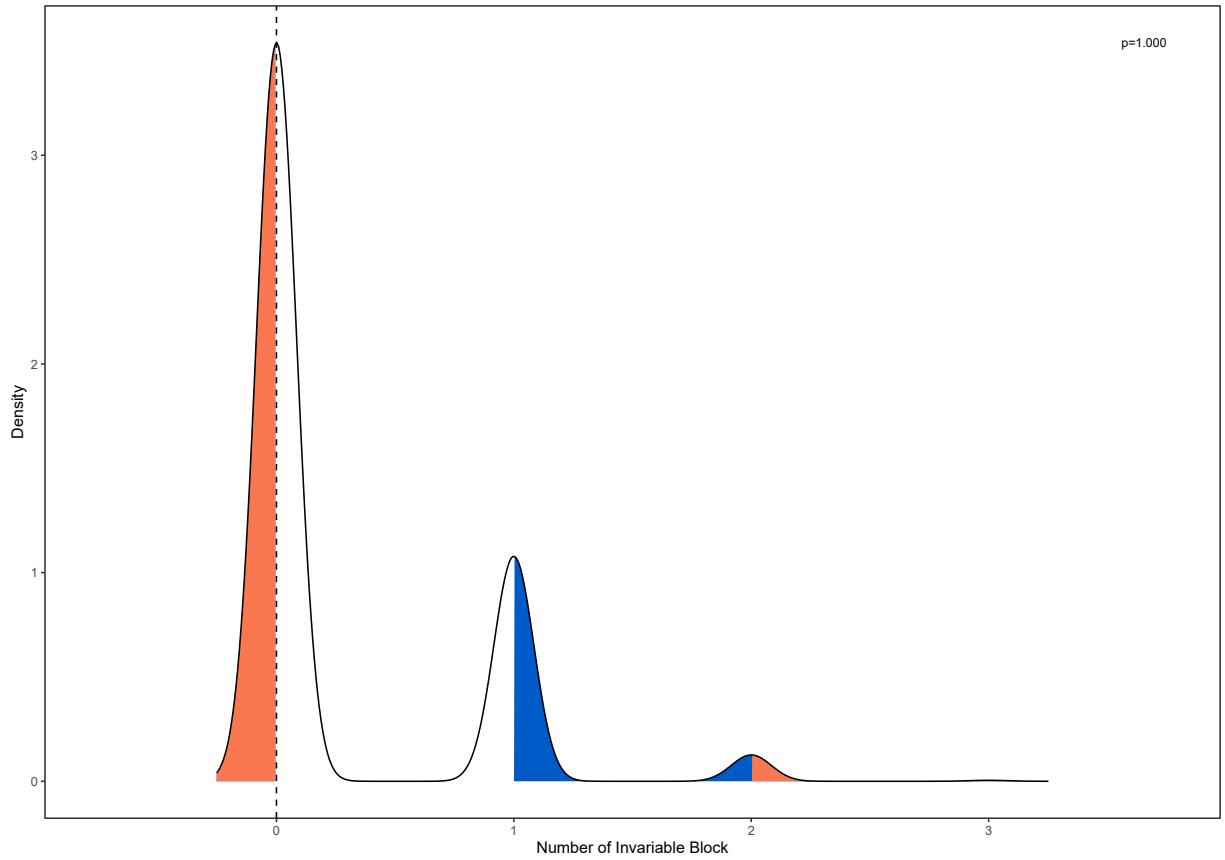

p=0.000
Density
0.6
0.4
0.2
0.0
155
160
Maximum Pairwise Difference Excluding Ambiguous

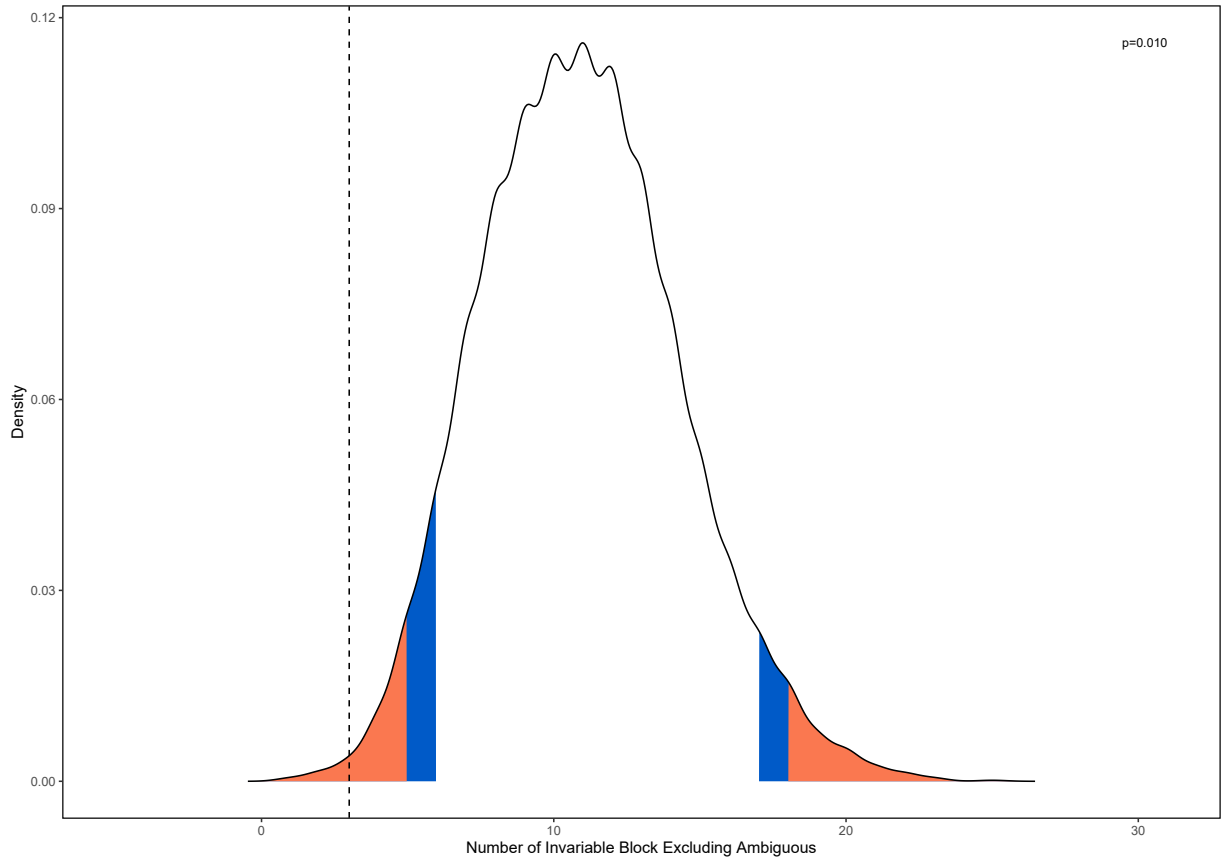

p=0.255
0.20
0.15
0.10
0.05
0.00
Density
100
125
150
175
Maximum Variable Block Length

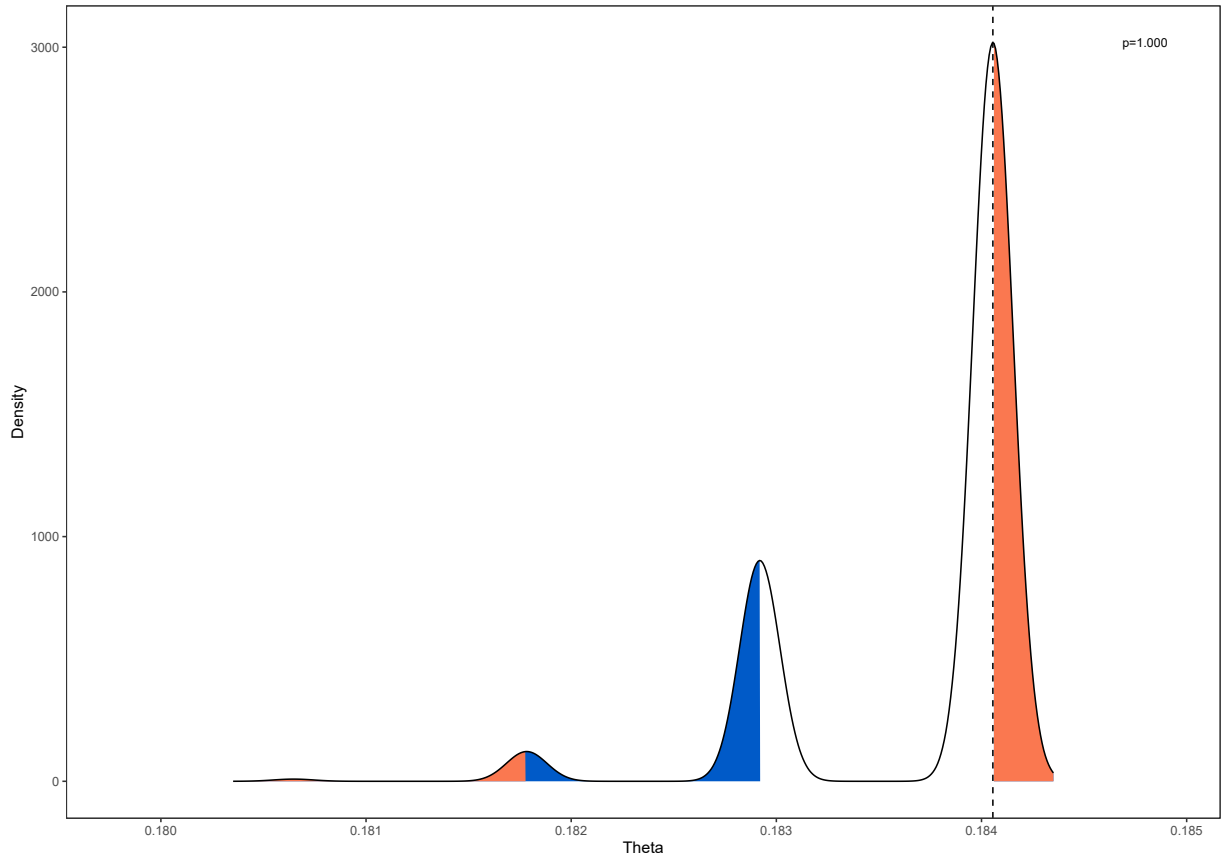

p=0.010
0.03
0.02
0.01
0.00
Density
0
50
100
150
200
Maximum Variable Block Length Excluding Ambiguous

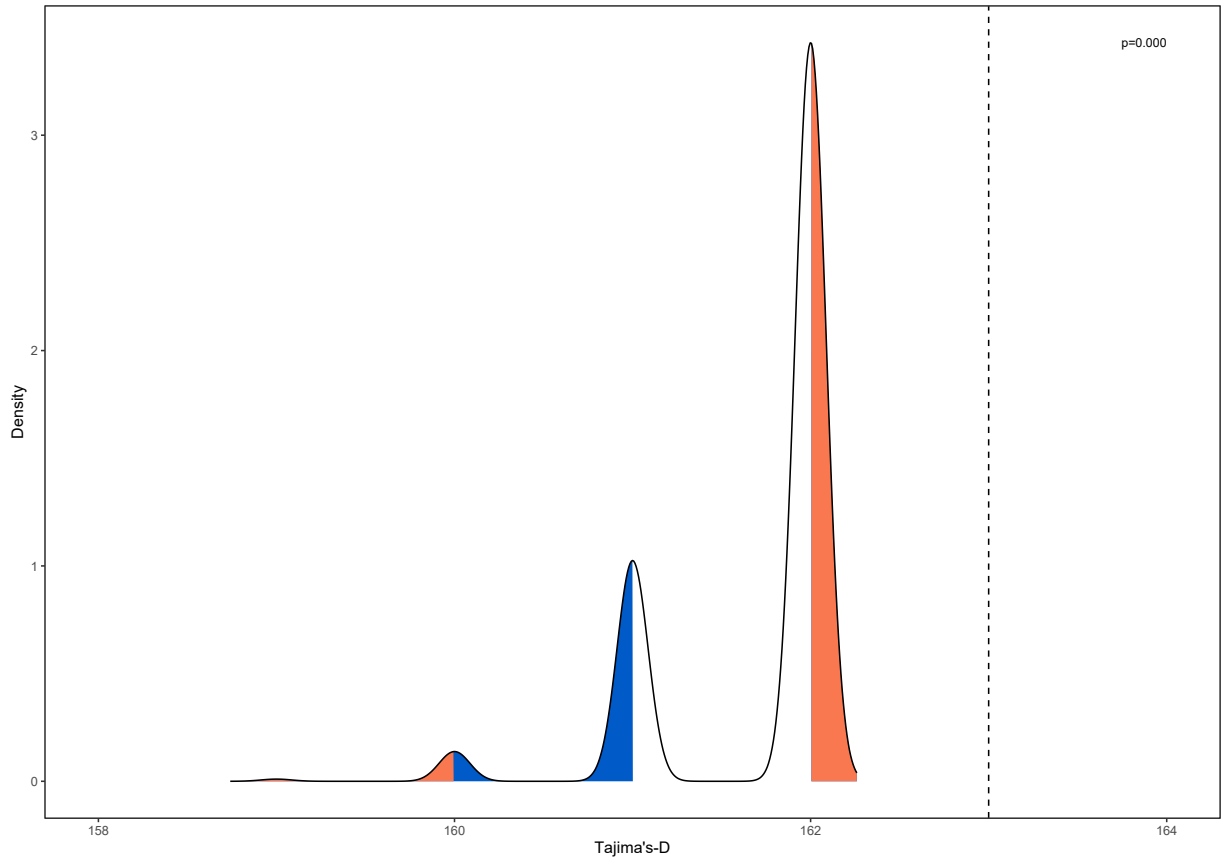

p=1.000
Density
25
20
15
10
5
0
0.0
0.5
1.0
Minimum Pairwise Difference

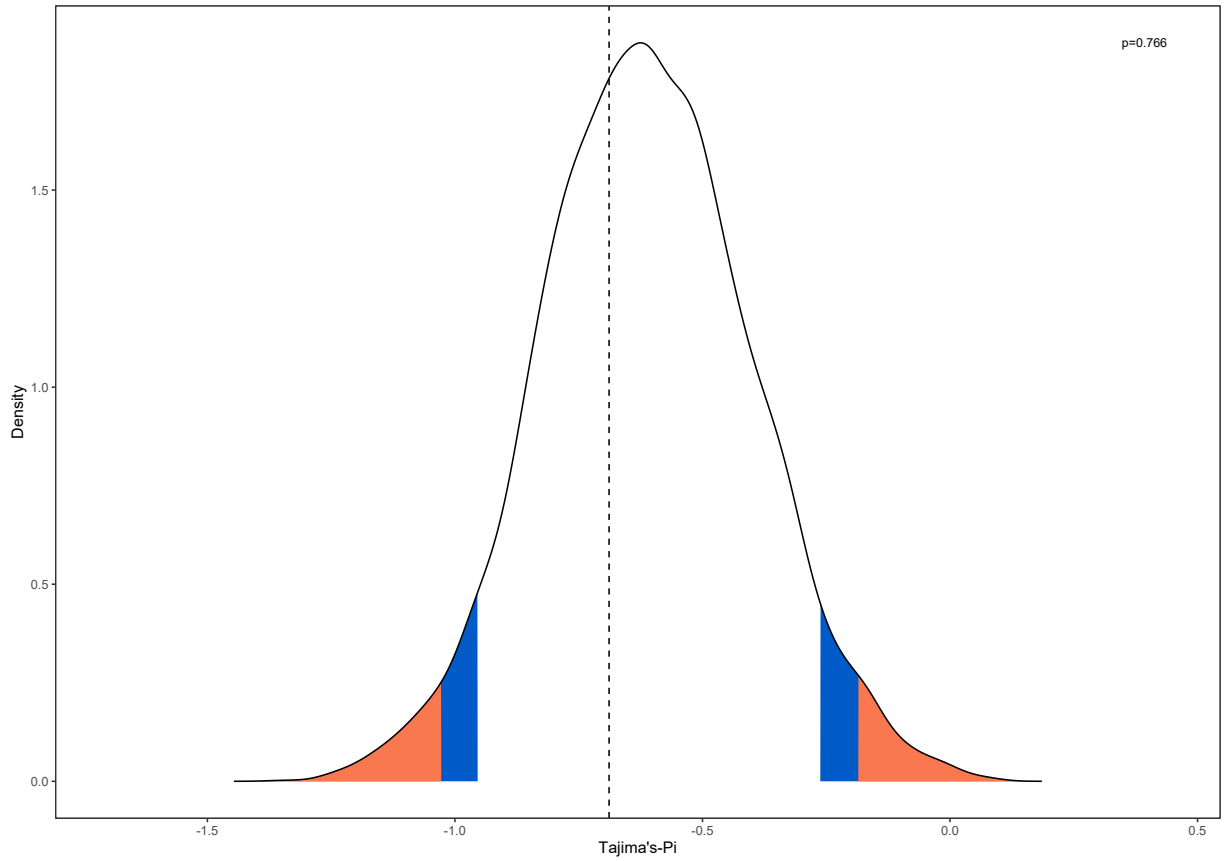

p=1.000
Density
Minimum Pairwise Difference Excluding Ambiguous
0
1
2
3

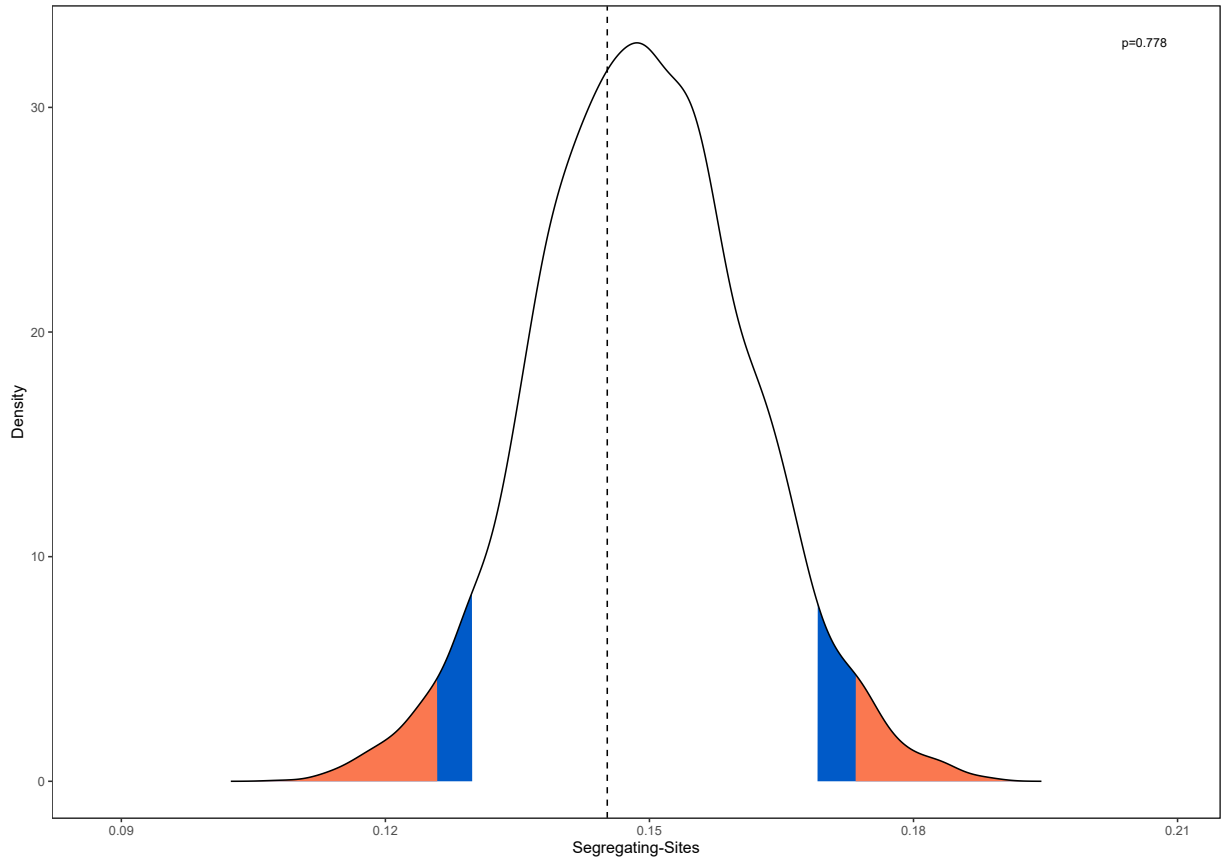

p=1.000
Density
0
1
2
3
0
1
2
3
Number of Invariable Block

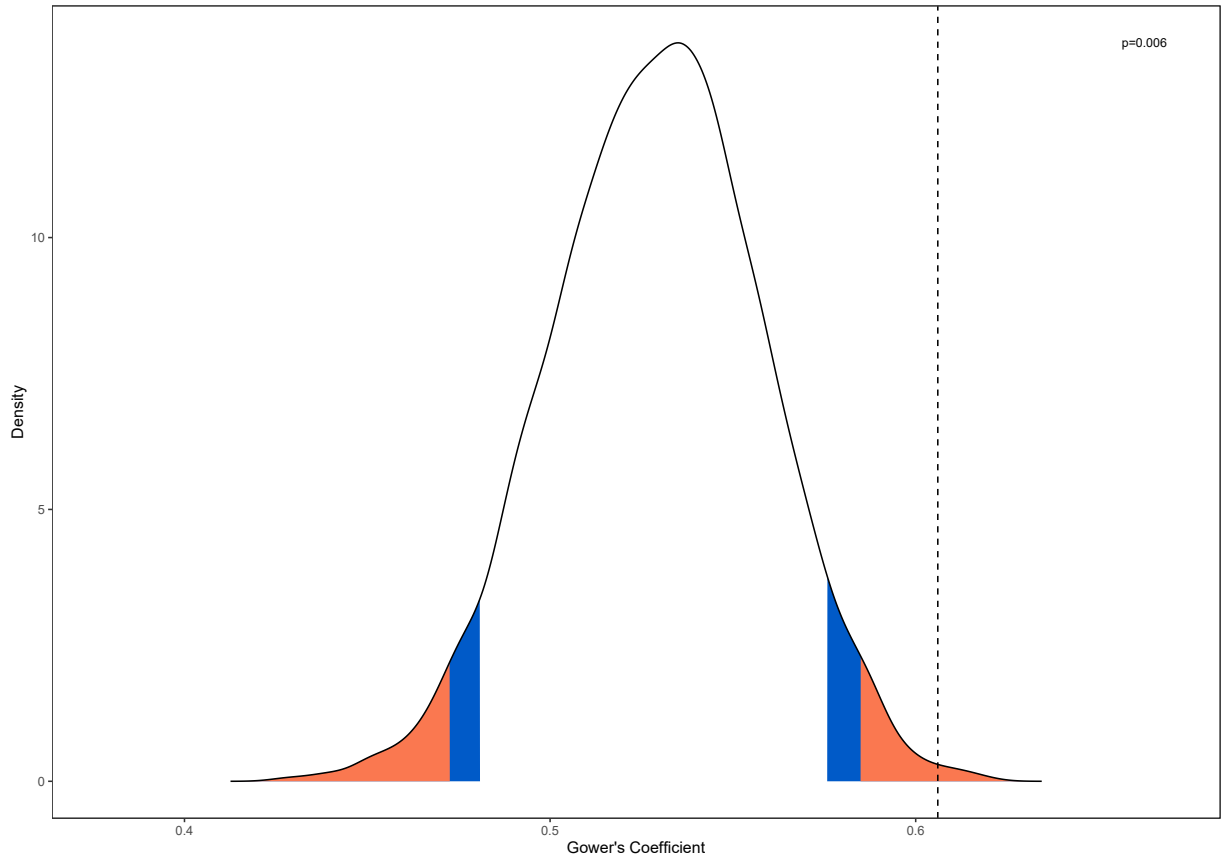

p=0.010
Density
0.12
0.09
0.06
0.03
0.00
0
10
20
30
Number of Invariable Block Excluding Ambiguous

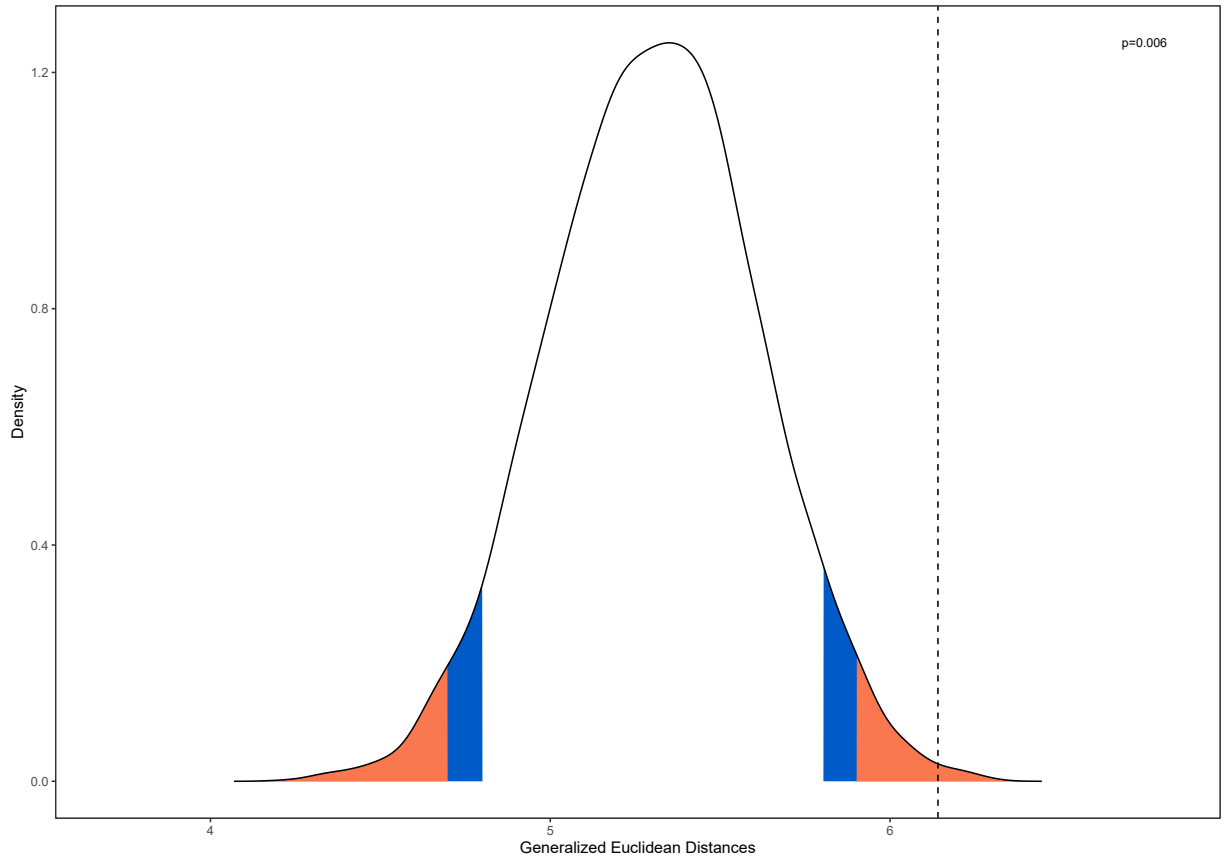

p=1.000
3000
2000
1000
0
Density
0.180
0.181
0.182
0.183
0.184
0.185
Theta


p=0.000
Density
0
1
2
3
158
160
162
164
Tajima's-D

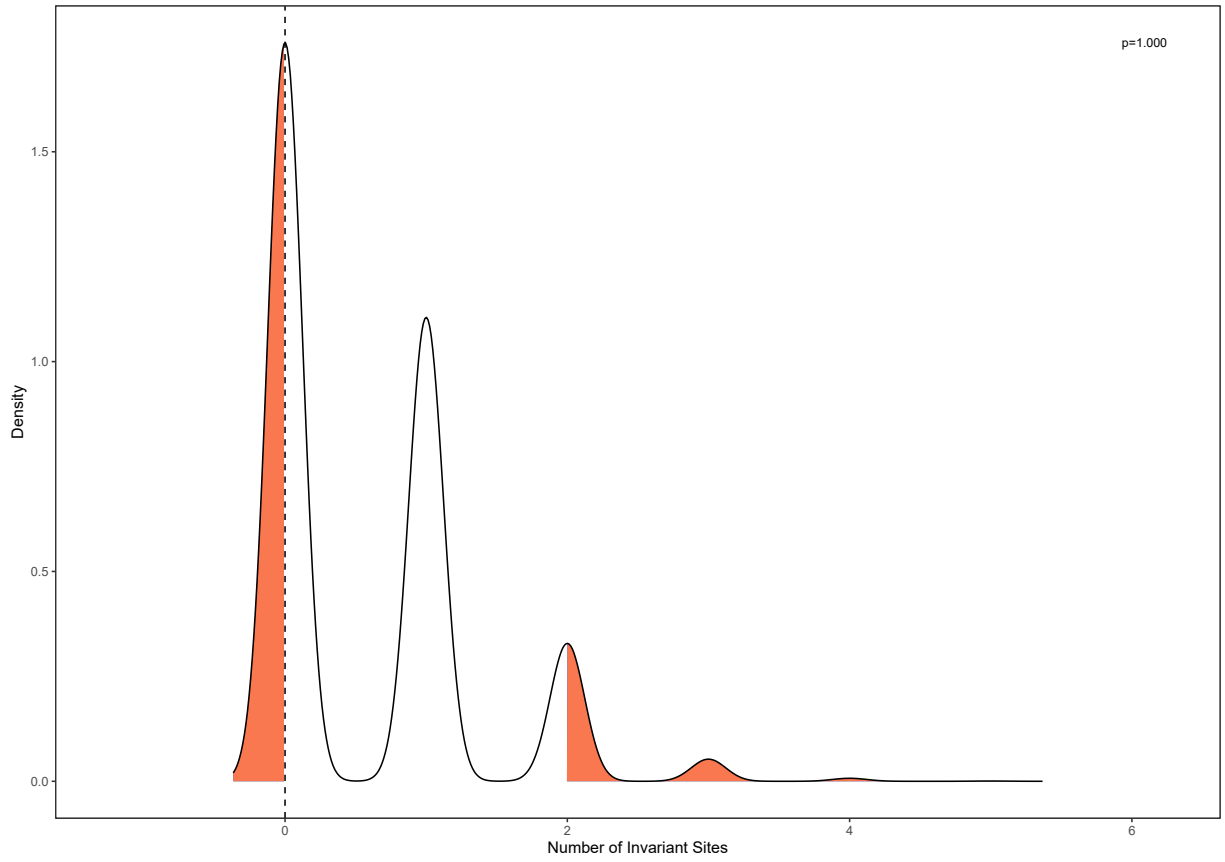

p=0.766
Density
1.5
1.0
0.5
0.0
-1.5
-1.0
-0.5
0.0
0.5
Tajima's-Pi

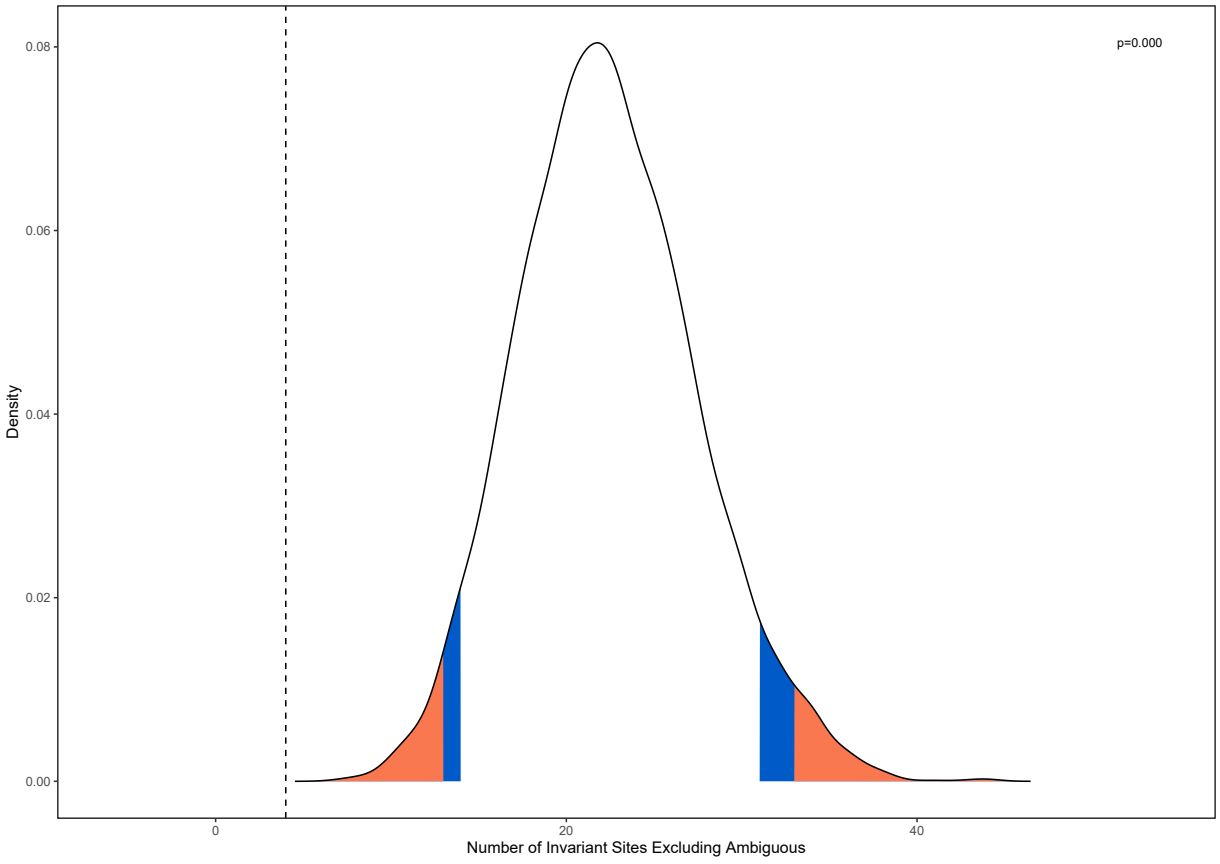

p=0.778
30
20
10
0
Density
0.09
0.12
0.15
0.18
0.21
Segregating-Sites

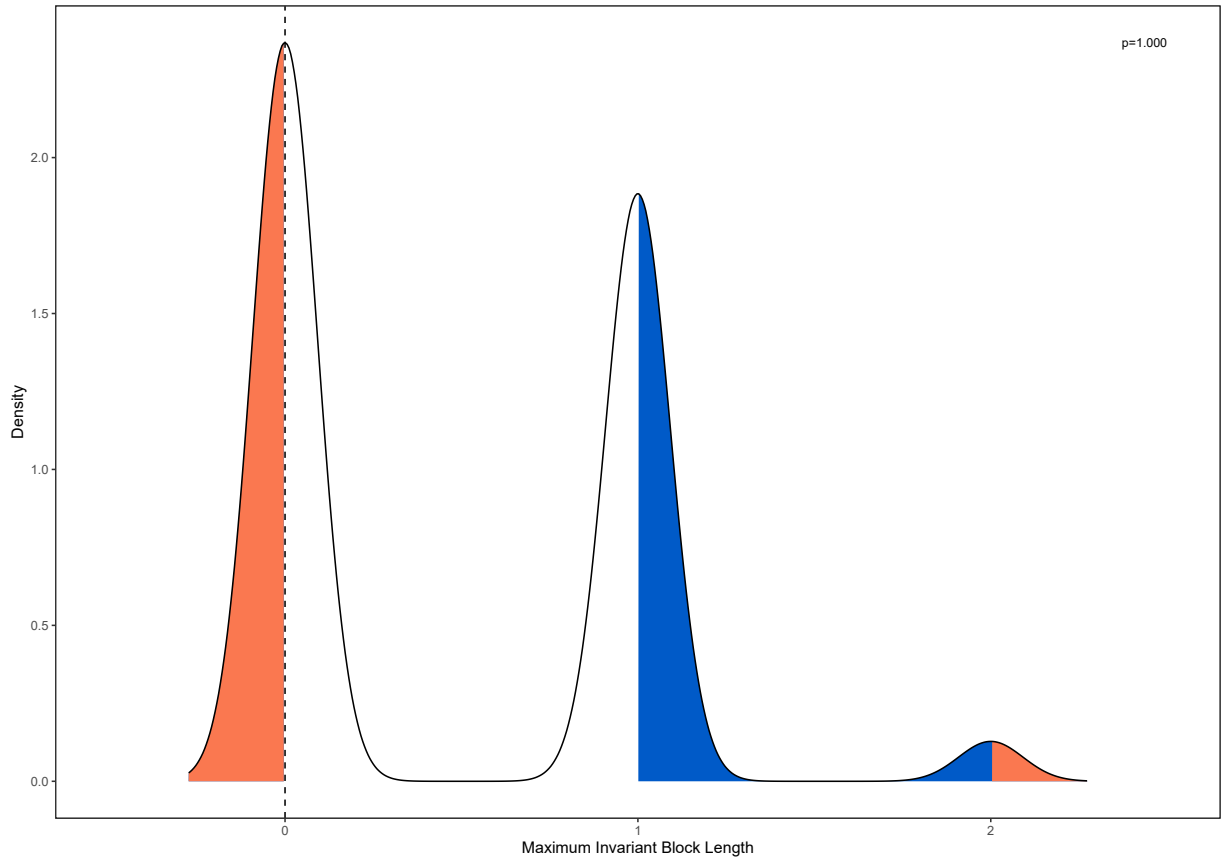

p=0.006
Density
10
5
0
0.4
0.5
0.6
Gower's Coefficient

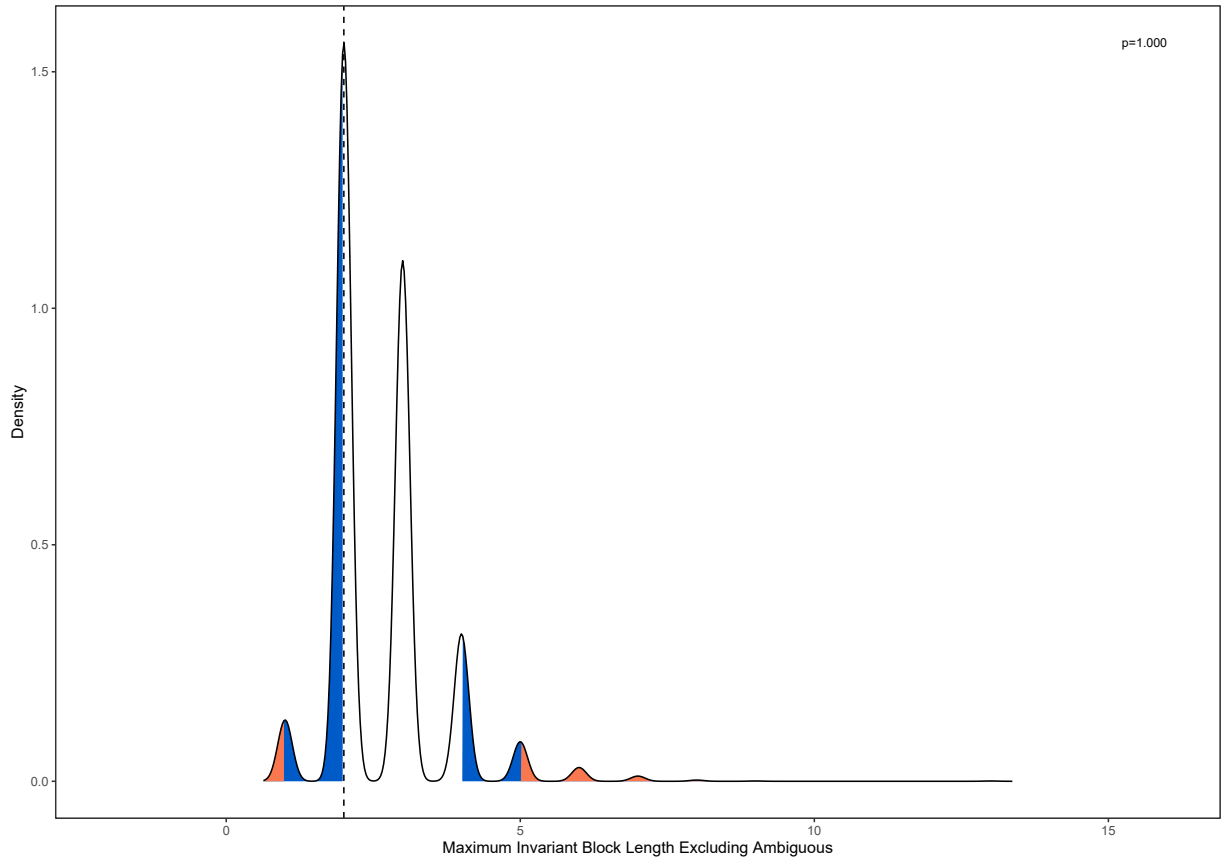

p=0.006
1.2
0.8
0.4
0.0
Density
4
5
6
Generalized Euclidean Distances

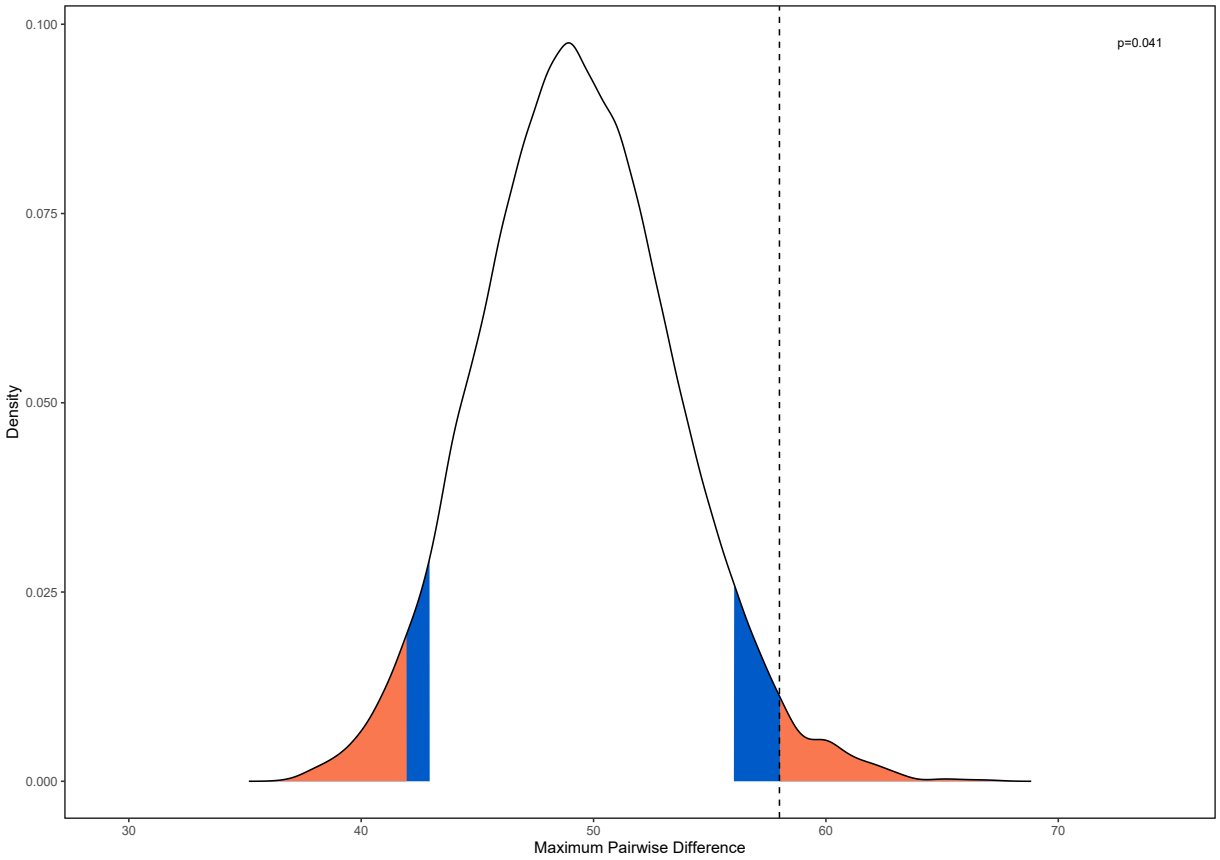
For SHDM model with 6 rate categories, the density graphs for all the summary statistics are as follows.

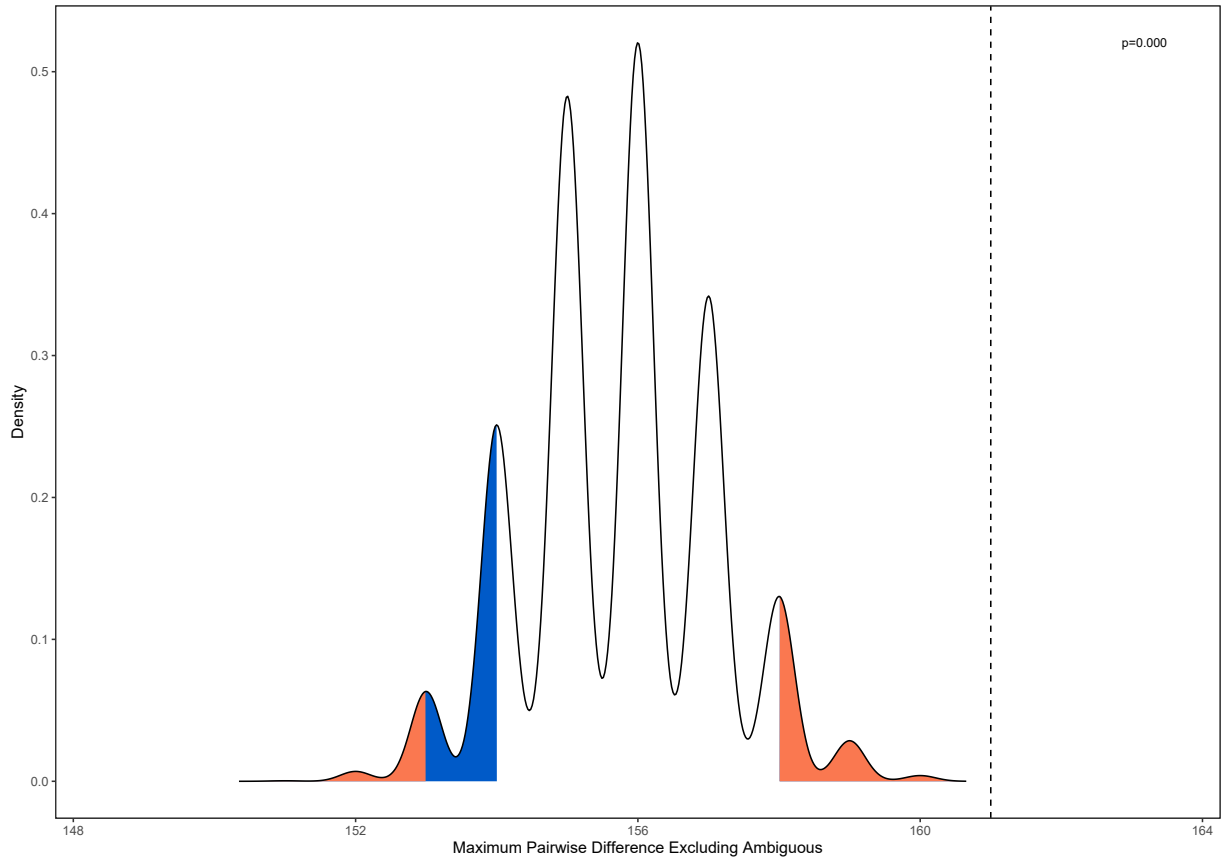

p=1.000
Density
1.5
1.0
0.5
0.0
0
2
4
6
Number of Invariant Sites

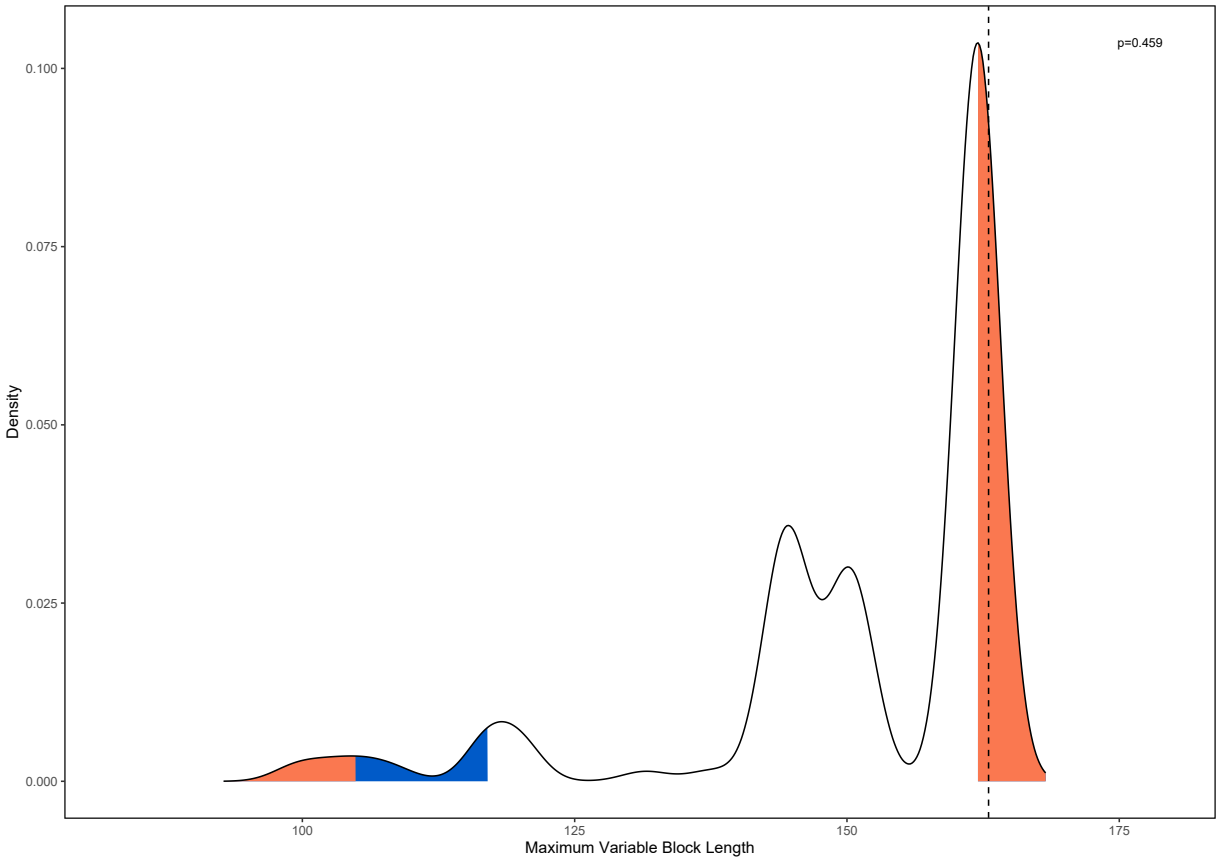

p=0.000
0.08
0.06
0.04
0.02
0.00
Density
0
20
40
Number of Invariant Sites Excluding Ambiguous

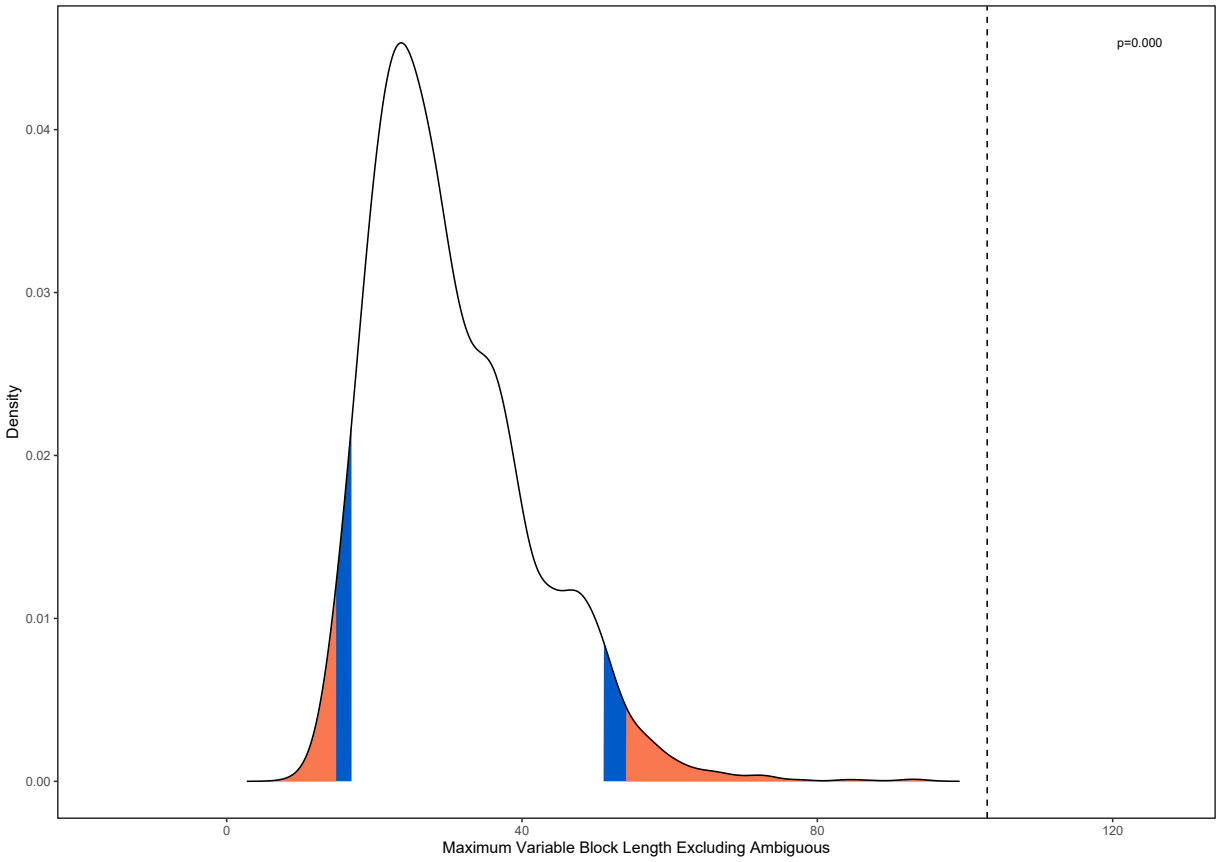

p=1.000
Density
2.0
1.5
1.0
0.5
0.0
0
1
2
Maximum Invariant Block Length

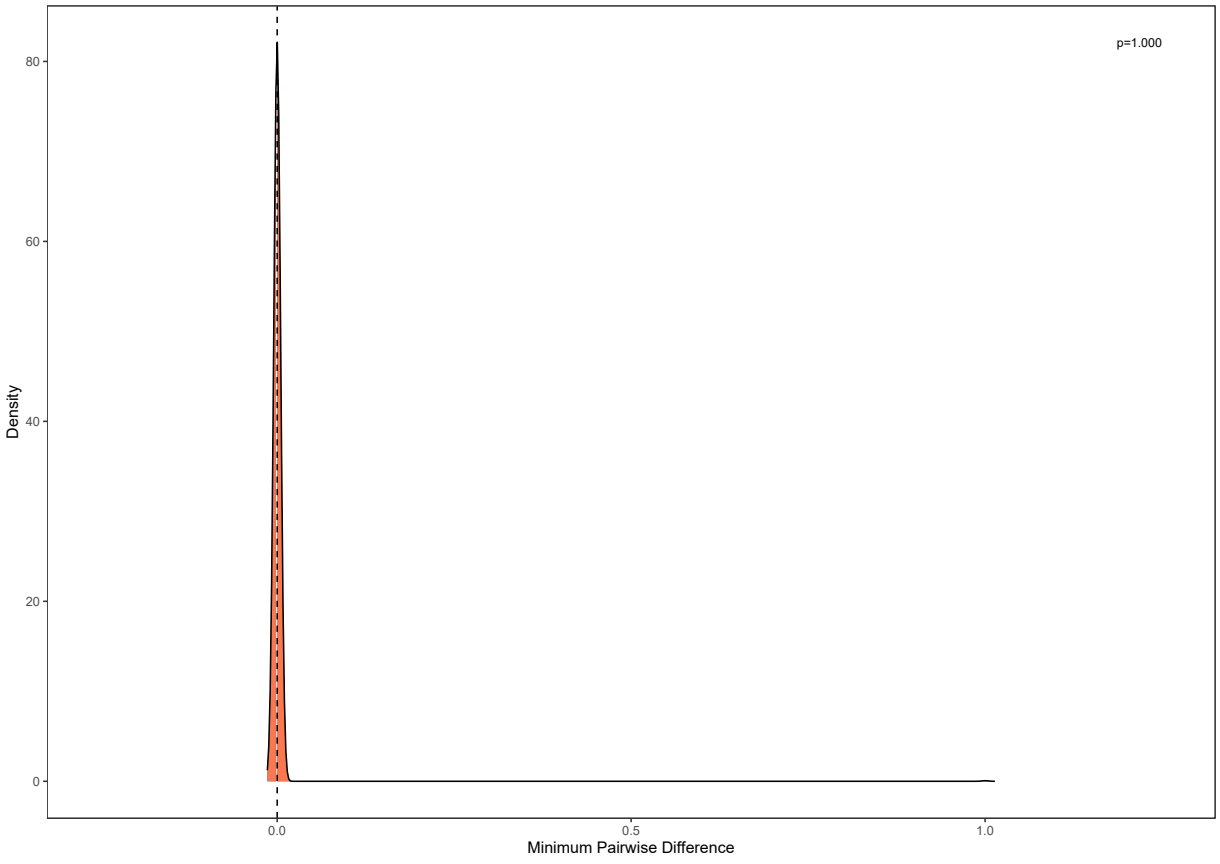

p=1.000
1.5
1.0
0.5
0.0
Density
0
5
10
15
Maximum Invariant Block Length Excluding Ambiguous

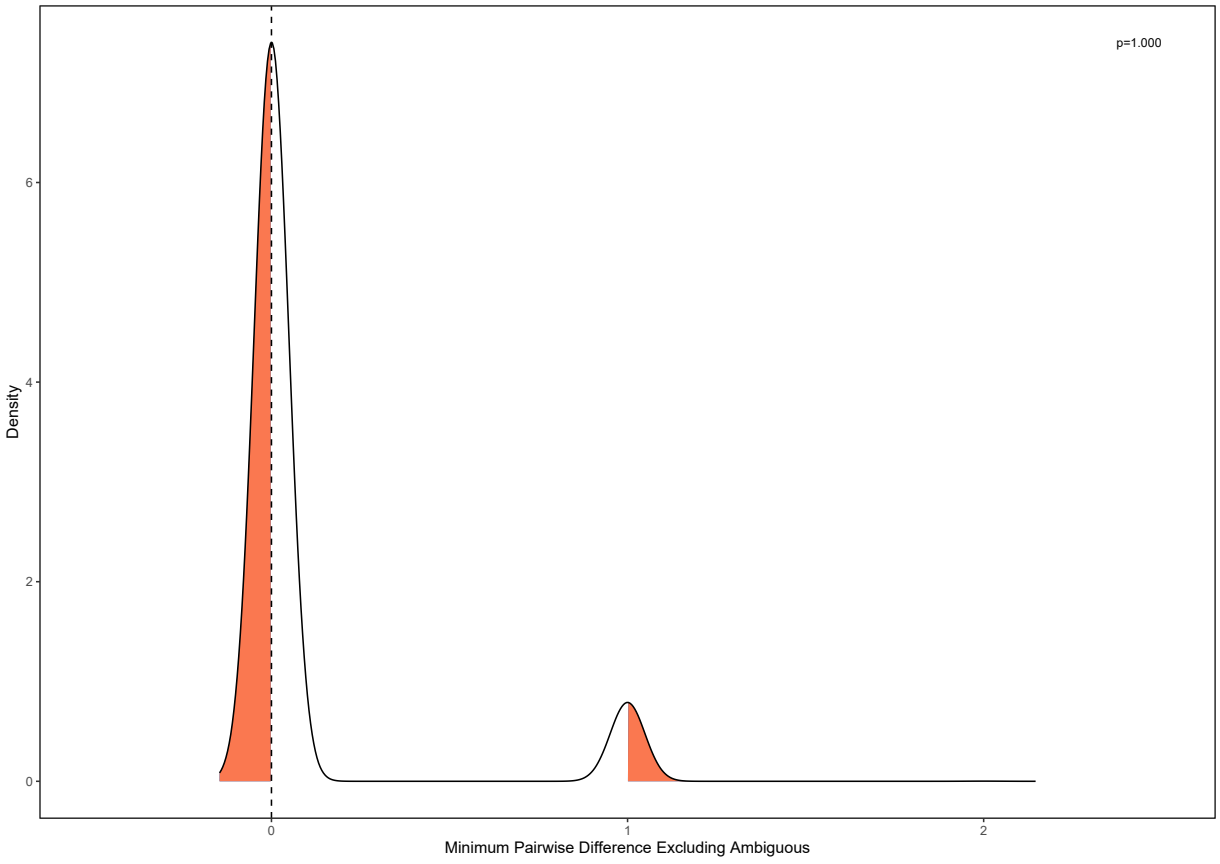

p=0.041
0.100
0.075
0.050
0.025
0.000
Density
30
40
50
60
70
Maximum Pairwise Difference

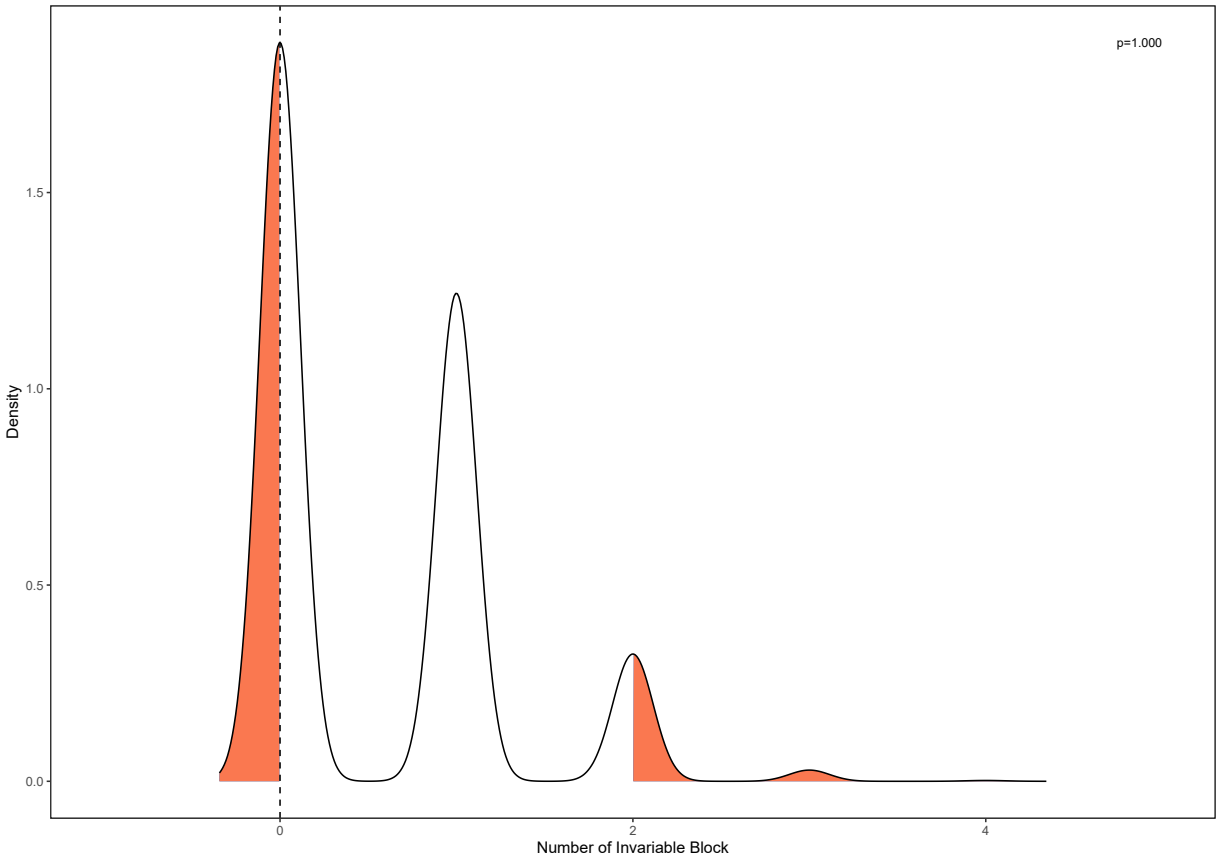

p=0.000
Density
0.0
0.1
0.2
0.3
0.4
0.5
148
152
156
160
164
Maximum Pairwise Difference Excluding Ambiguous

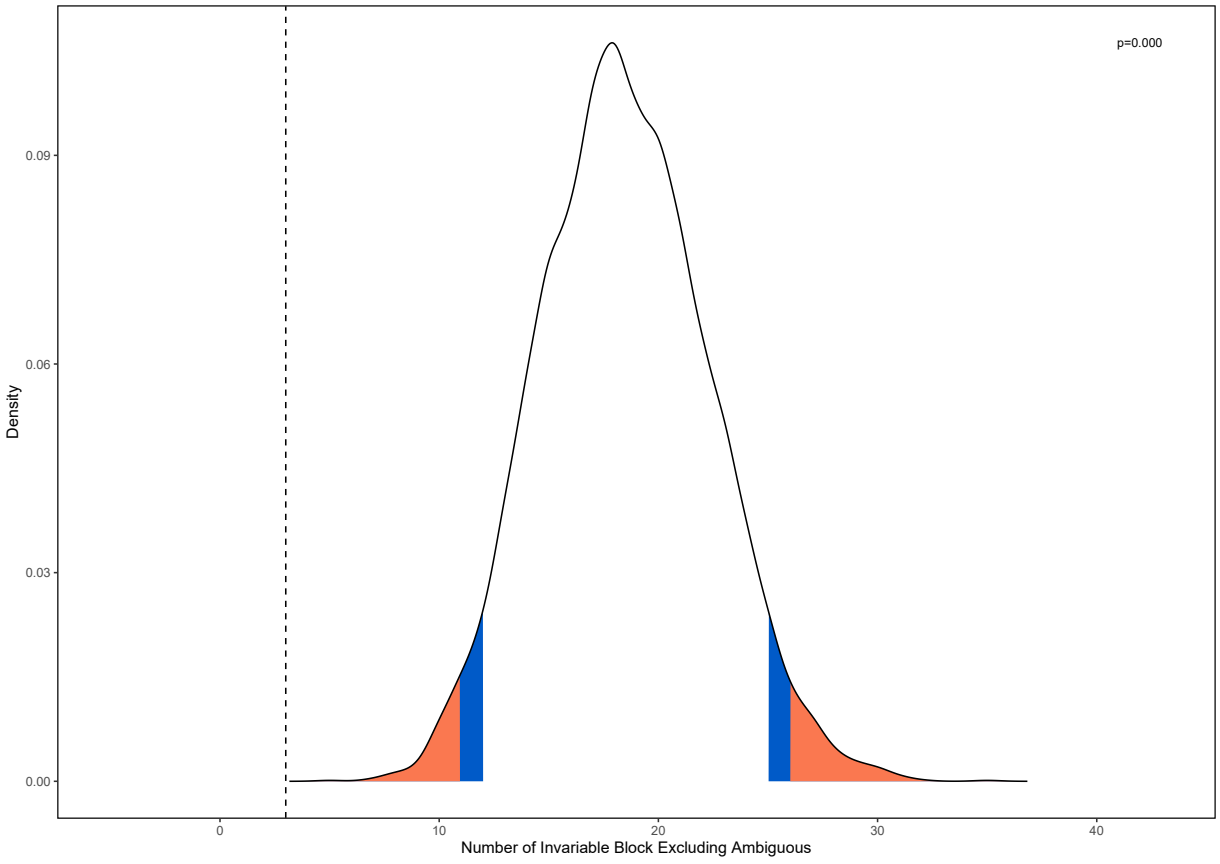

p=0.459
0.100
0.075
0.050
0.025
0.000
Density
100
125
150
175
Maximum Variable Block Length

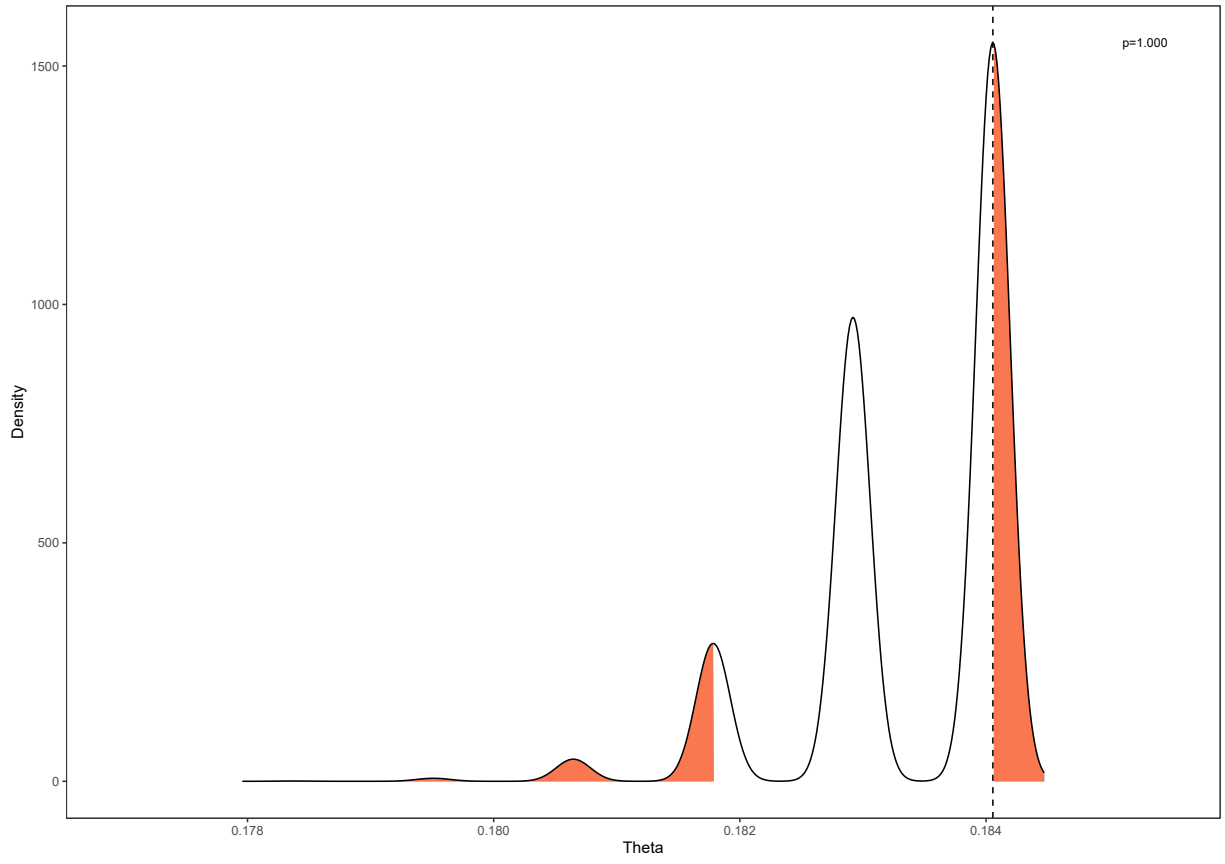

p=0.000
0.04
0.03
0.02
0.01
0.00
Density
0
40
80
120
Maximum Variable Block Length Excluding Ambiguous

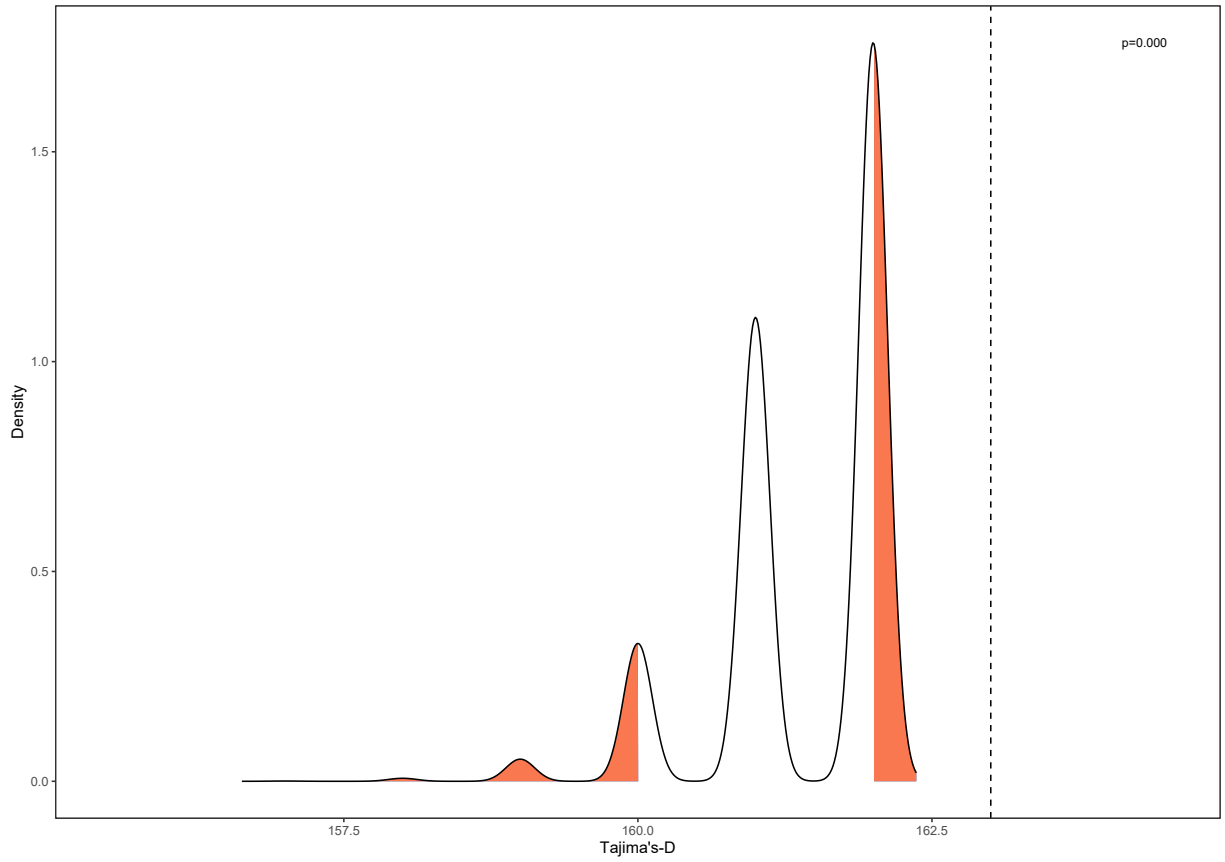

p=1.000
Density
80
60
40
20
0
0.0
0.5
1.0
Minimum Pairwise Difference

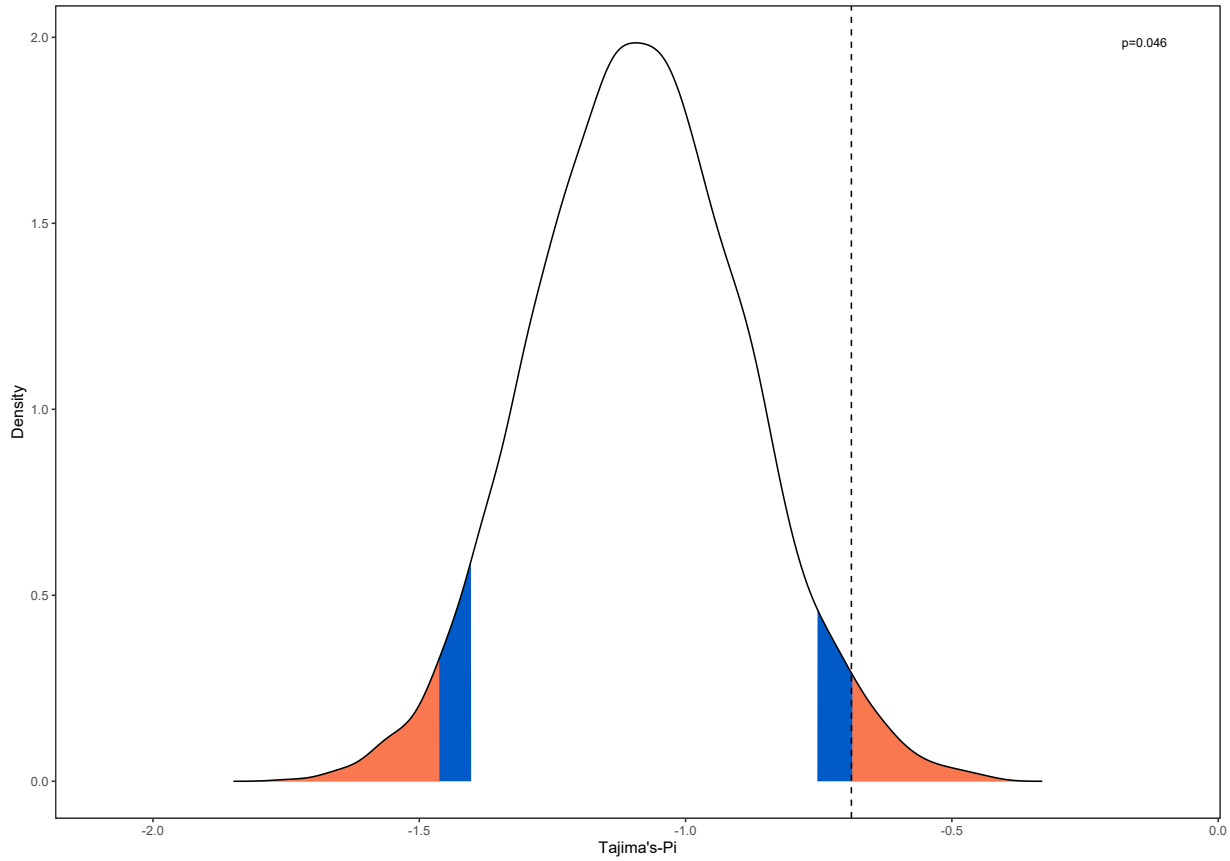

p=1.000
Density
6
4
2
0
0
1
2
Minimum Pairwise Difference Excluding Ambiguous

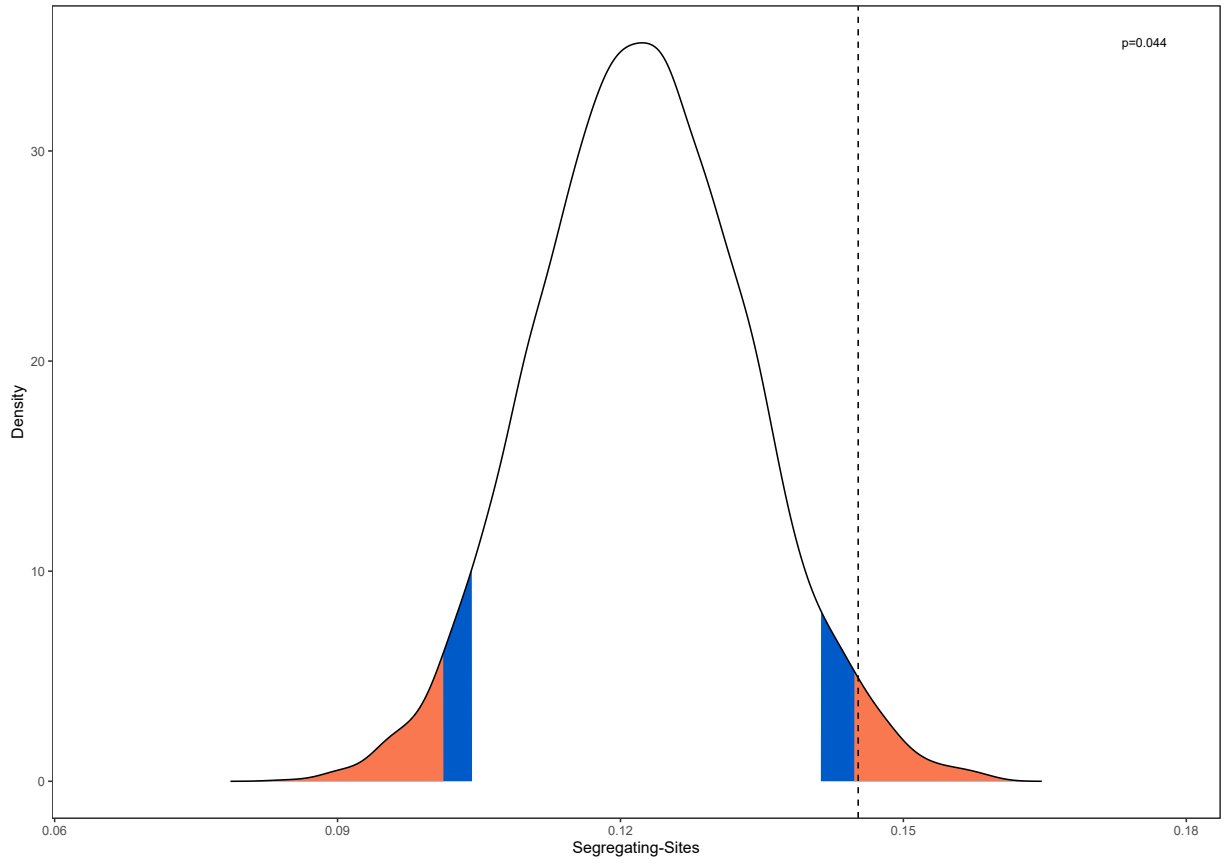

p=1.000
Density
1.5
1.0
0.5
0.0
0
2
4
Number of Invariable Block

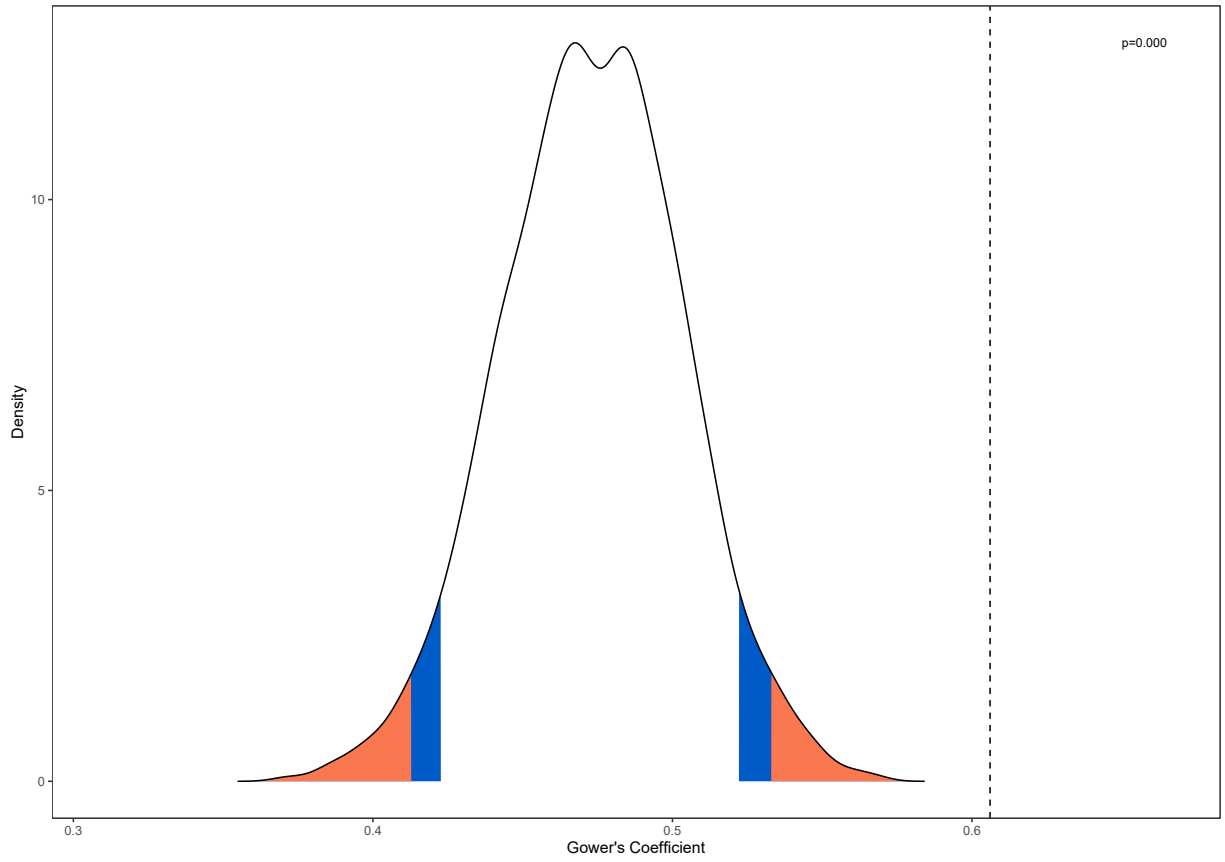

p=0.000
0.09
0.06
0.03
0.00
Density
0
10
20
30
40
Number of Invariable Block Excluding Ambiguous

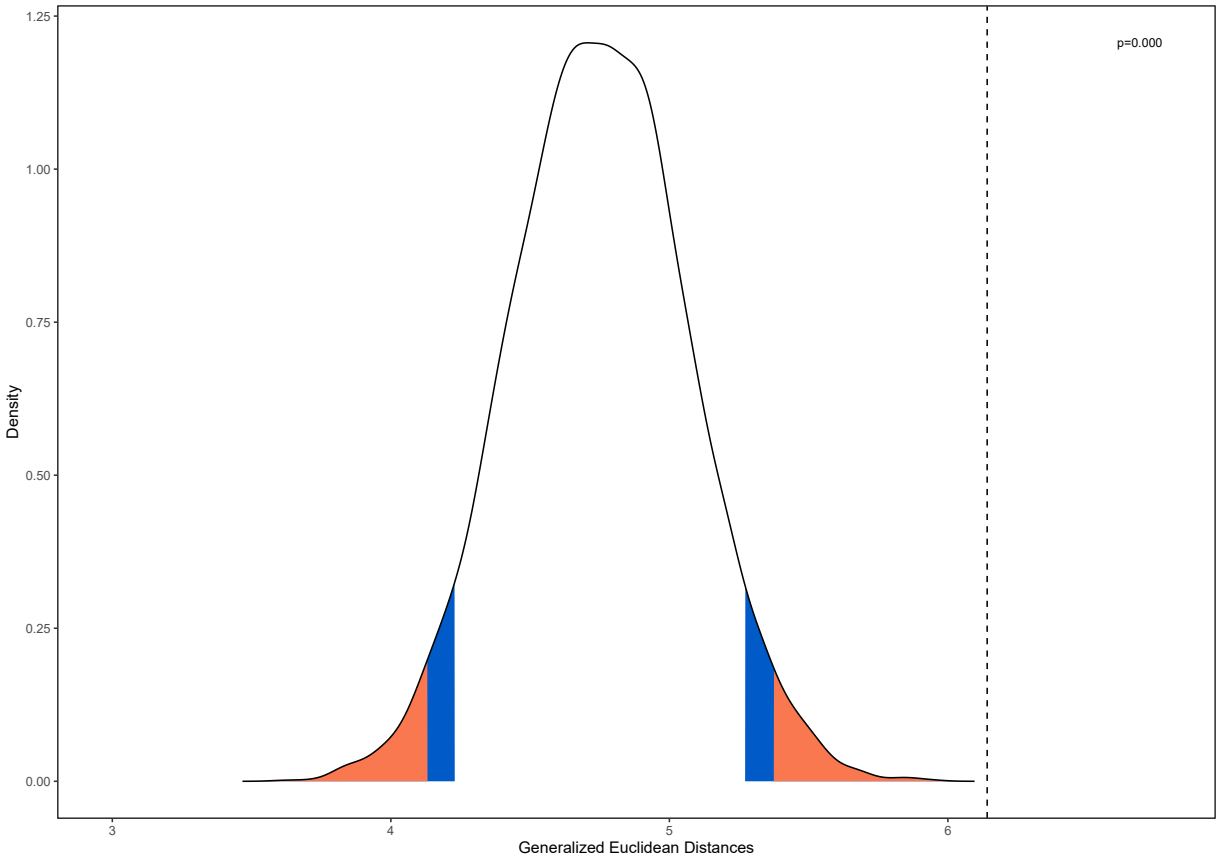

p=1.000
1500
1000
500
0
Density
0.178
0.180
0.182
0.184
Theta


p=0.000
1.5
1.0
0.5
0.0
Density
157.5
160.0
162.5
Tajima's-D

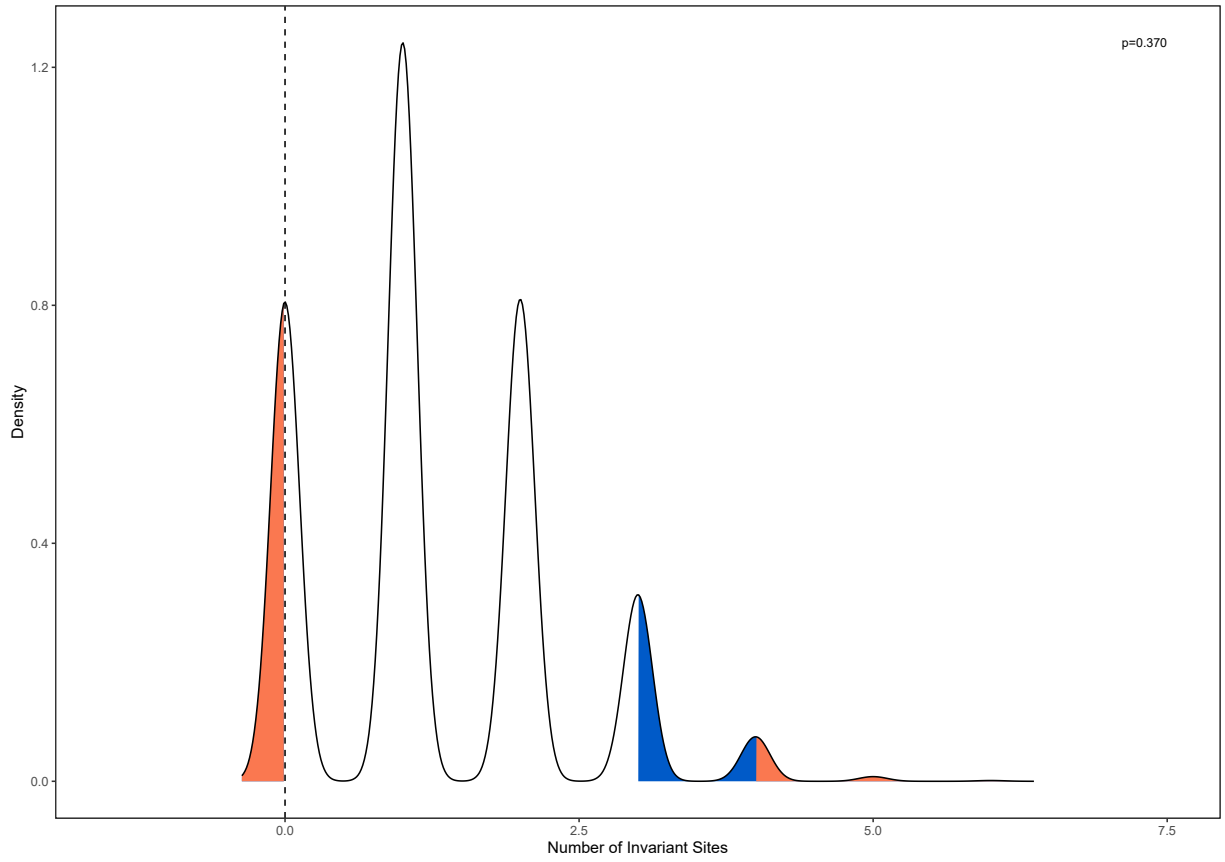

p=0.046
Density
2.0
1.5
1.0
0.5
0.0
-2.0
-1.5
-1.0
-0.5
0.0
Tajima's-Pi

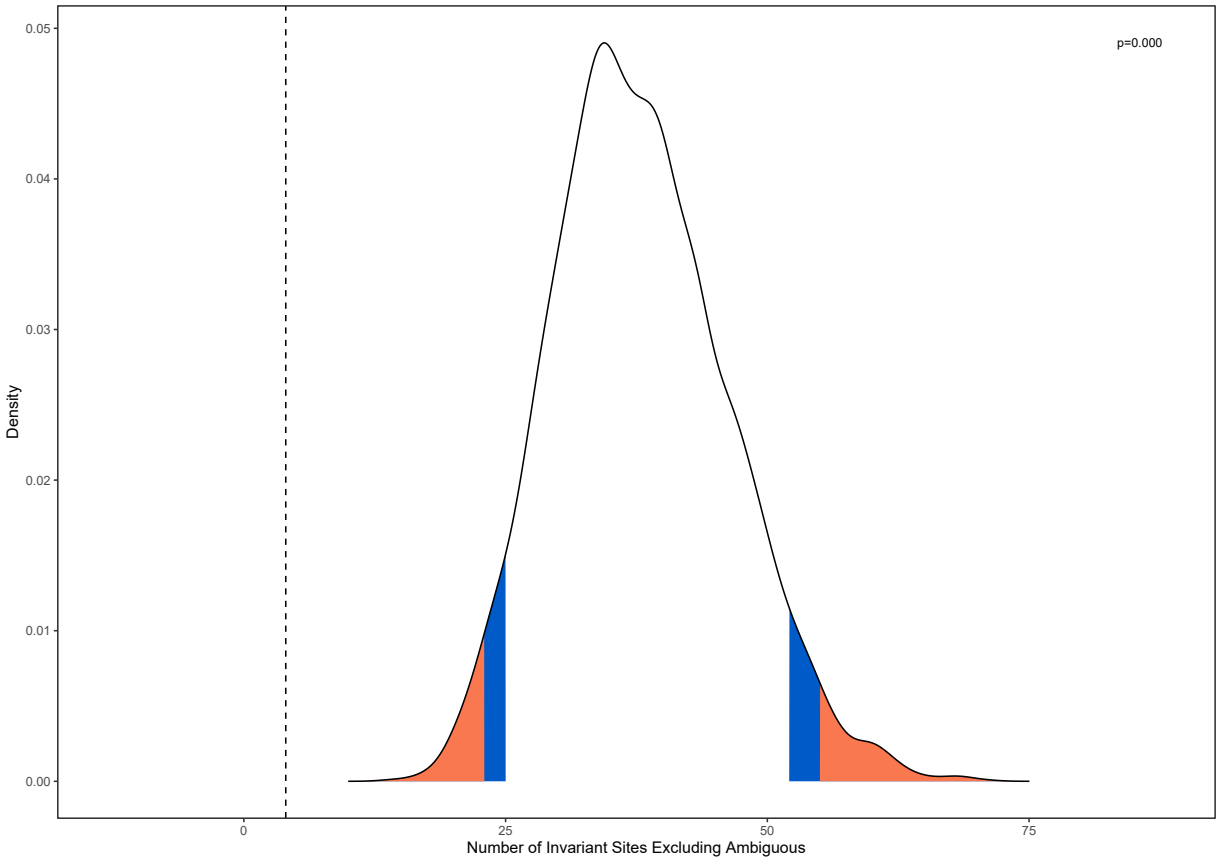

p=0.044
Density
30
20
10
0
0.06
0.09
0.12
0.15
0.18
Segregating-Sites

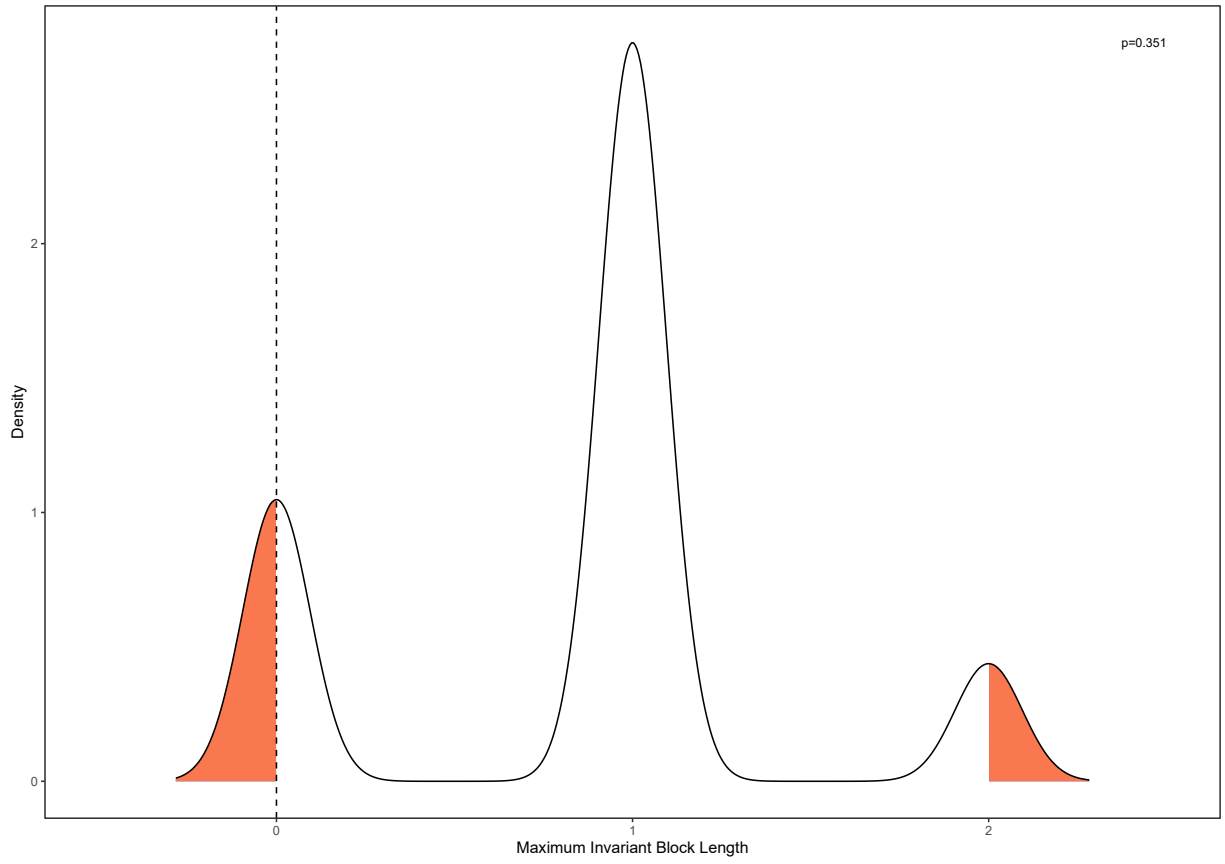

p=0.000
Density
10
5
0
0.3
0.4
0.5
0.6
Gower's Coefficient

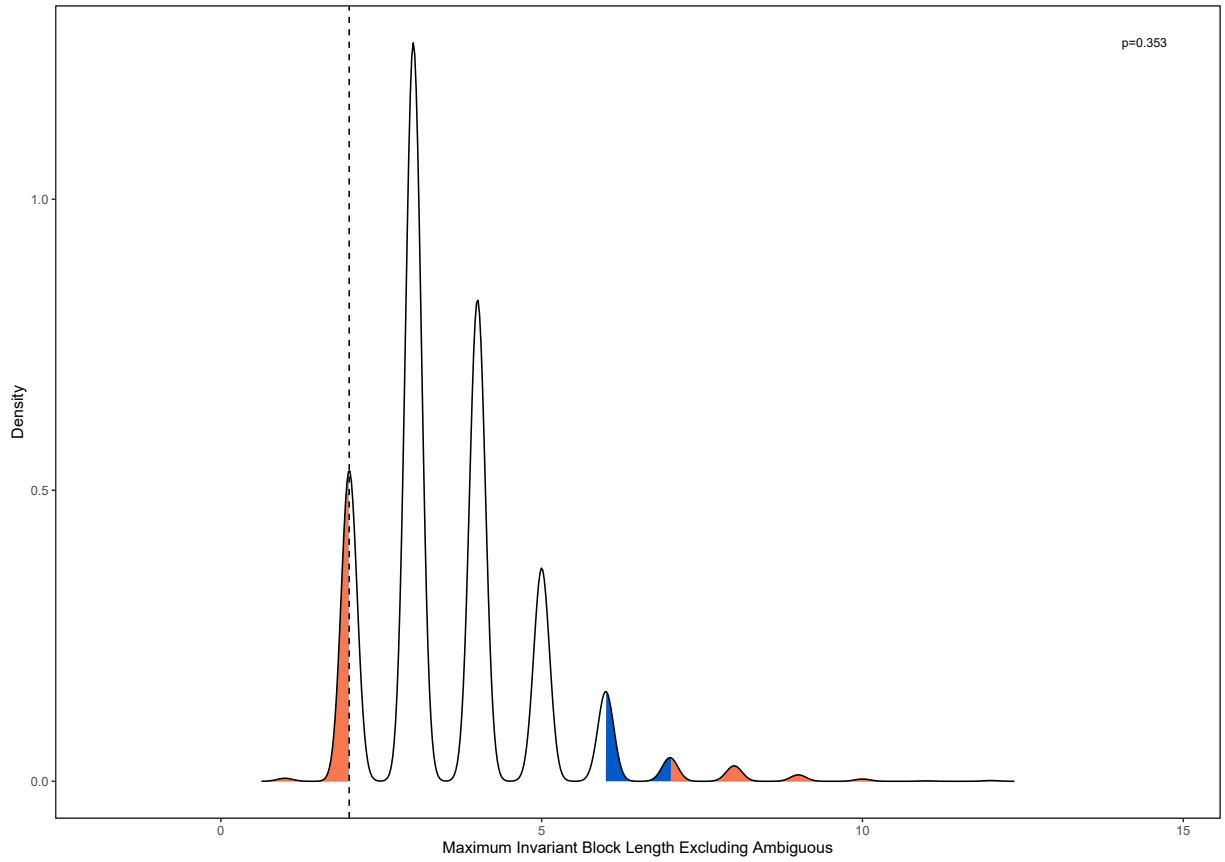

p=0.000
Density
1.25
1.00
0.75
0.50
0.25
0.00
3
4
5
6
Generalized Euclidean Distances

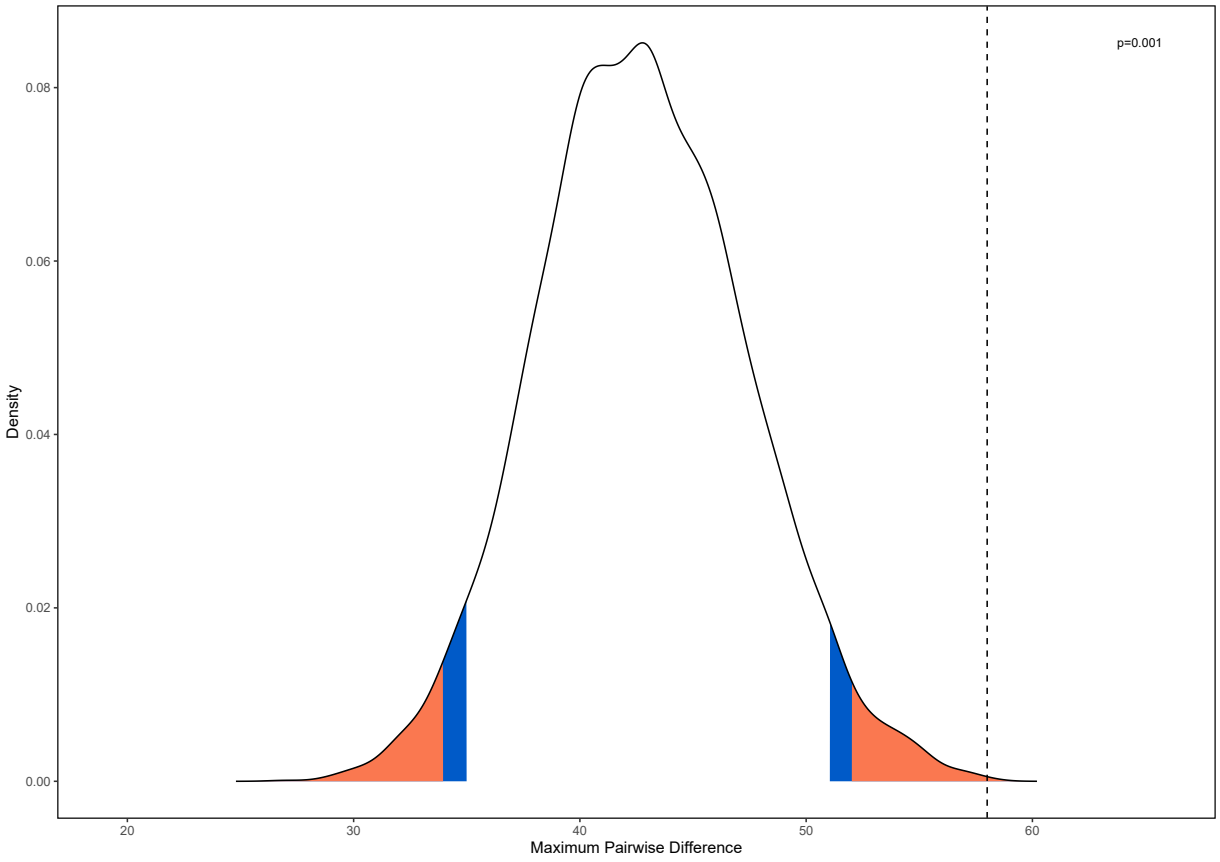
For SHDM model with 8 rate categories, the density graphs for all the summary statistics are as follows.

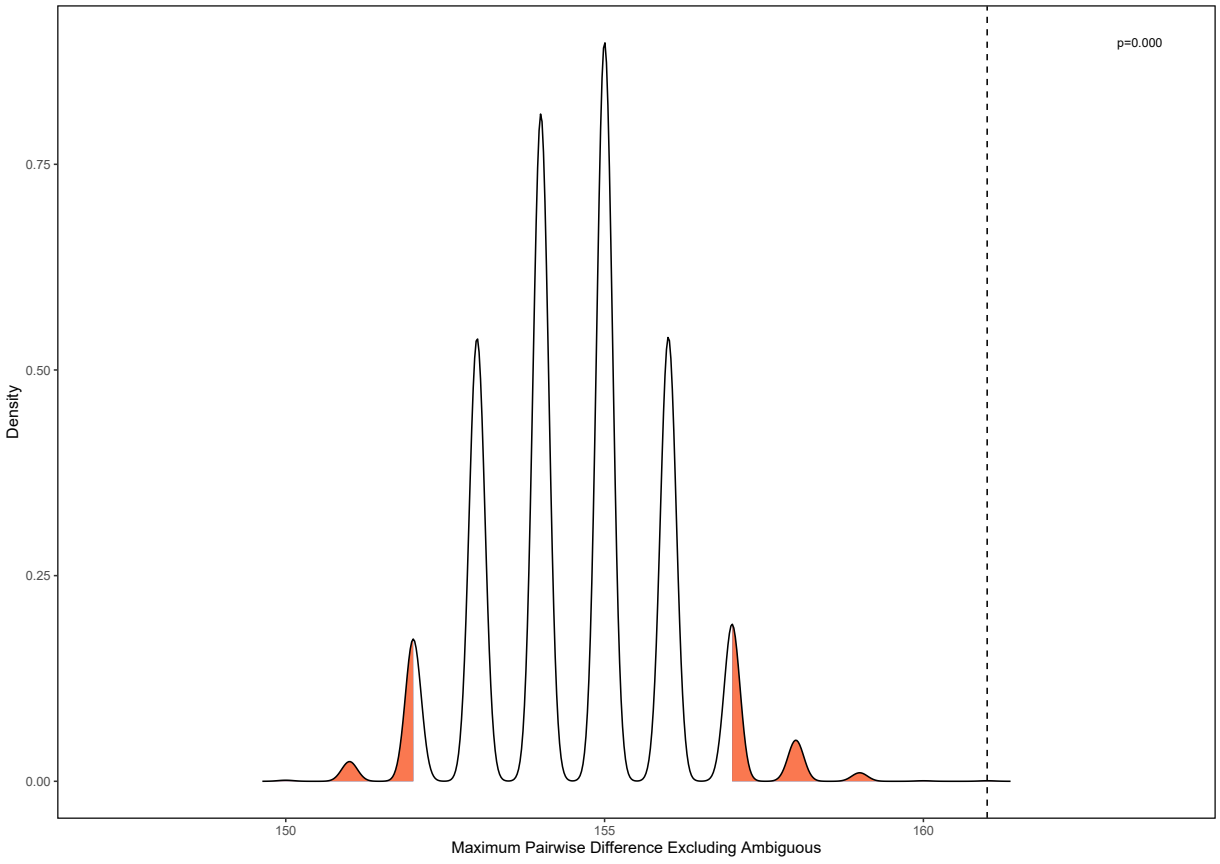

p=0.370
1.2
0.8
0.4
0.0
Density
0.0
2.5
5.0
7.5
Number of Invariant Sites

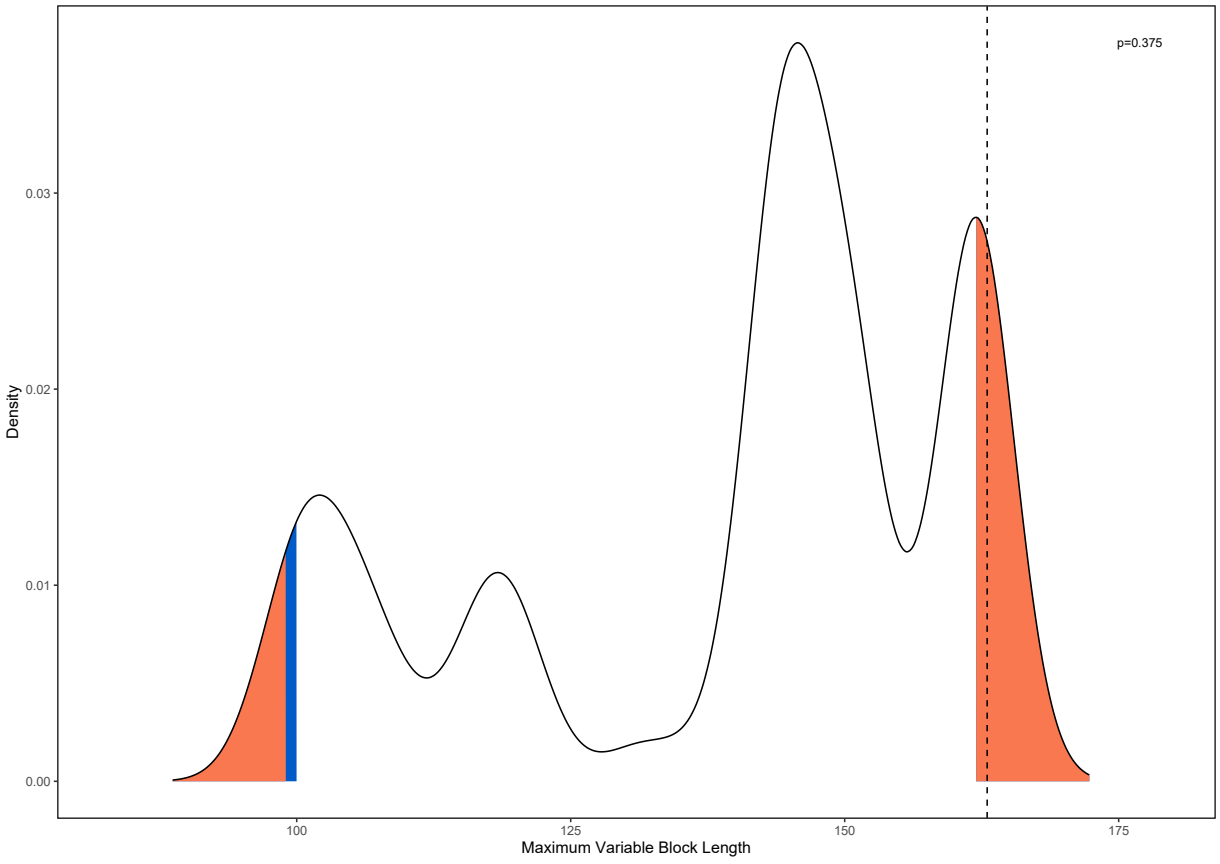

p=0.000
Density
0.05
0.04
0.03
0.02
0.01
0.00
0
25
50
75
Number of Invariant Sites Excluding Ambiguous

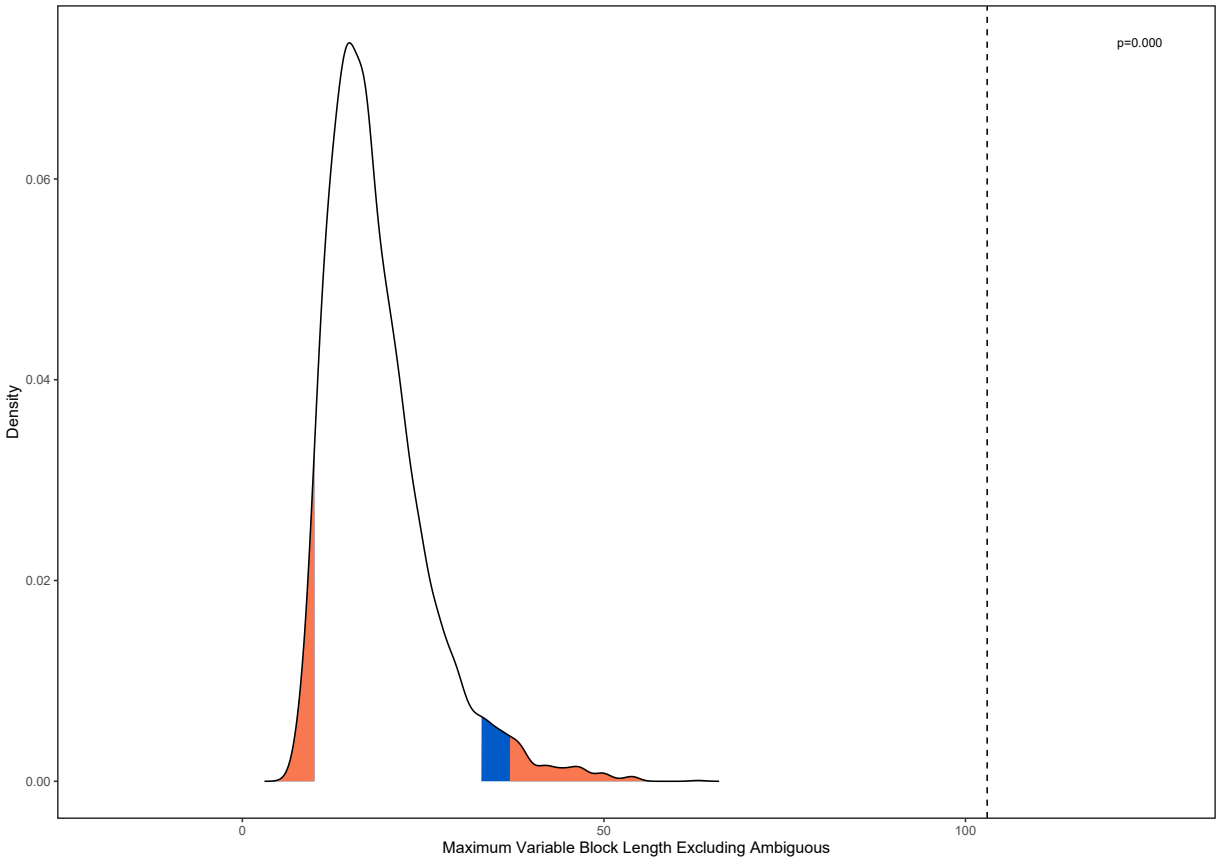

p=0.351
Density
2
1
0
0
1
2
Maximum Invariant Block Length

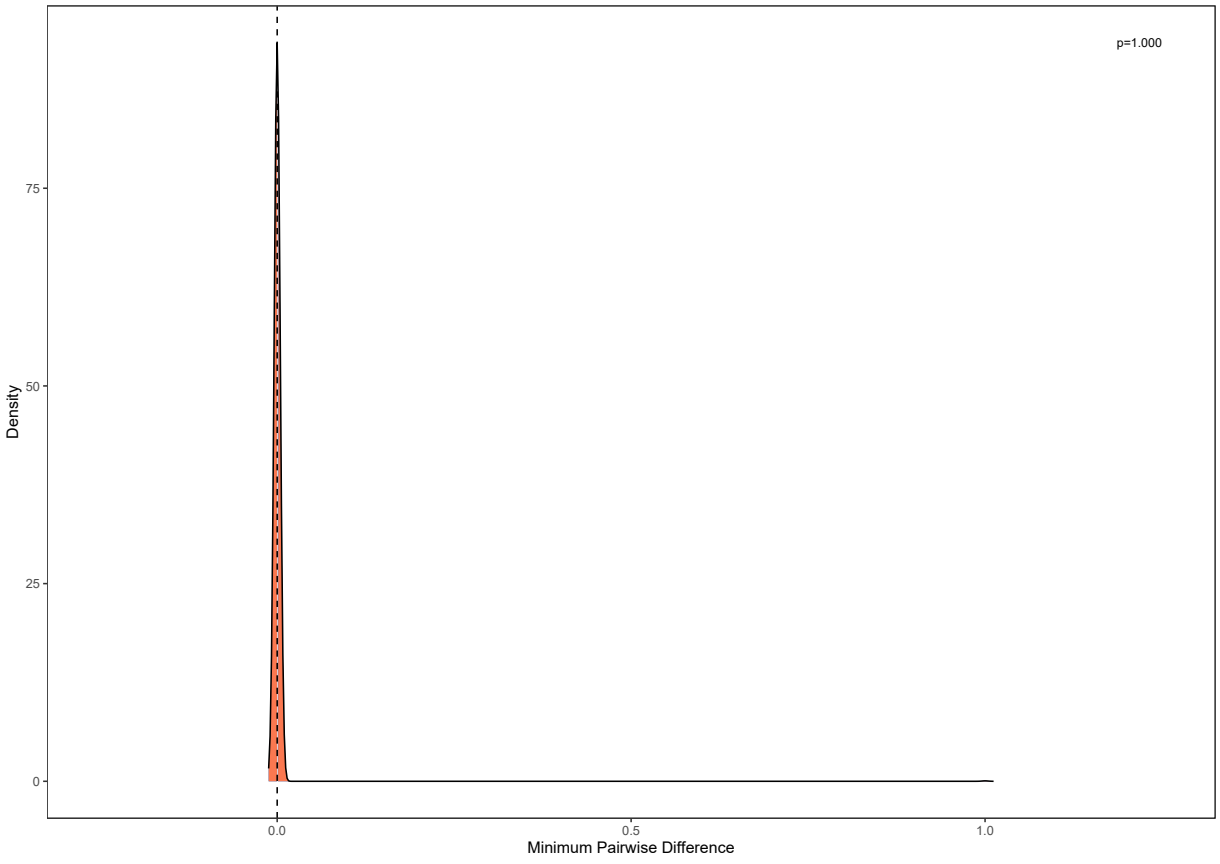

p=0.353
Density
1.0
0.5
0.0
0
5
10
15
Maximum Invariant Block Length Excluding Ambiguous

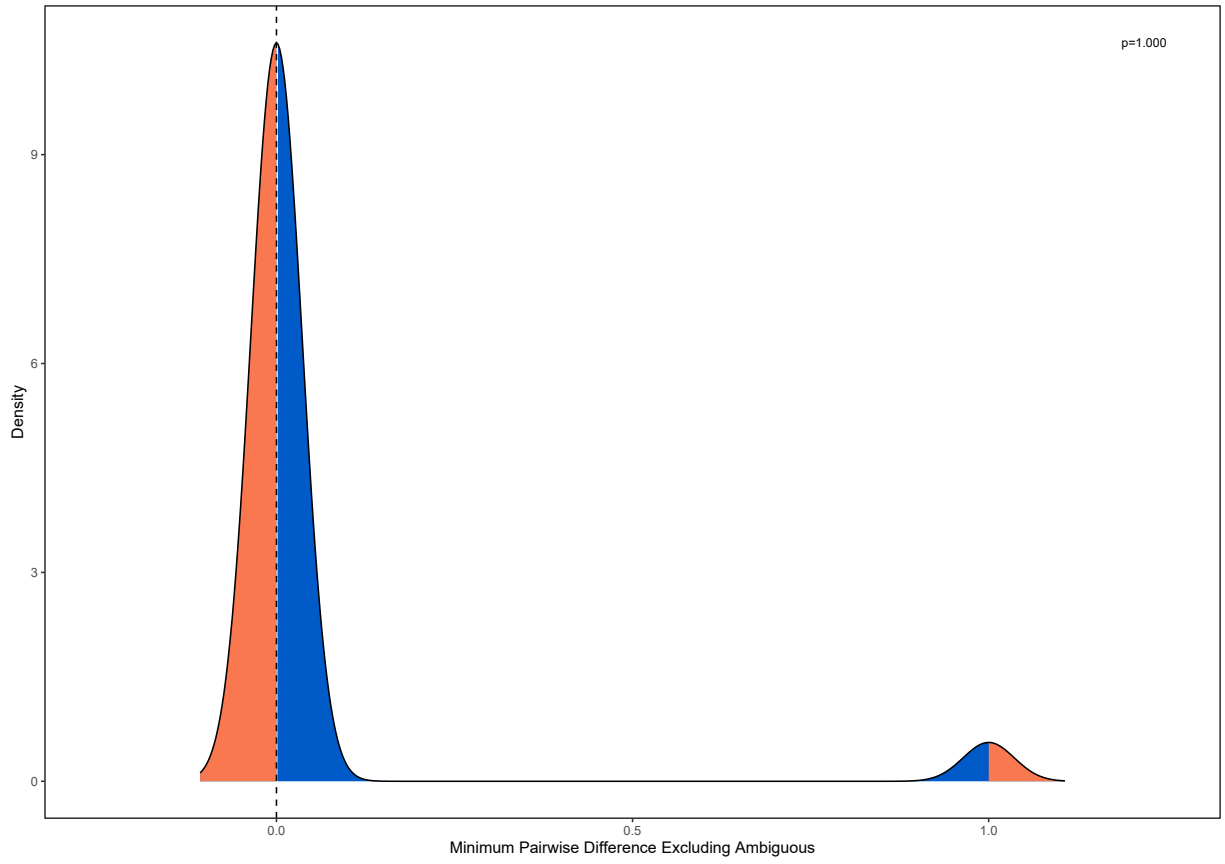

p=0.001
Density
0.08
0.06
0.04
0.02
0.00
20
30
40
50
60
Maximum Pairwise Difference

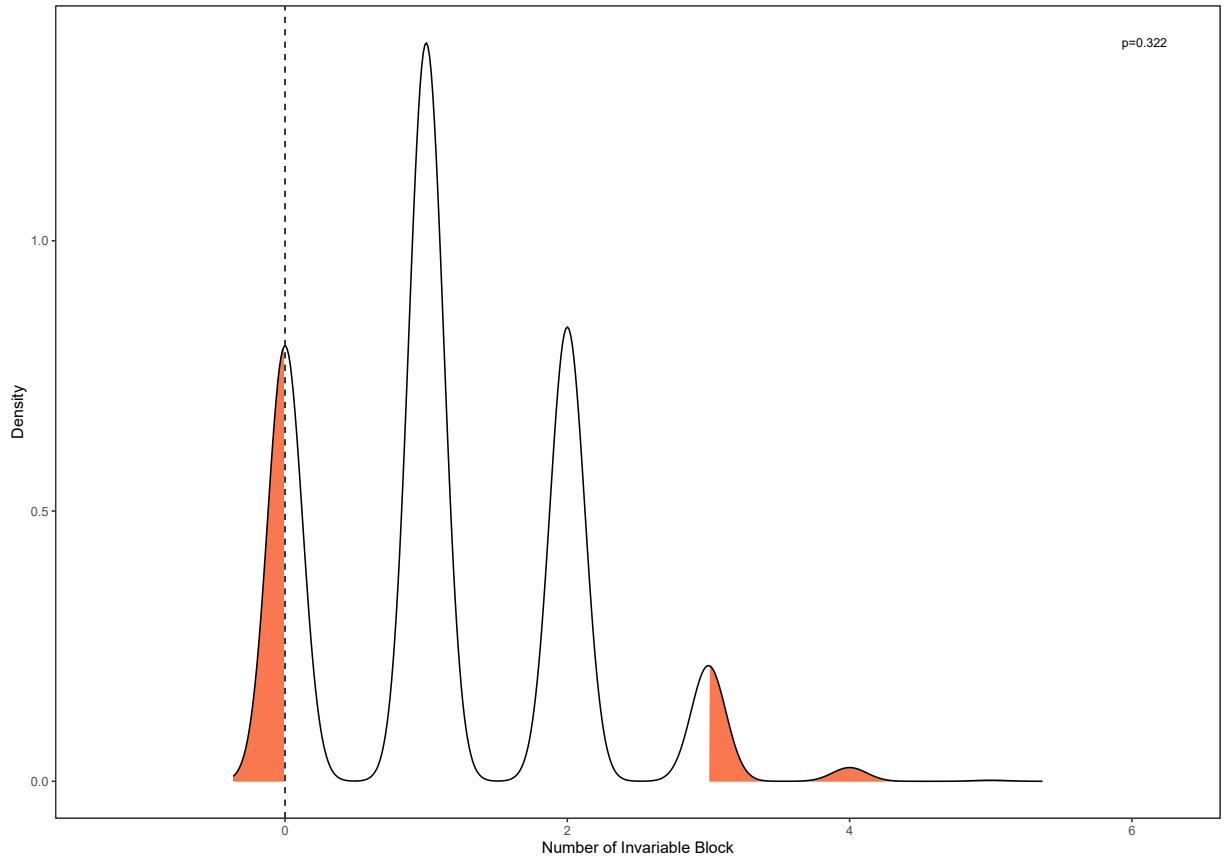

p=0.000
0.75
0.50
0.25
0.00
Density
150
155
160
Maximum Pairwise Difference Excluding Ambiguous

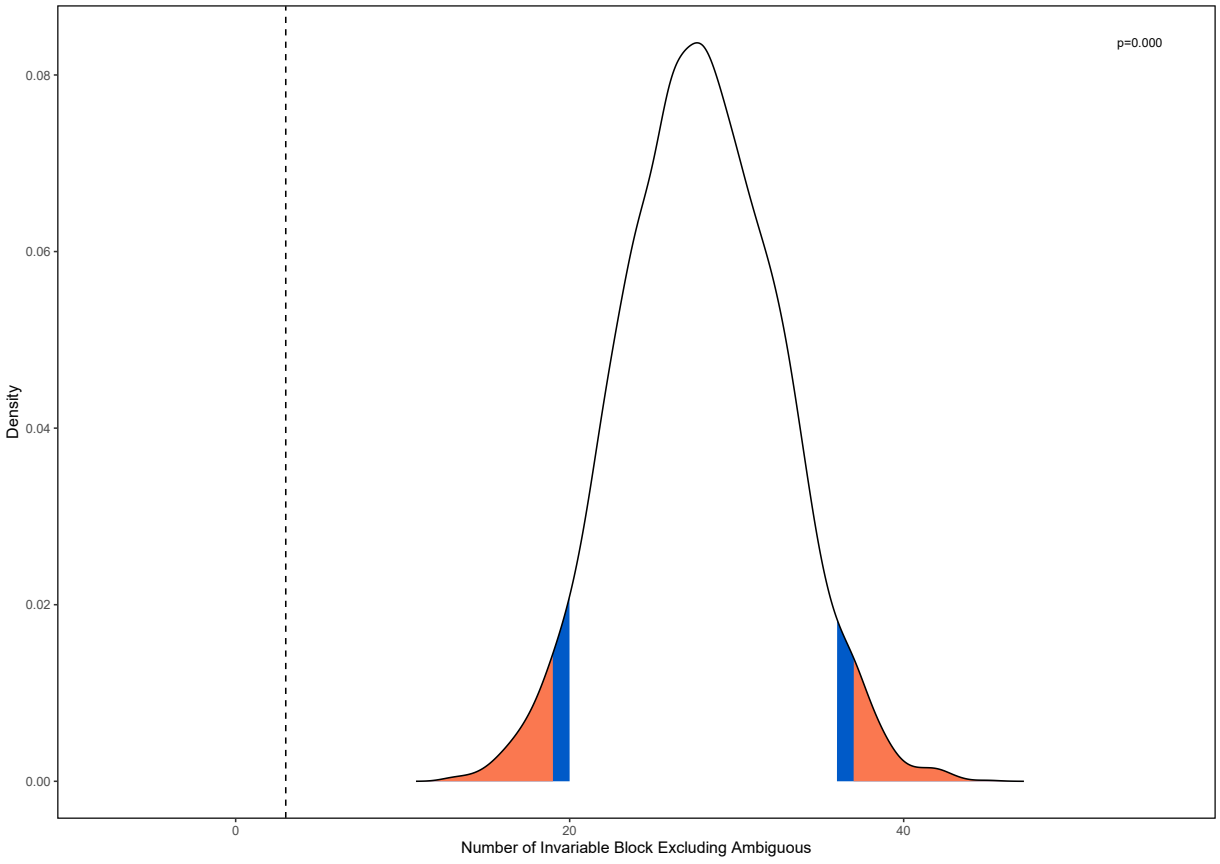

p=0.375
0.03
0.02
0.01
0.00
Density
100
125
150
175
Maximum Variable Block Length

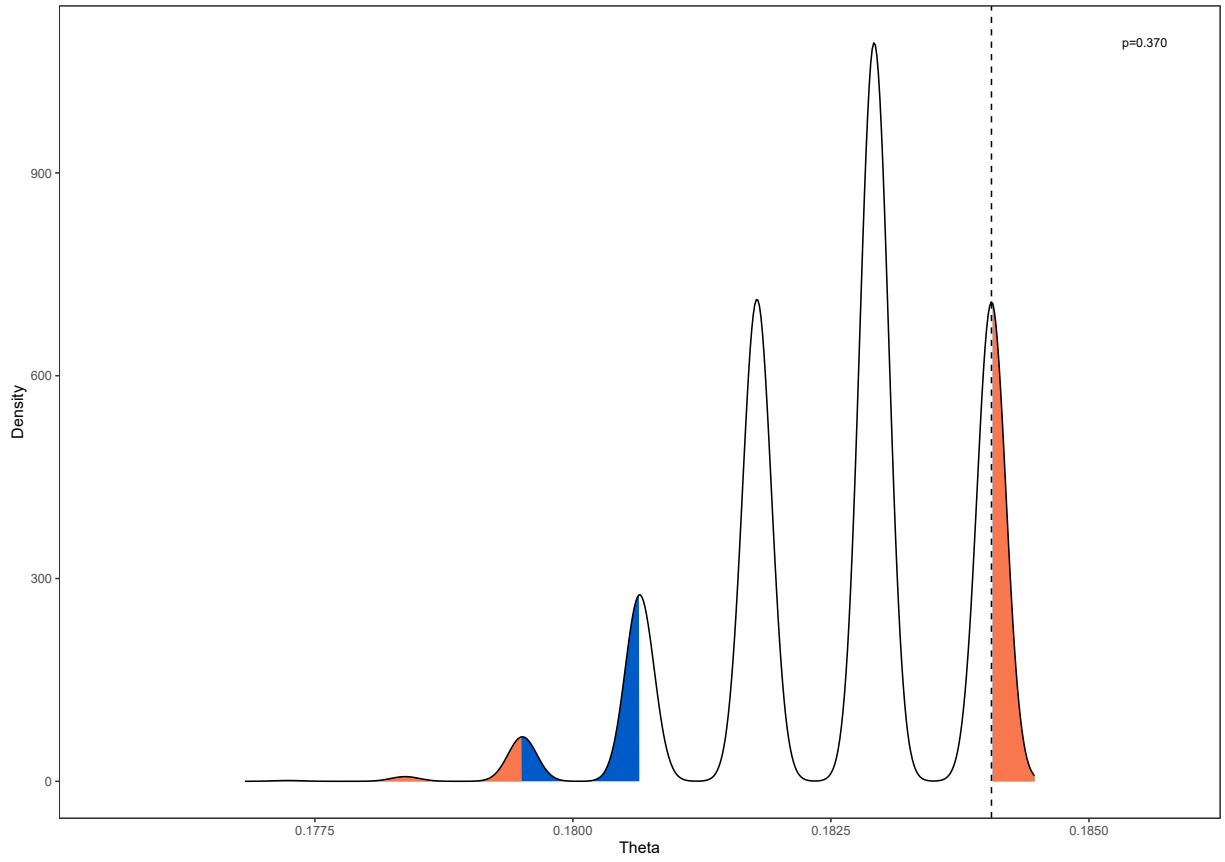

p=0.000
0.06
0.04
0.02
0.00
Density
0
50
100
Maximum Variable Block Length Excluding Ambiguous

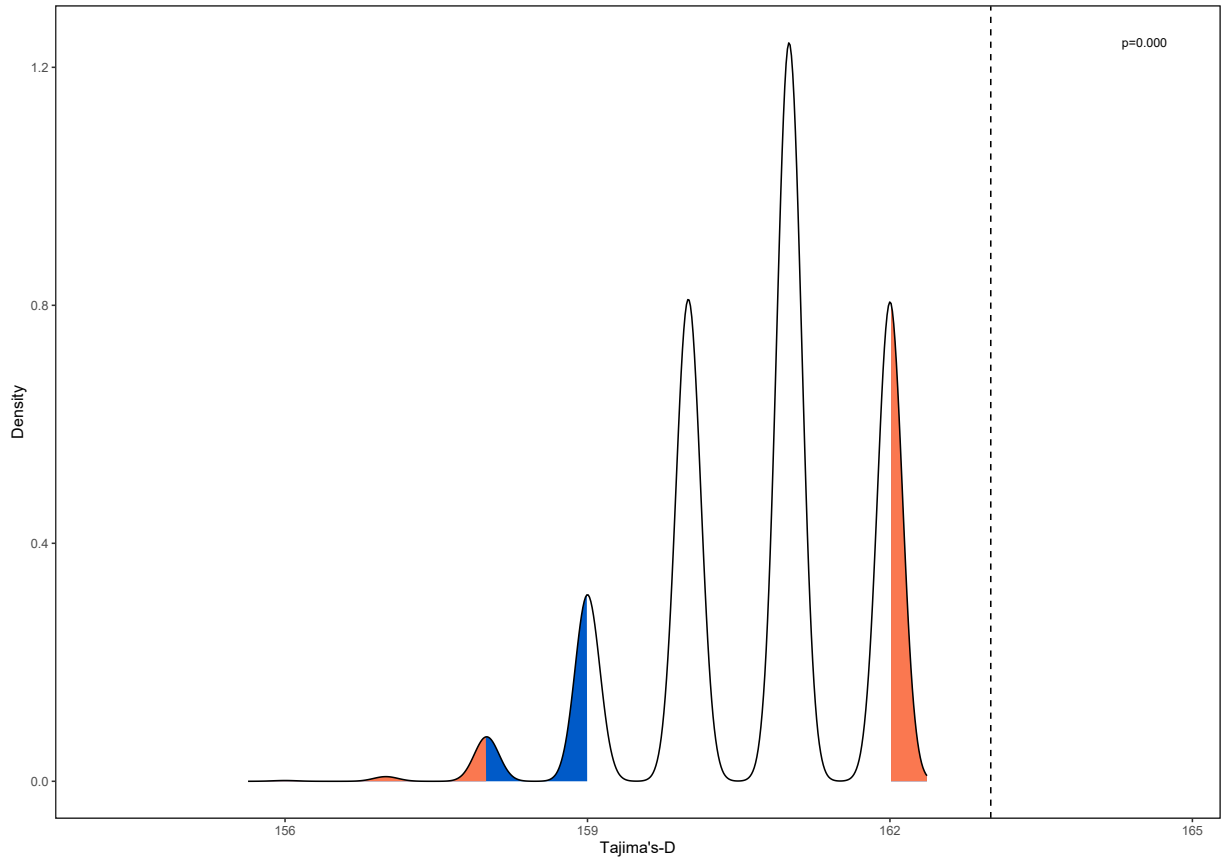

p=1.000
75
50
25
0
Density
0.0
0.5
1.0
Minimum Pairwise Difference

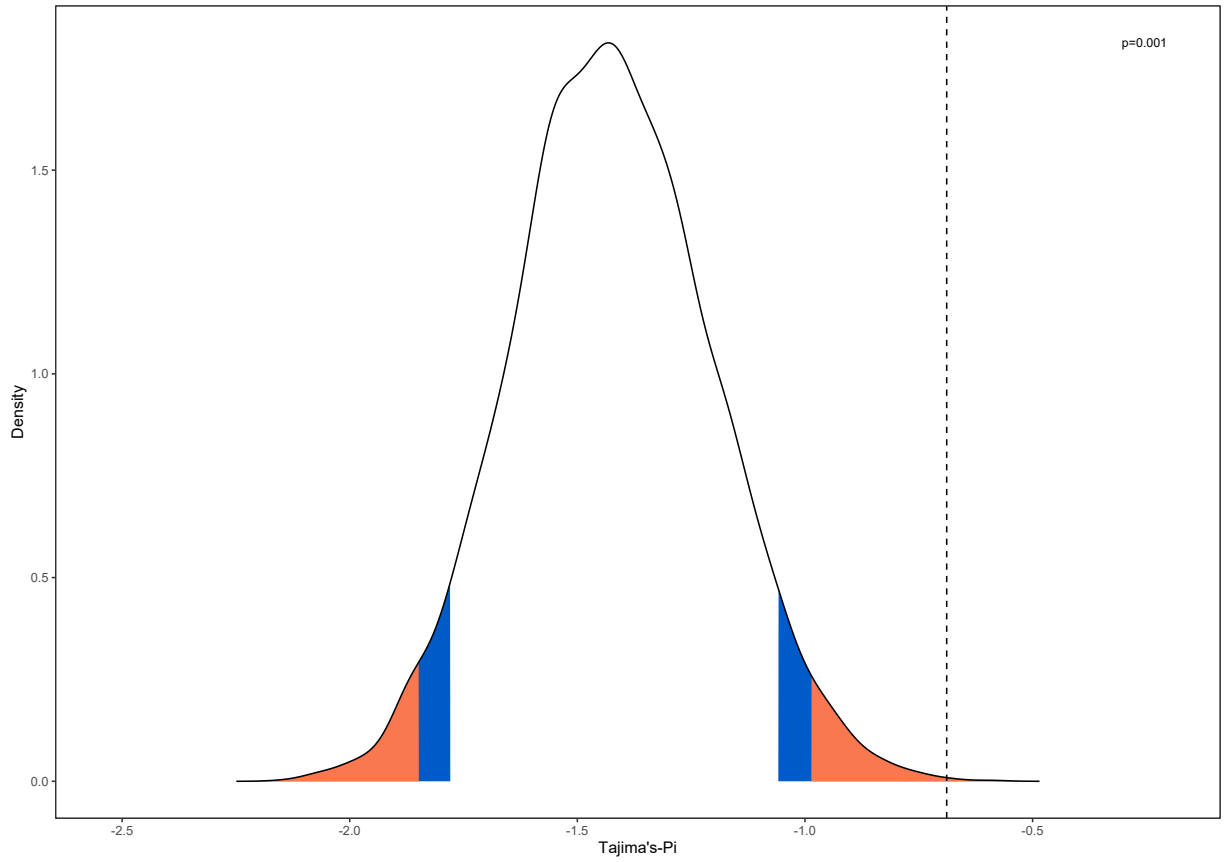

p=1.000
Density
9
6
3
0
0.0
0.5
1.0
Minimum Pairwise Difference Excluding Ambiguous

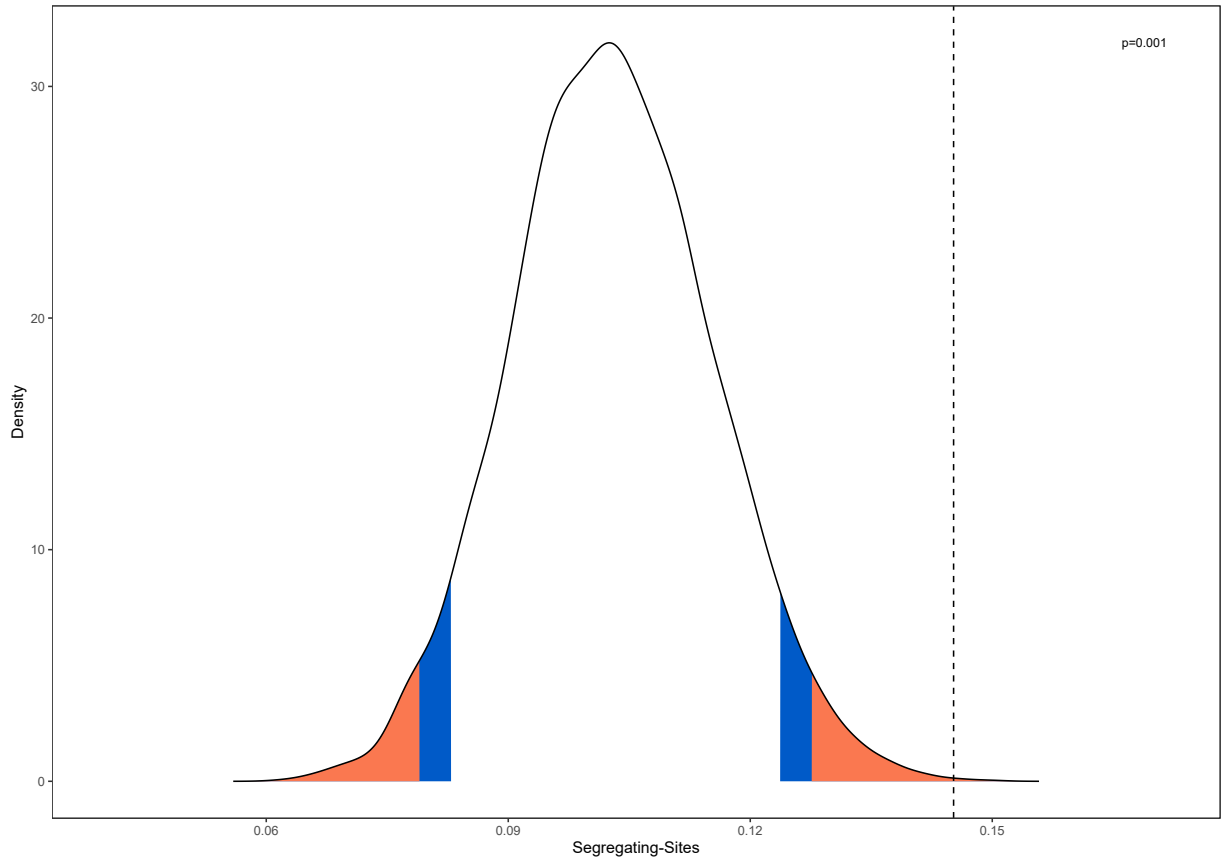

p=0.322
Density
1.0
0.5
0.0
0
2
4
6
Number of Invariable Block

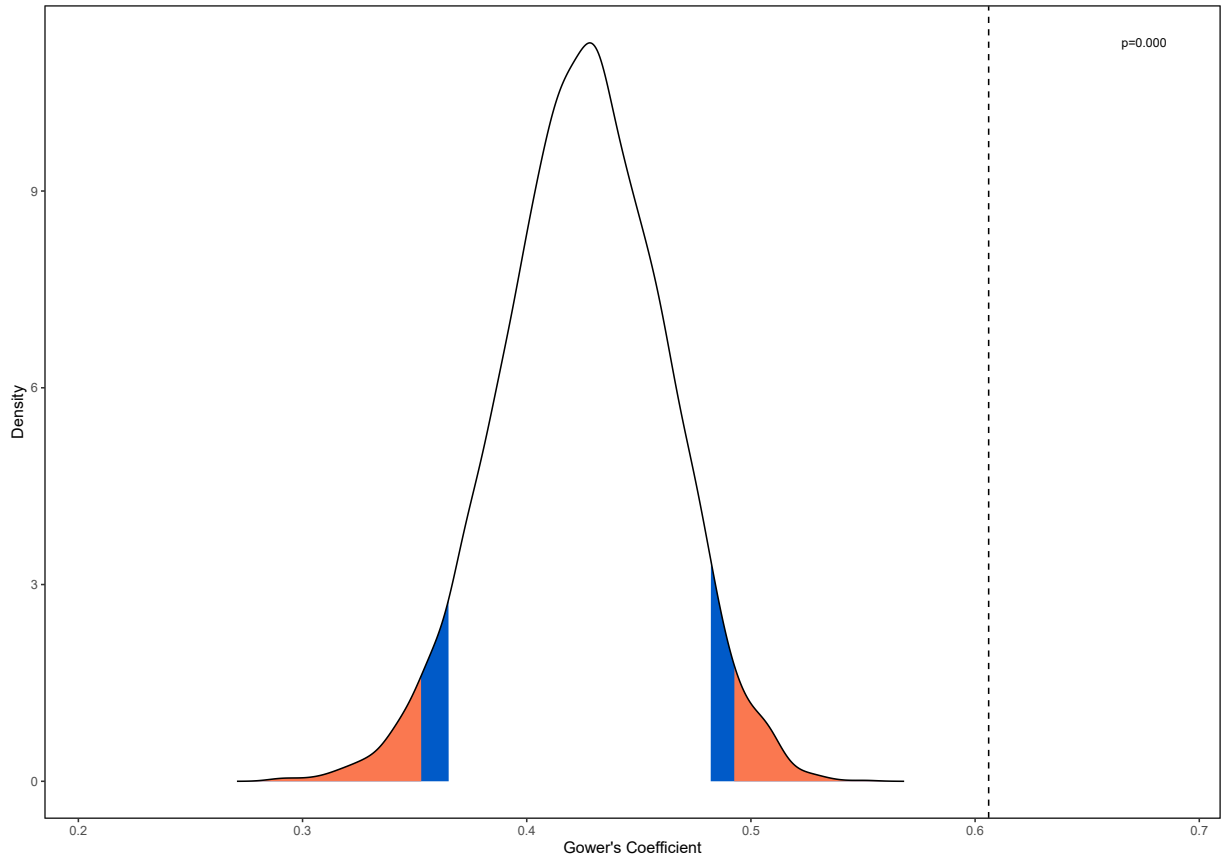

p=0.000
0.08
0.06
0.04
0.02
0.00
Density
0
20
40
Number of Invariable Block Excluding Ambiguous

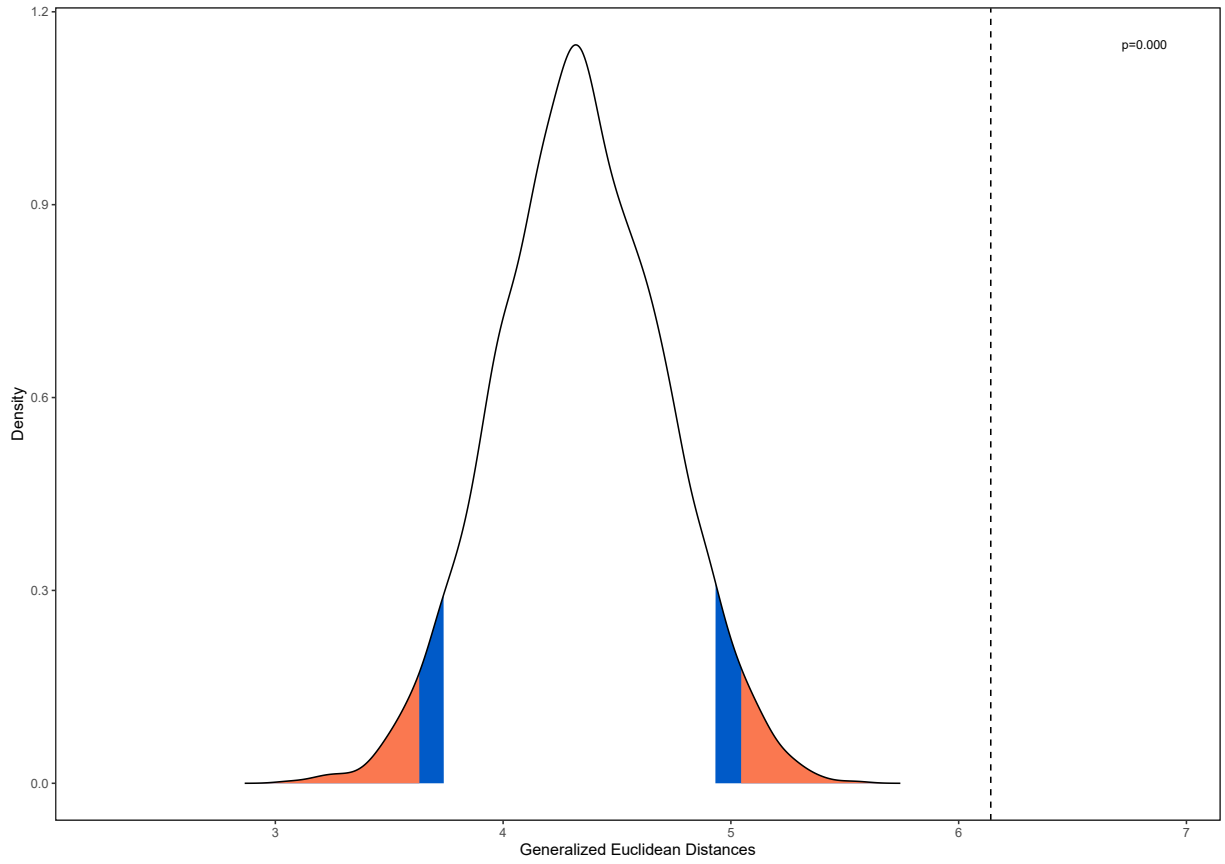

p=0.370
900
600
300
0
Density
0.1775
0.1800
0.1825
0.1850
Theta


p=0.000
1.2
0.8
0.4
0.0
Density
156
159
162
165
Tajima's-D


p=0.001
Density
1.5
1.0
0.5
0.0
-2.5
-2.0
-1.5
-1.0
-0.5
Tajima's-Pi


p=0.001
30
20
10
0
Density
0.06
0.09
0.12
0.15
Segregating-Sites


p=0.000
Density
0
3
6
9
0.2
0.3
0.4
0.5
0.6
0.7
Gower's Coefficient


p=0.000
Density
1.2
0.9
0.6
0.3
0.0
3
4
5
6
7
Generalized Euclidean Distances

Appendix B

Summary Statistics in Inference PPS

Listed below are the summary statistics that were used for the Inference PPS.

- Mean Robinson Foulds - The Robinson-Foulds metric shows the topological distance scaled by the branch lengths on the trees. RF is calculated from each tree in the posterior distribution and a mean is taken to obtain this metric.
- Quantile 25 - It is the Robinson-Foulds score for the first quartile of the trees in the posterior distribution.
- Quantile 50 - It is the Robinson-Foulds score for the median of the trees in the posterior distribution.
- Quantile 75 - It is the Robinson-Foulds score for the third quantile of the trees in the posterior distribution.

- Quantile 99 - It is the Robinson-Foulds score for the 99^th percentile of the trees in the posterior distribution.

- Quantile 999 - It is the Robinson-Foulds score for the 999^th permillage of the trees in the posterior distribution.

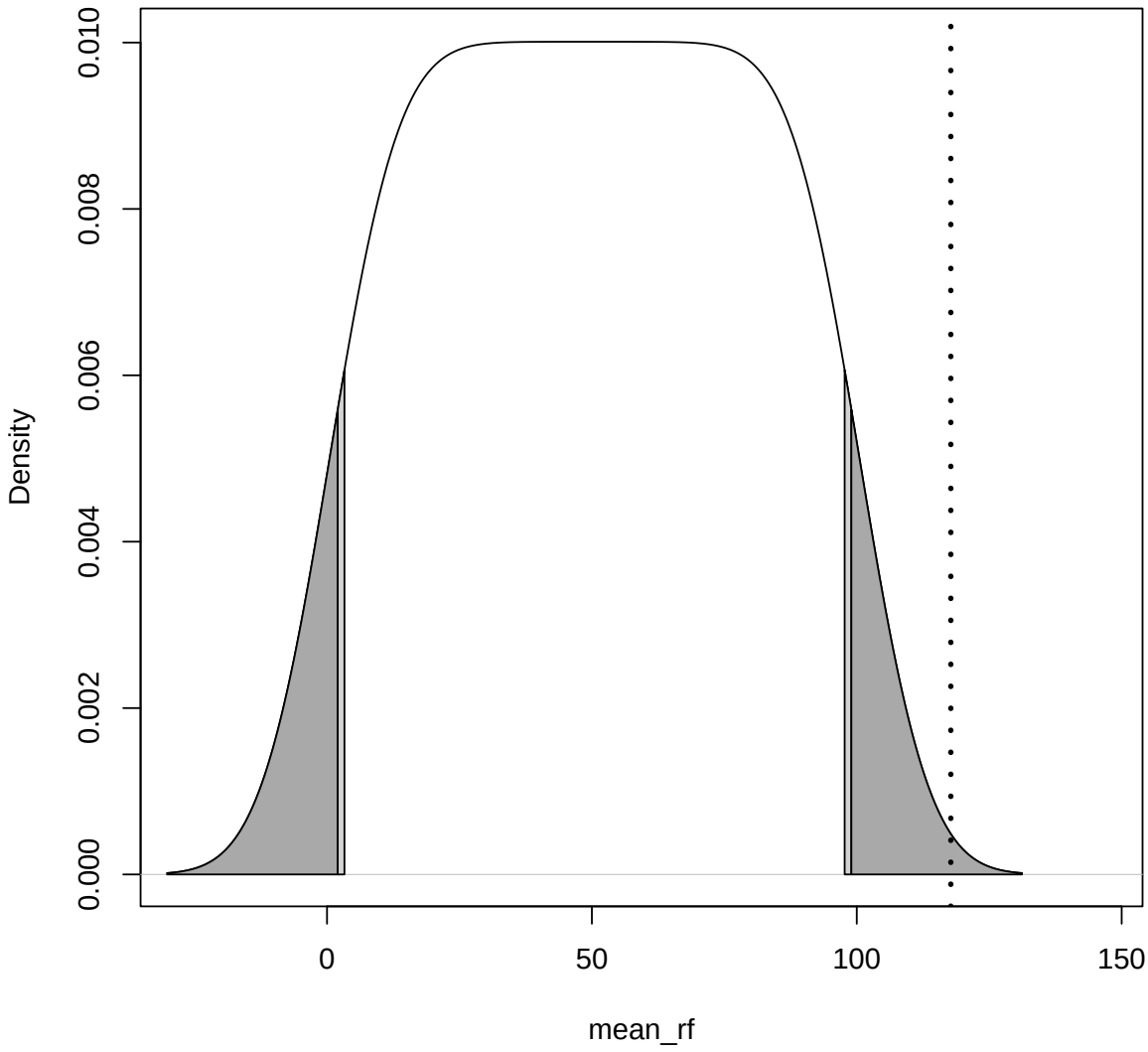
- Mean Tree Length - This test statistic is the sum of all branch lengths which is averaged across the posterior distribution of the trees. This will indicate the number of evolutionary changes in the inferred tree.

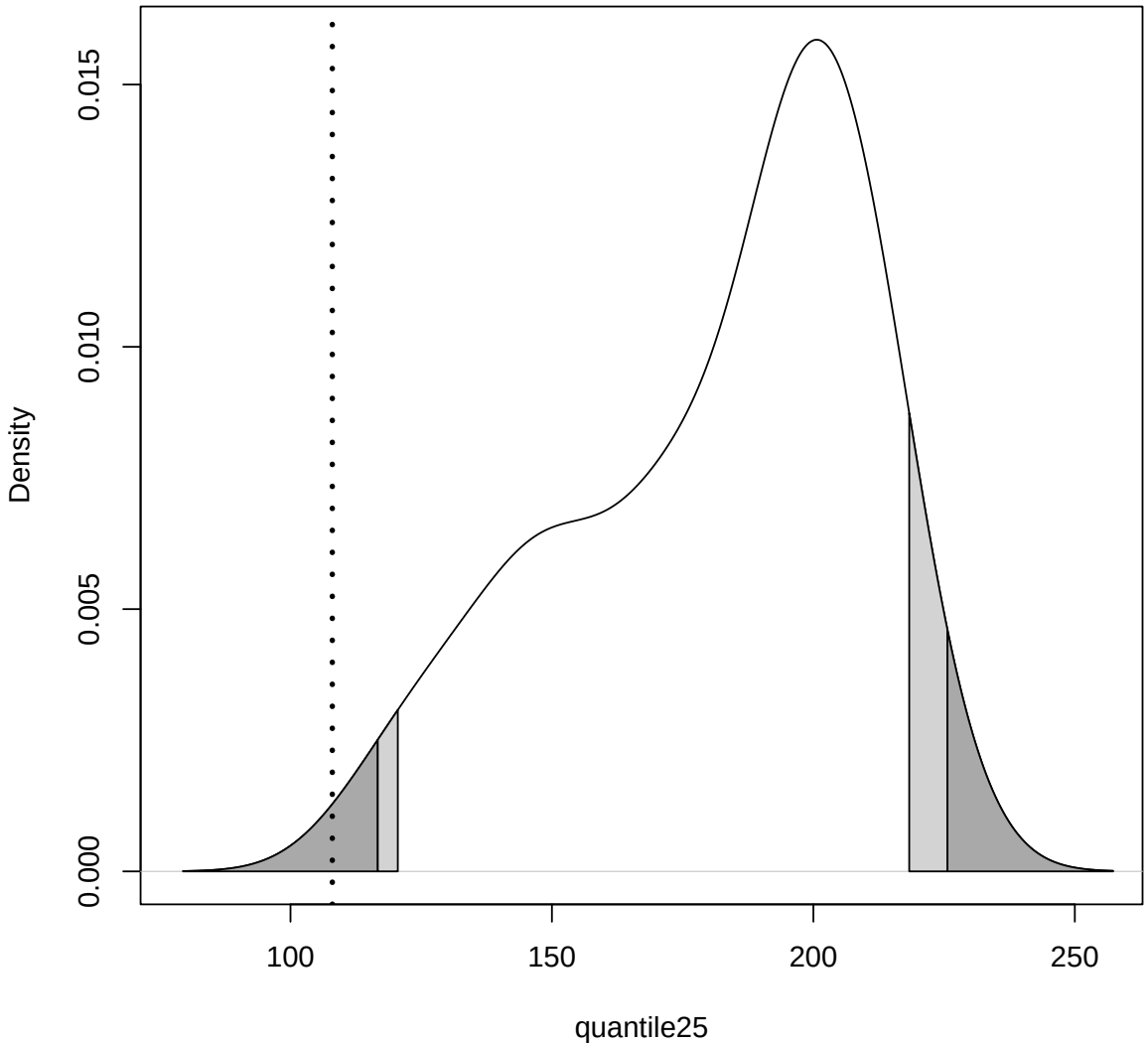
- Variance in Tree Length - This test statistic is designed to capture the uncertainty in the posterior distribution of the branch lengths.

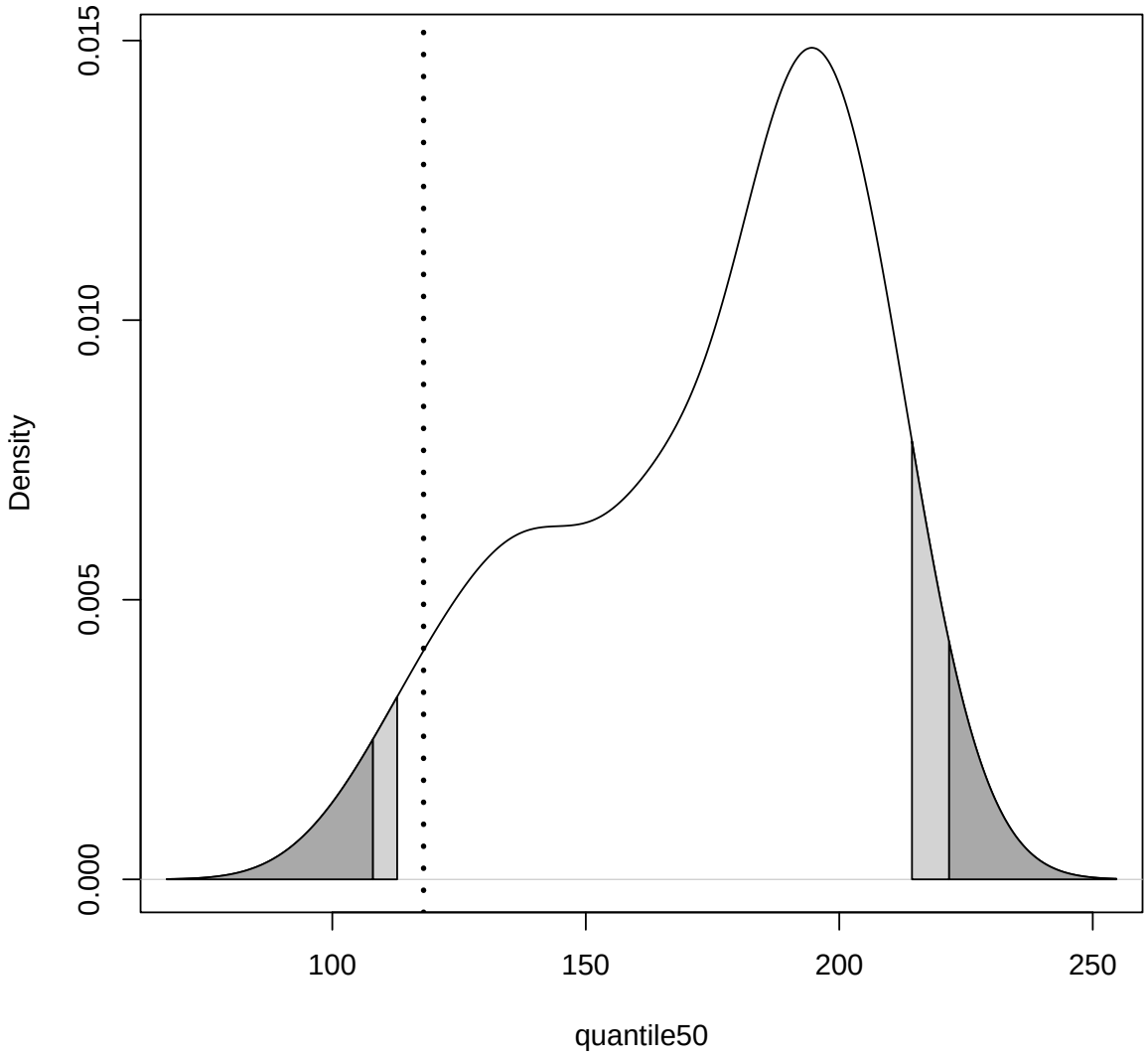
- Entropy - The entropy of the tree captures the information gain between the marginal prior and the posterior distribution of the tree topologies.

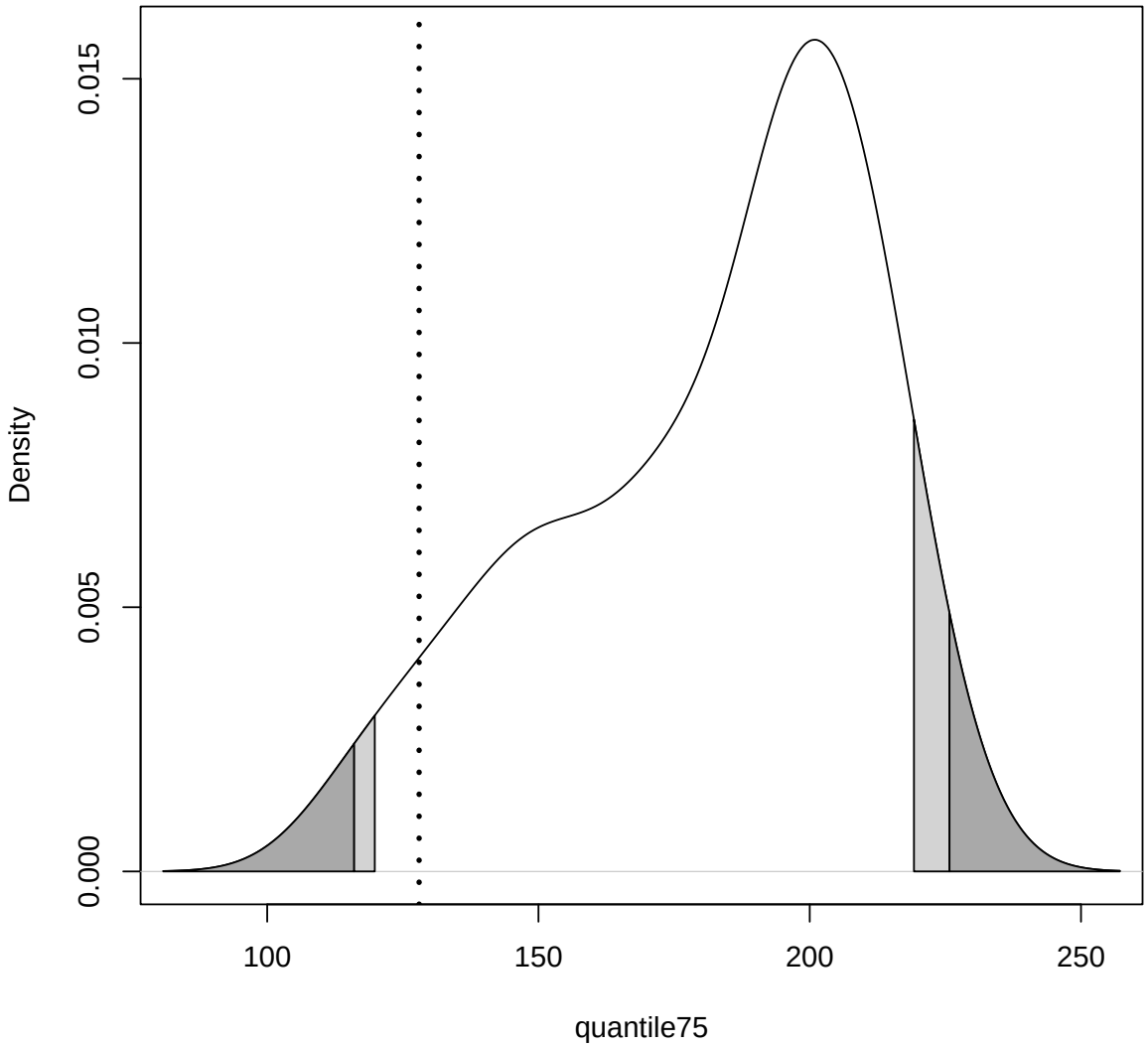
The figures for above listed test statistics are in the following pages.

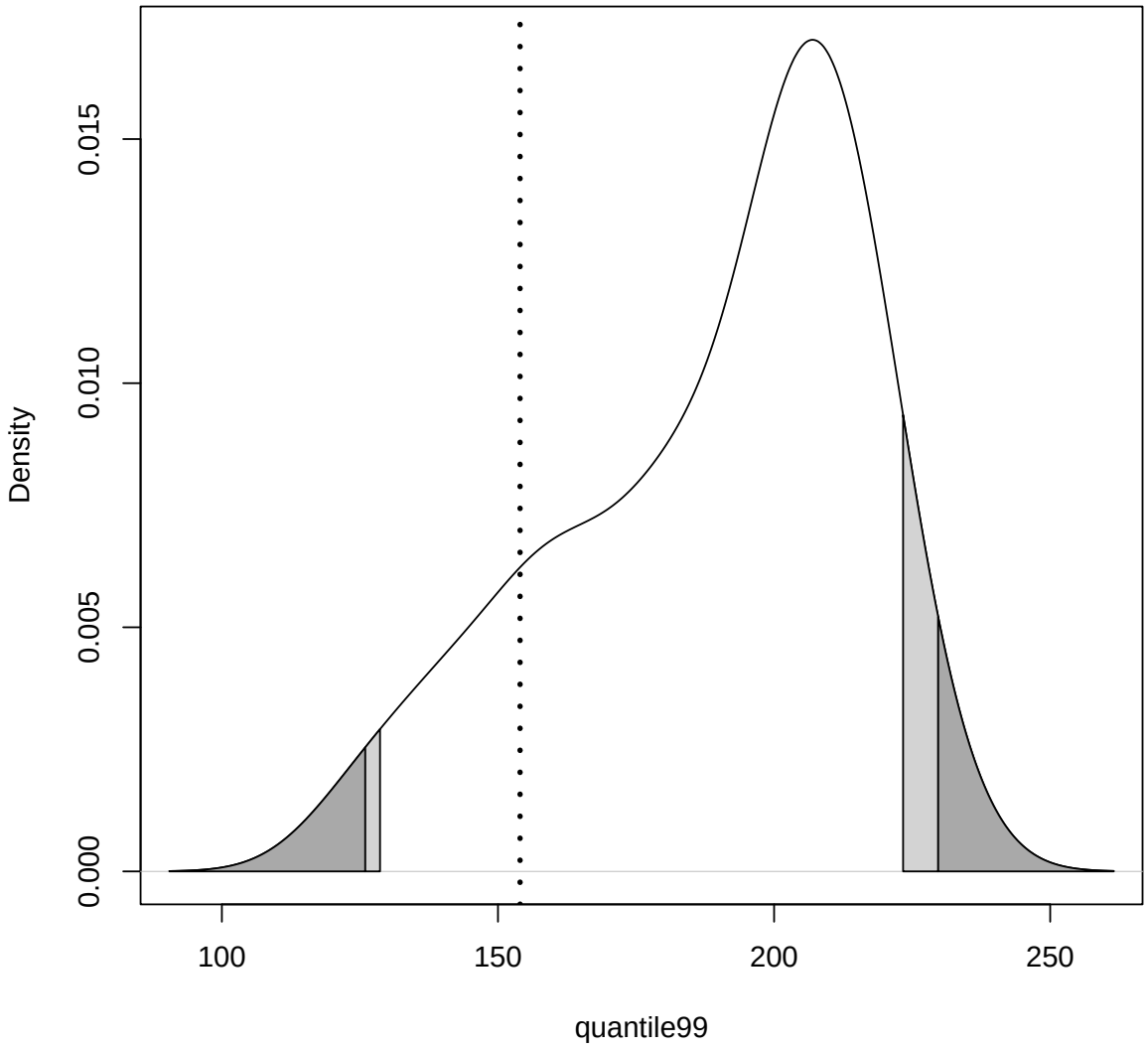
For the Mk model, the density graphs for all the summary statistics are as follows.

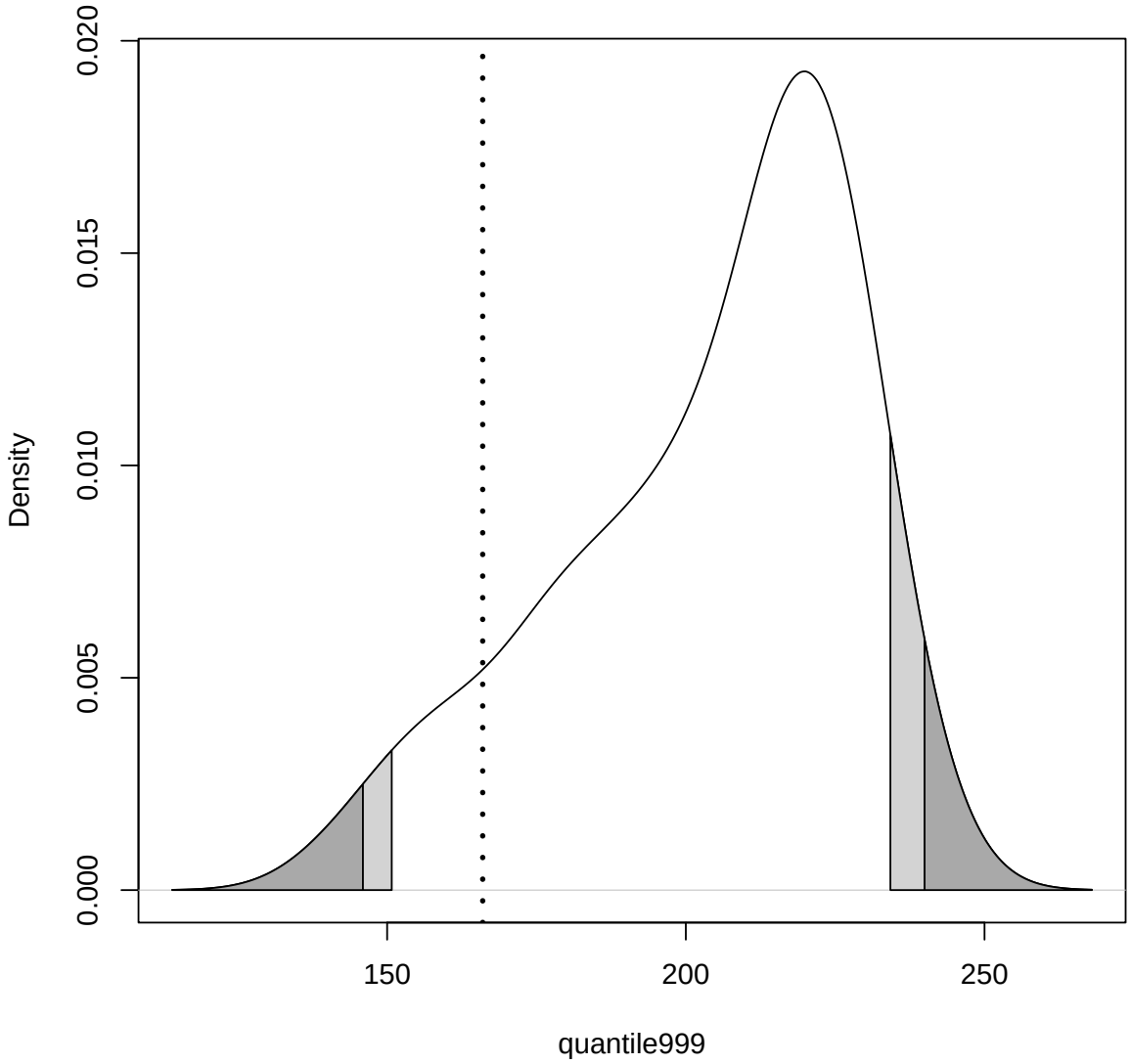

Density
0.000
0.002
0.004
0.006
0.008
0.010
0
50
100
150
mean_rf

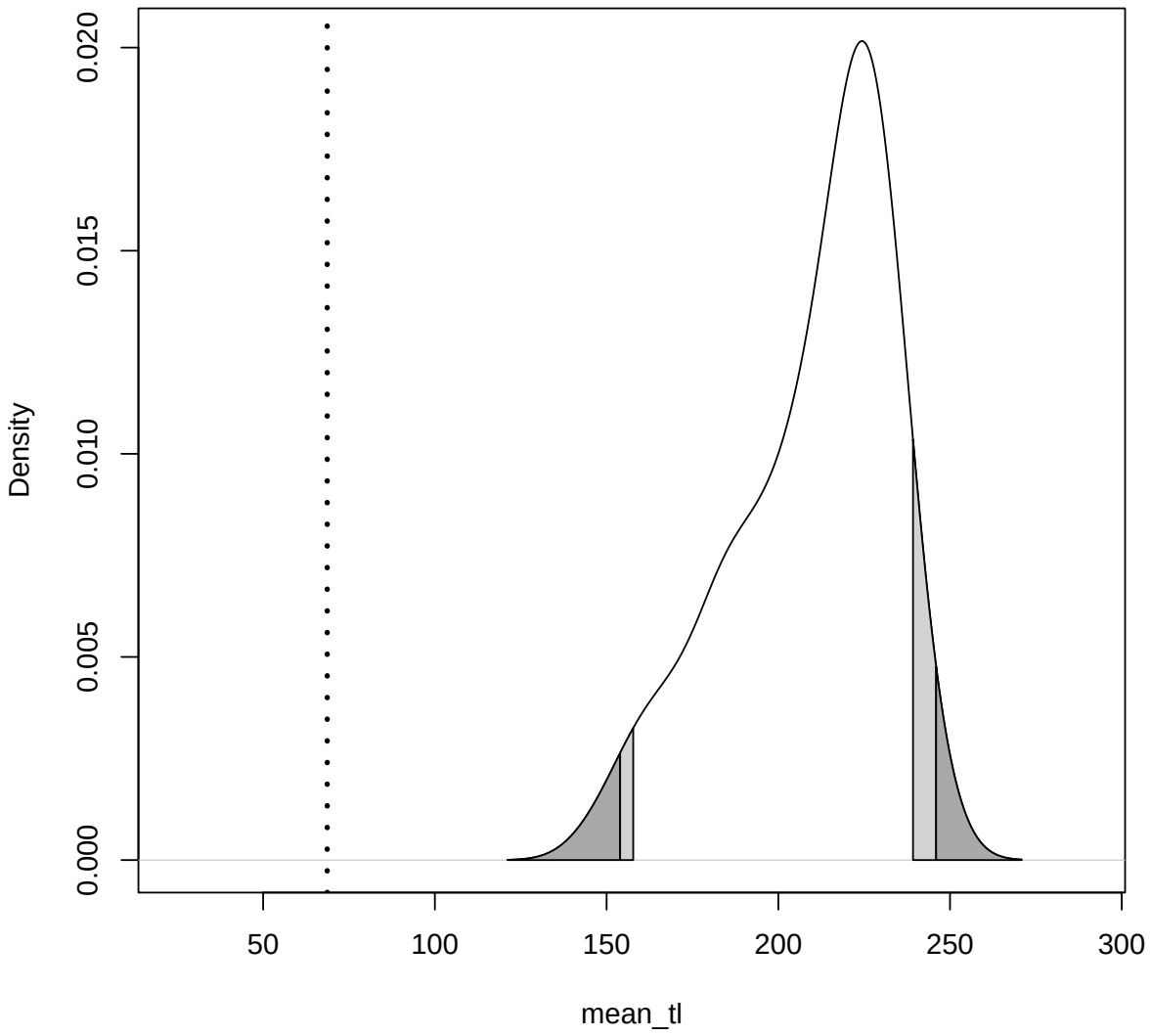

Density
0.000
0.005
0.010
0.015
100
150
200
250
quantile25

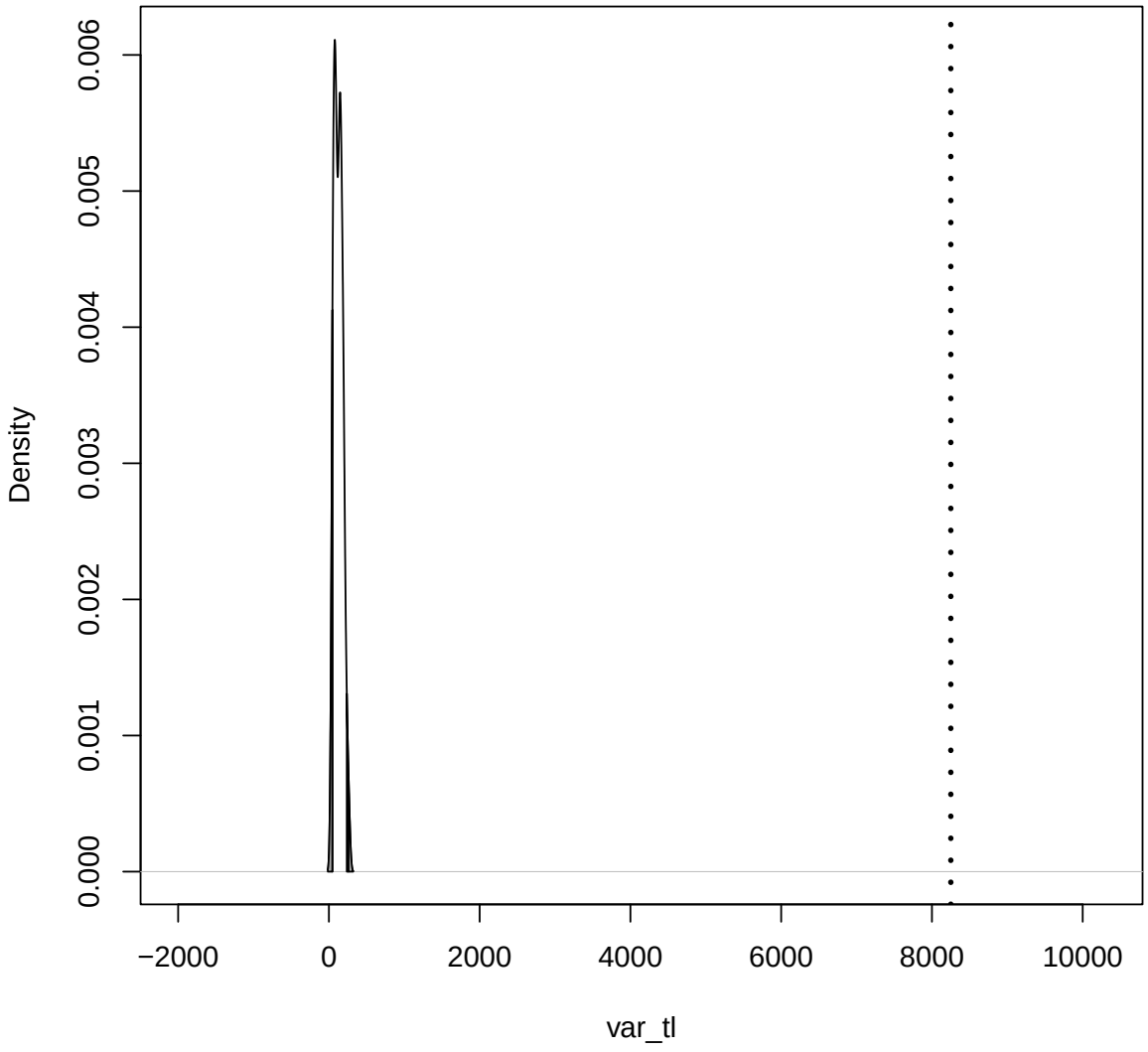

0.015
0.010
0.005
0.000
Density
100
150
200
250
quantile50

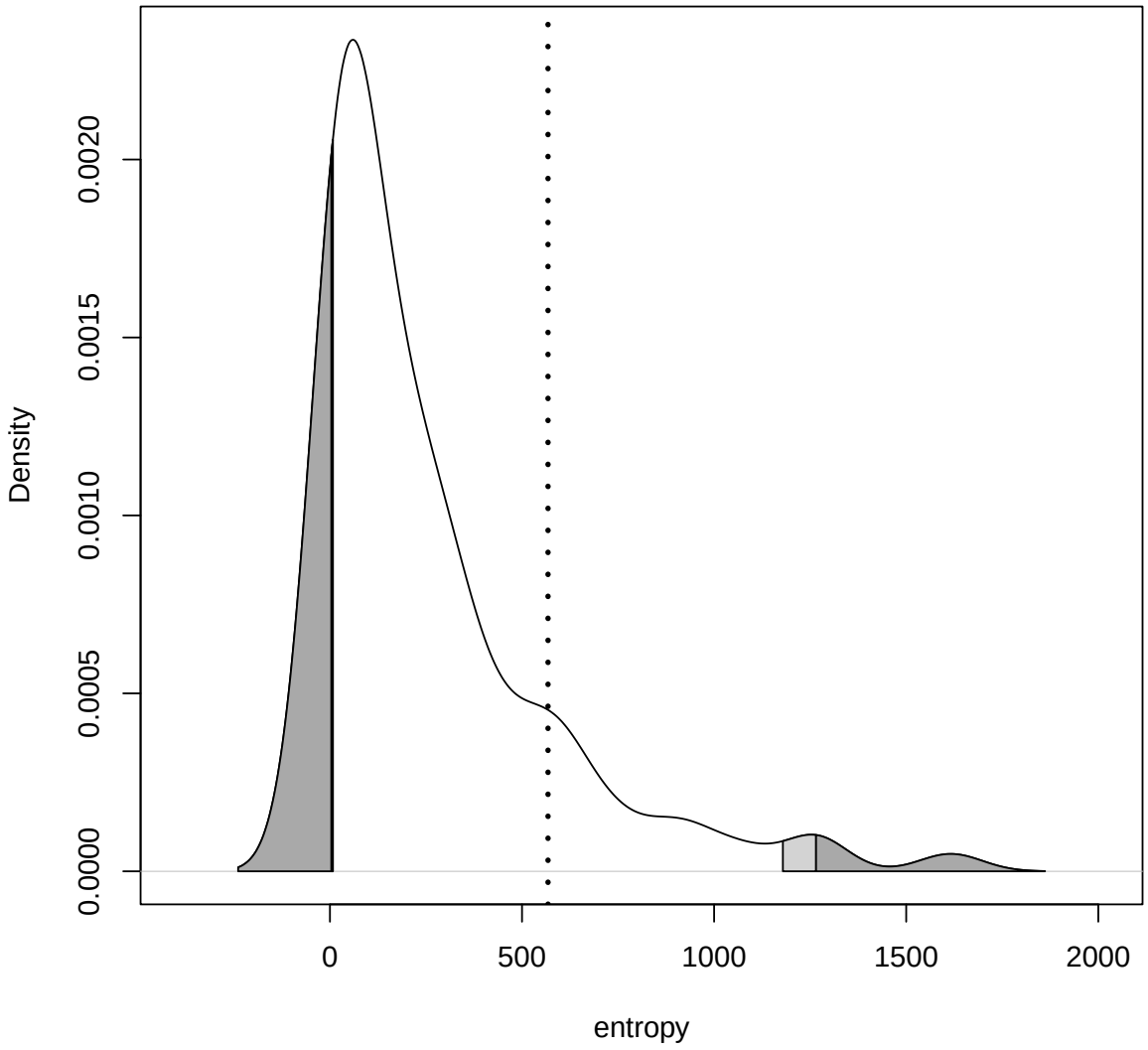

Density
0.000
0.005
0.010
0.015
100
150
200
250
quantile75


Density
0.000
0.005
0.010
0.015
100
150
200
250
quantile99

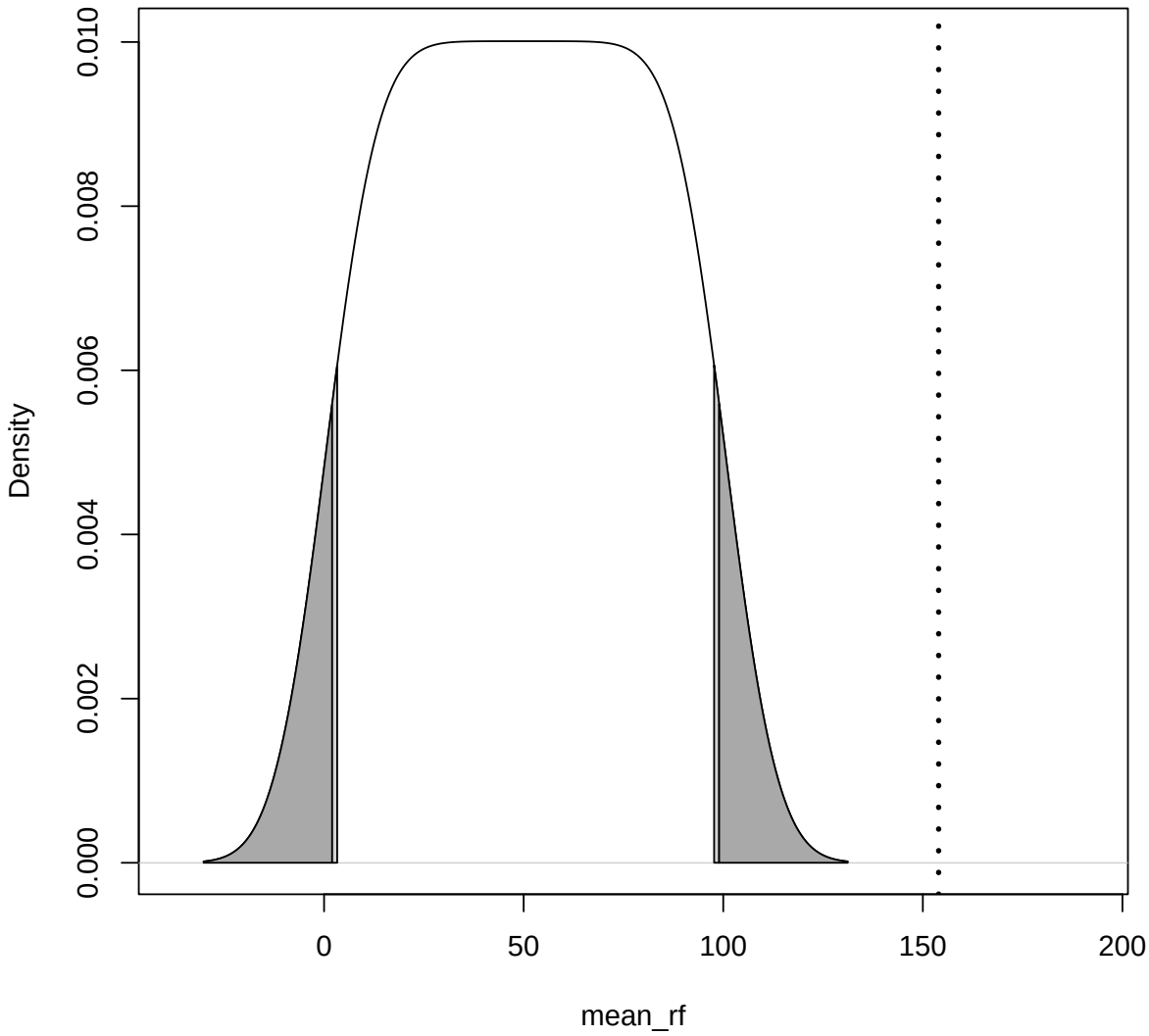

0.020
0.015
0.010
0.005
0.000
Density
150
200
250
quantile999

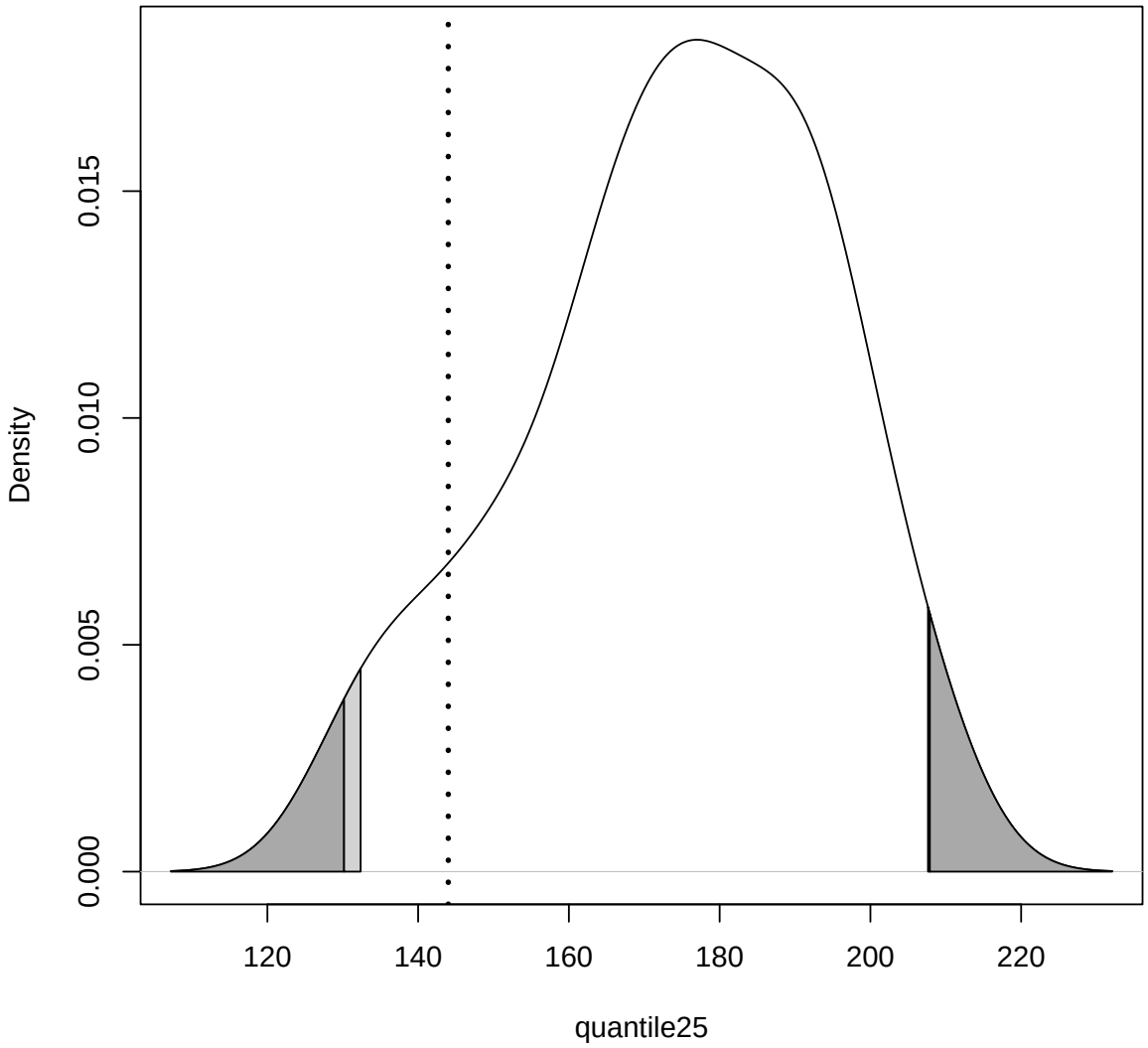

Density
0.000
0.005
0.010
0.015
0.020
50
100
150
200
250
300
mean_tl

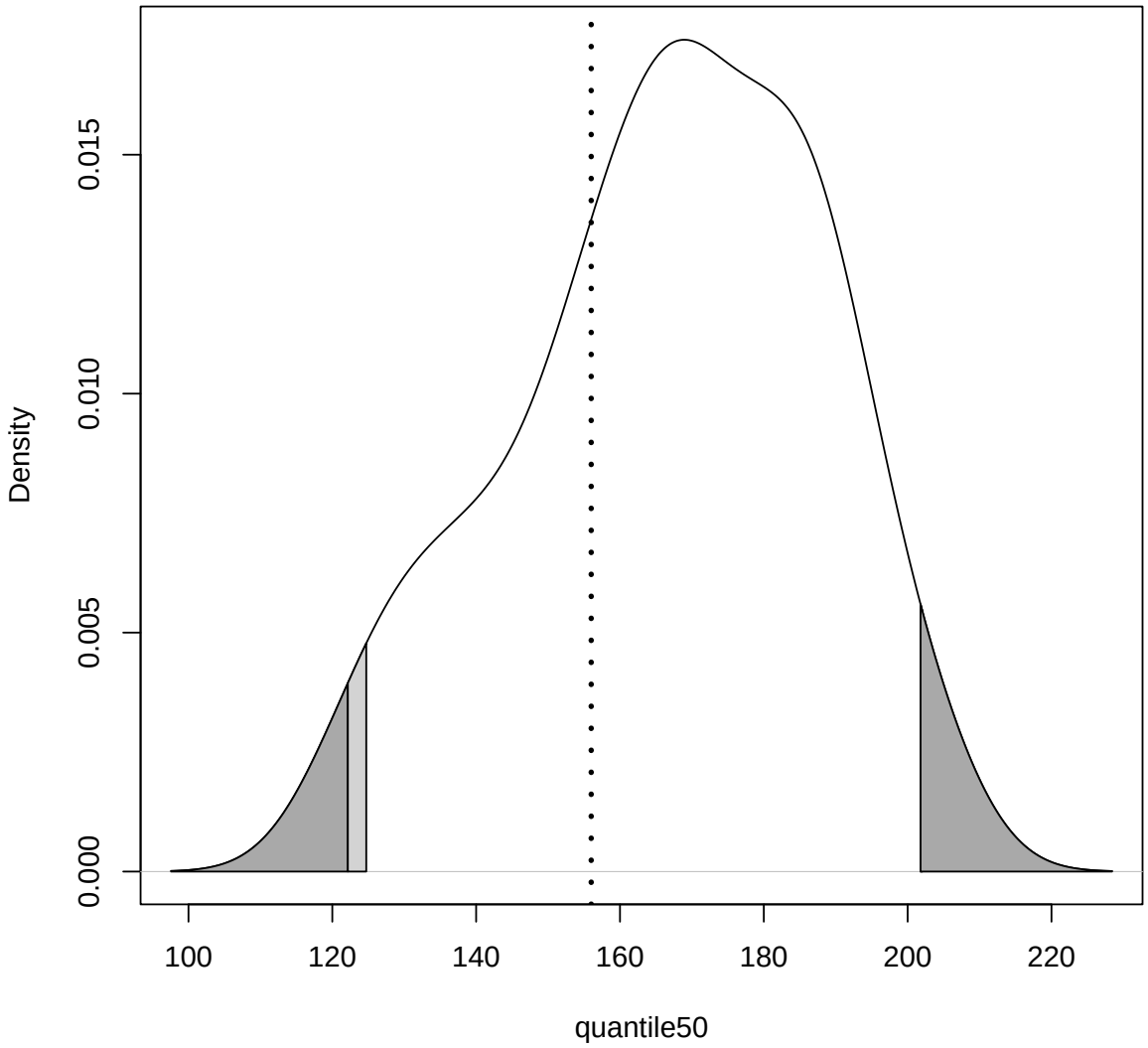

Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
−2000
0
2000
4000
6000
8000
10000
var_tl

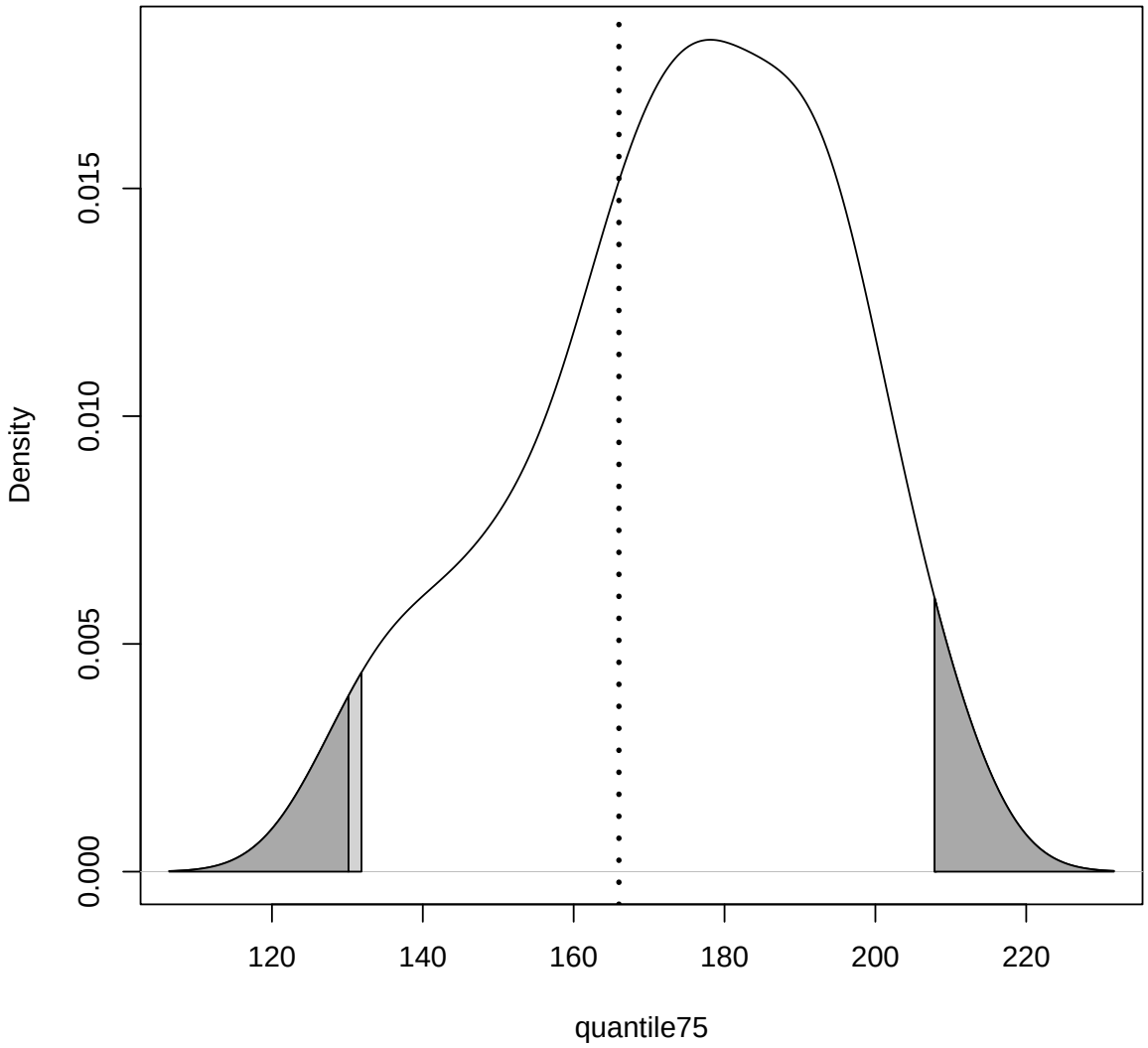

Density
0.0000
0.0005
0.0010
0.0015
0.0020
0
500
1000
1500
2000
entropy

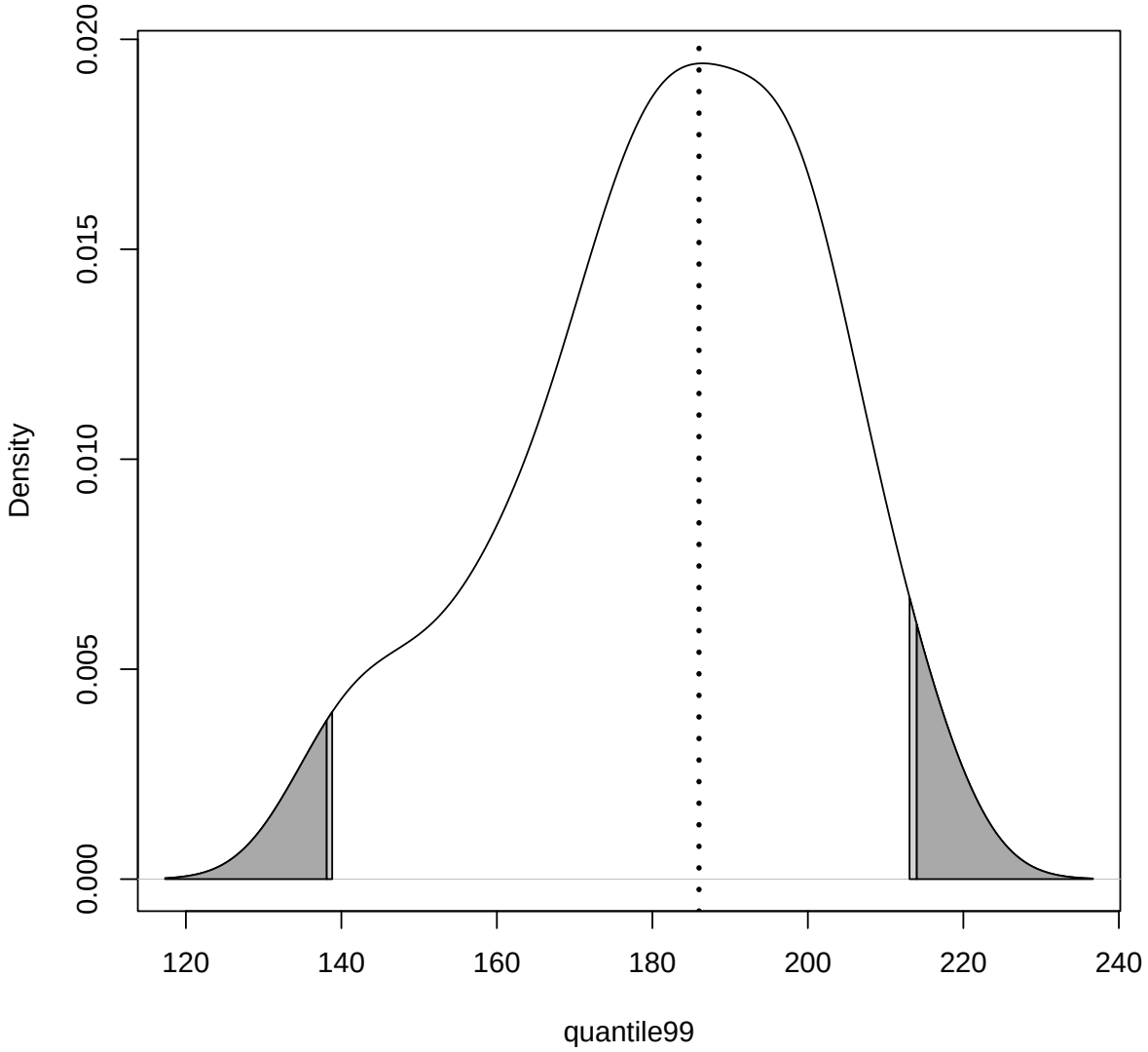
For the SHDM model, the density graphs for all the summary statistics are as follows.

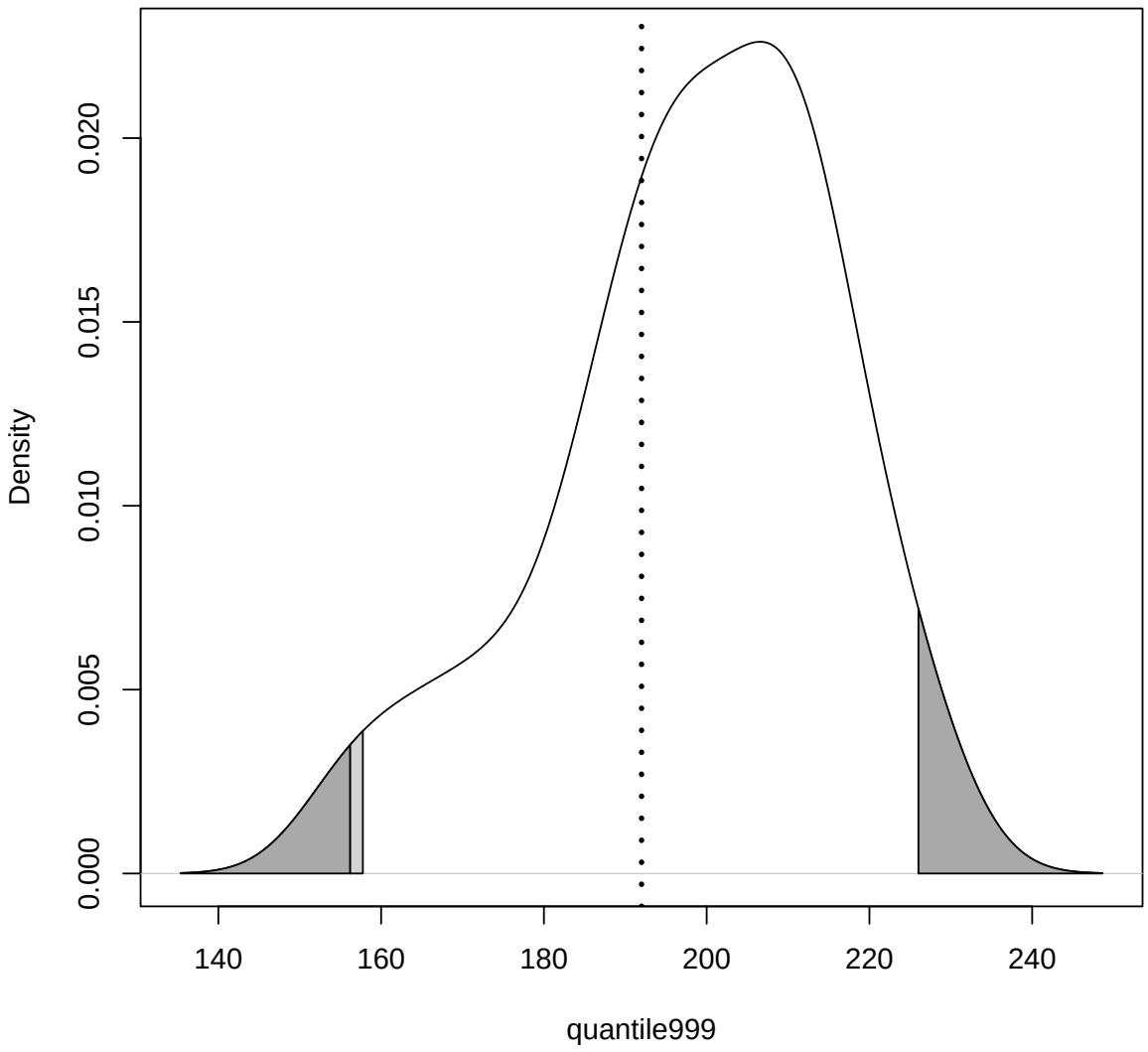

Density
0.000
0.002
0.004
0.006
0.008
0.010
0
50
100
150
200
mean_rf

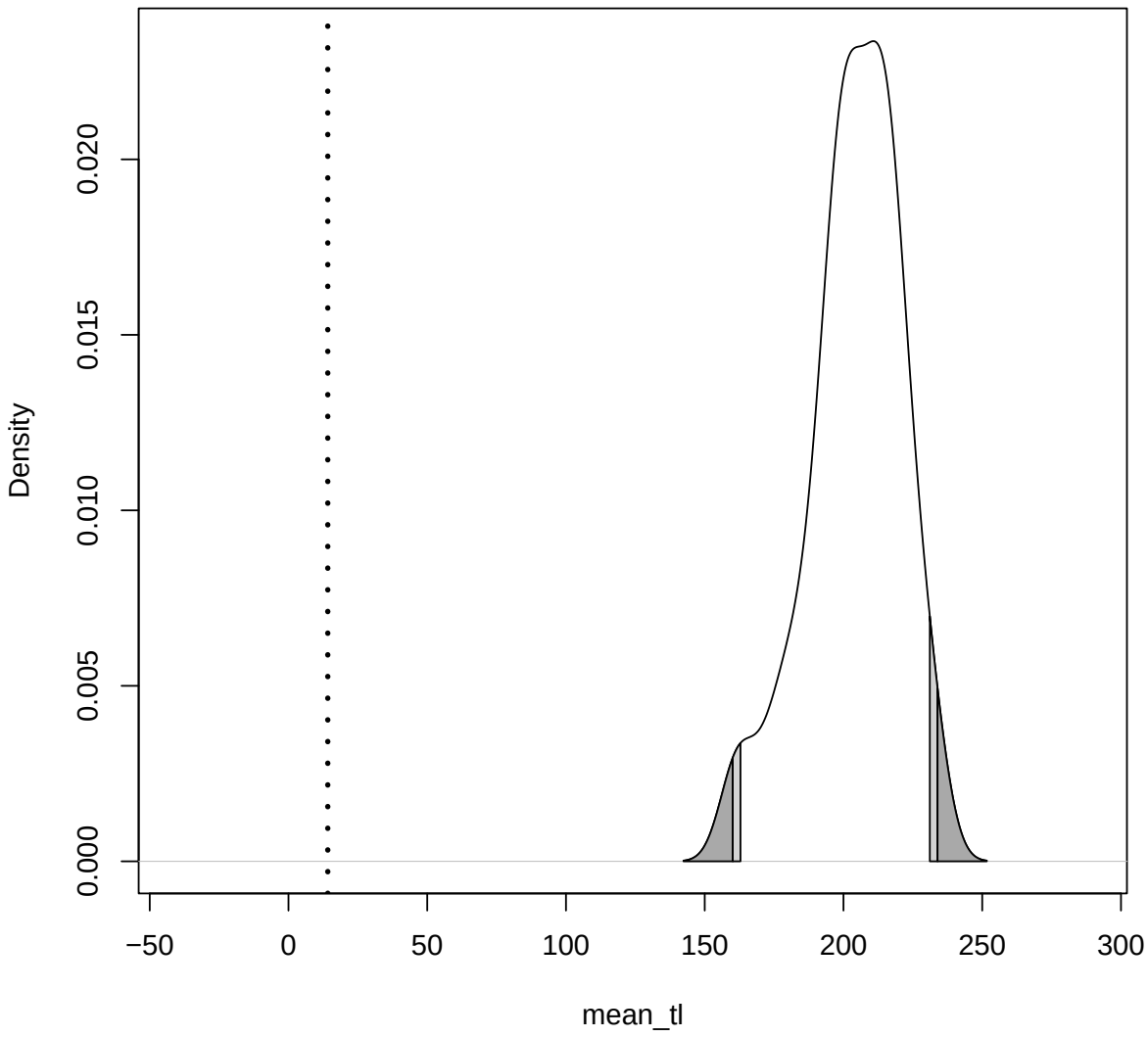

Density
0.000
0.005
0.010
0.015
120
140
160
180
200
220
quantile25

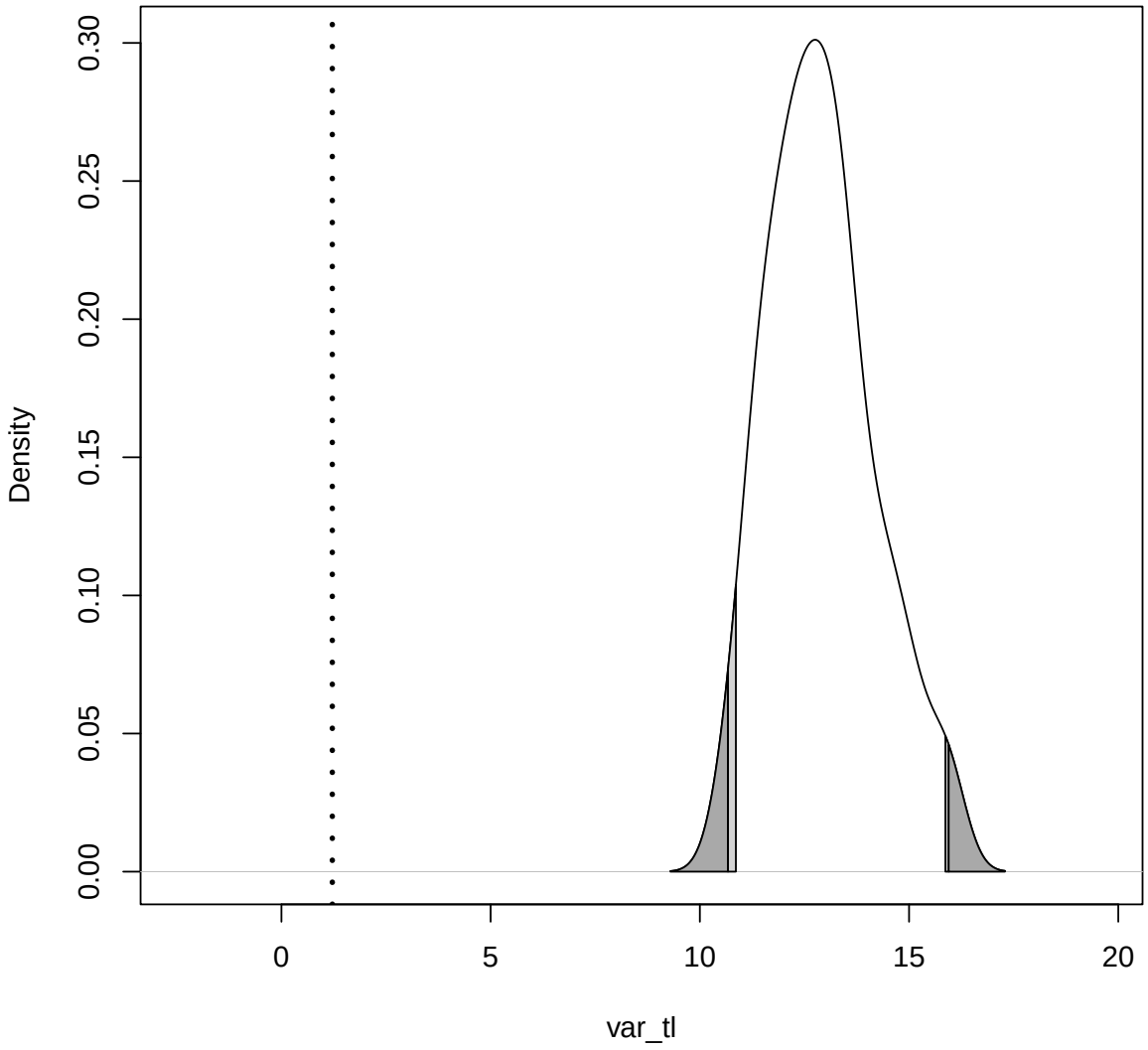

Density
0.000
0.005
0.010
0.015
100
120
140
160
180
200
220
quantile50

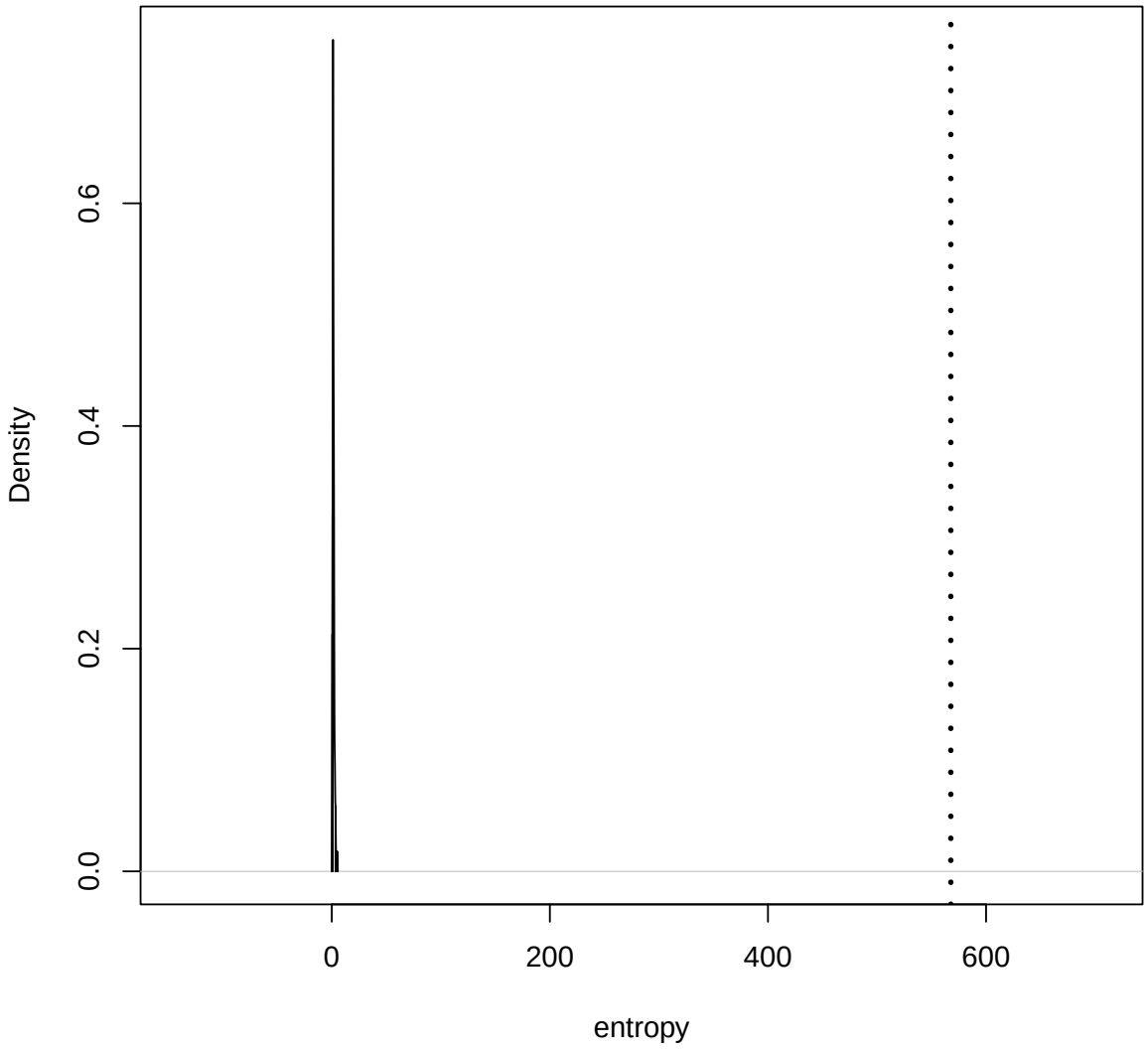

Density
0.000
0.005
0.010
0.015
120
140
160
180
200
220
quantile75


Density
0.000
0.005
0.010
0.015
0.020
120
140
160
180
200
220
240
quantile99


Density
0.000
0.005
0.010
0.015
0.020
140
160
180
200
220
240
quantile999


Density
0.000
0.005
0.010
0.015
0.020
−50
0
50
100
150
200
250
300
mean_tl


Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
5
10
15
20
var_tl


Density
0.0
0.2
0.4
0.6
0
200
400
600
entropy