Bayesian Phylogenetic Inference using RevBayes:

Model Adequacy Testing

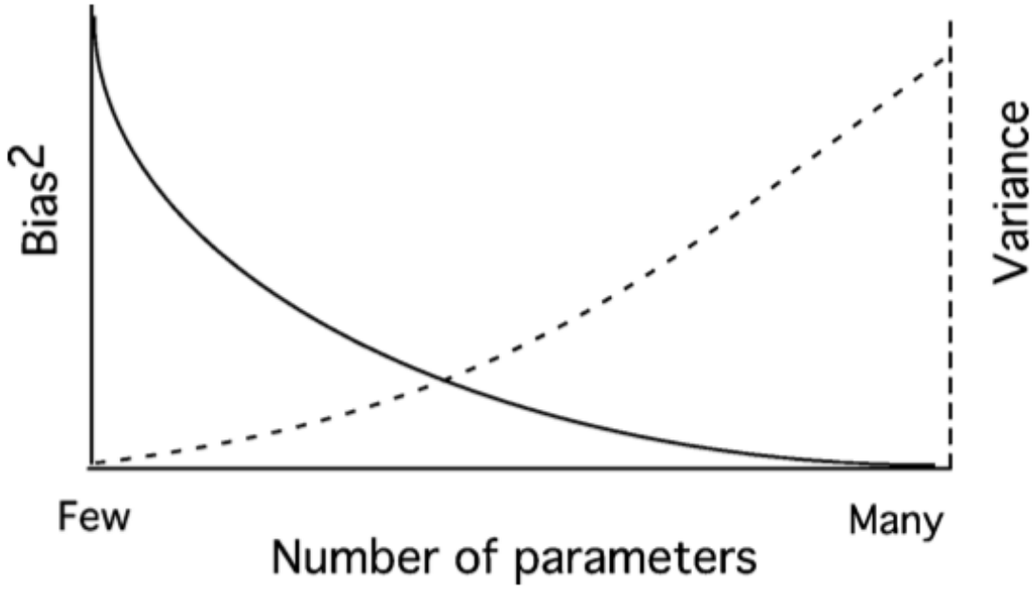
Sebastian Höhna

Division of Evolutionary Biology
Ludwig-Maximilians Universität, München

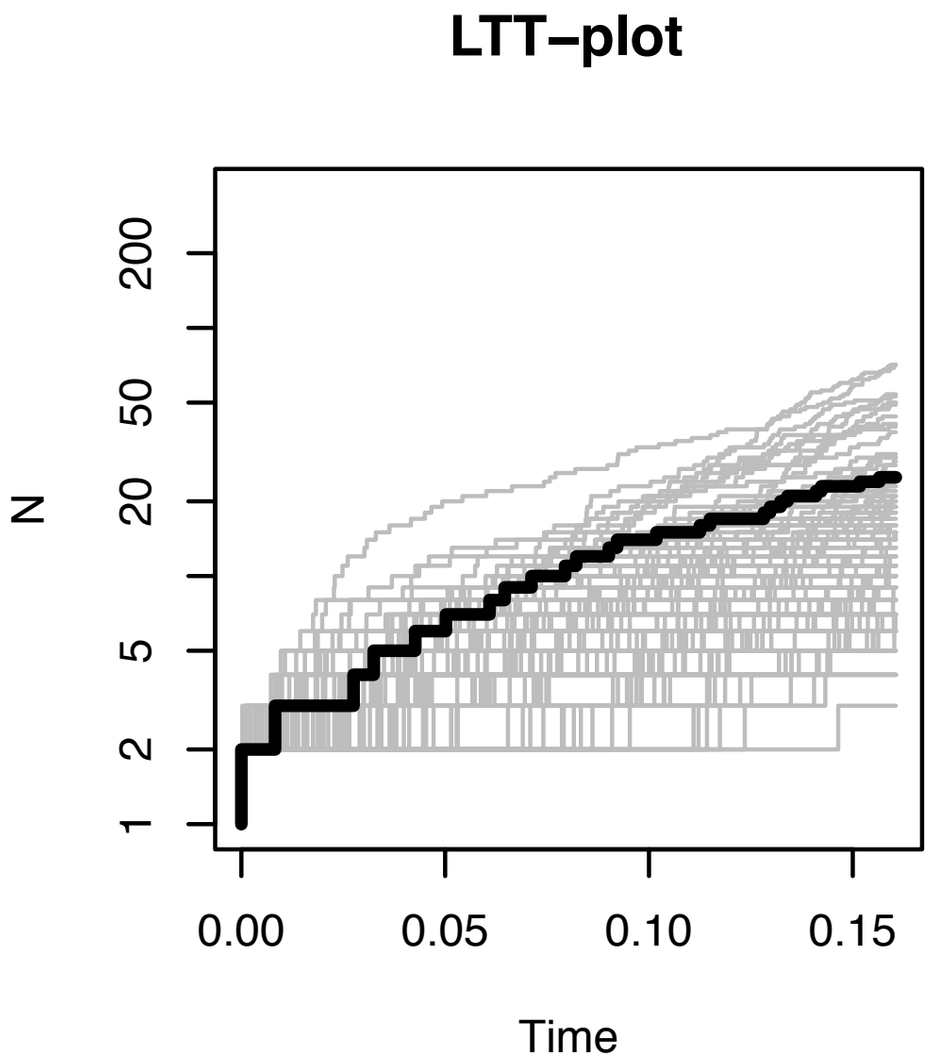
Model Specification Issues

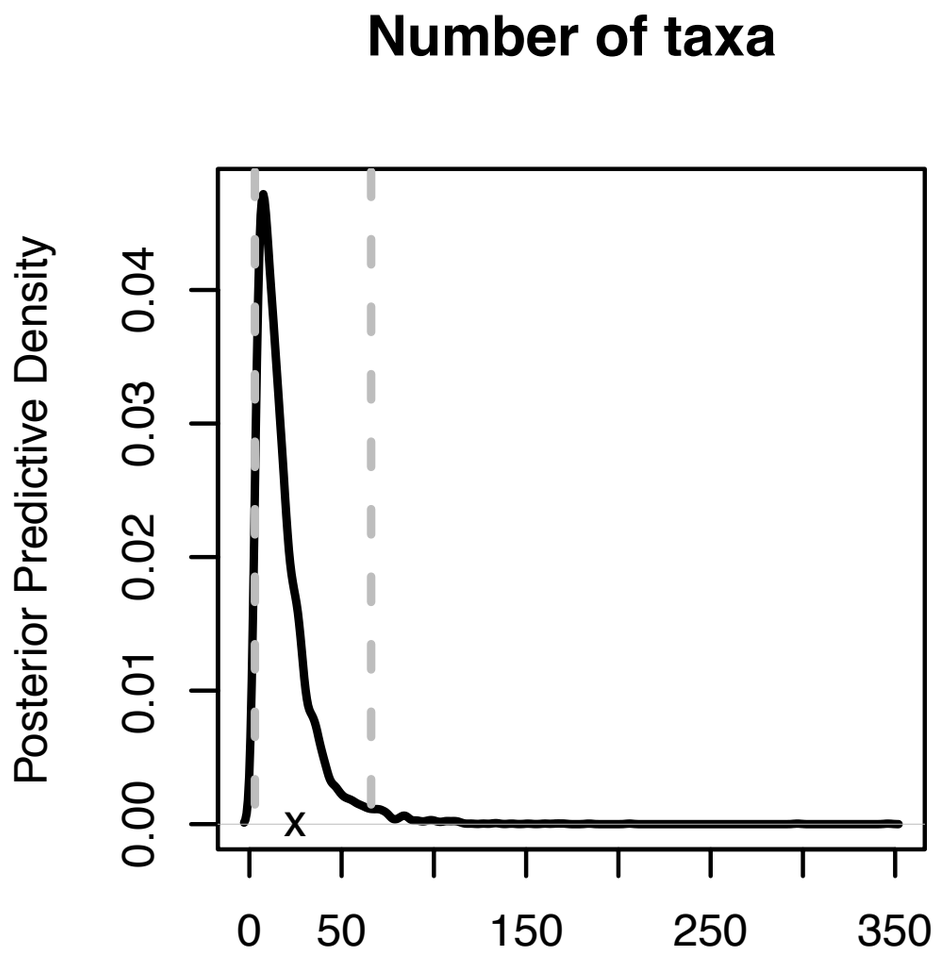
Model selection, adequacy, and related issues

The model is central to parametric estimation of phylogeny: an under parameterized model will cause estimates to be biased (e.g., under estimation of branch lengths, topological error, inflated estimates of nodal support...); however, an over parameterized model will inflate estimation error (error variance, etc.).

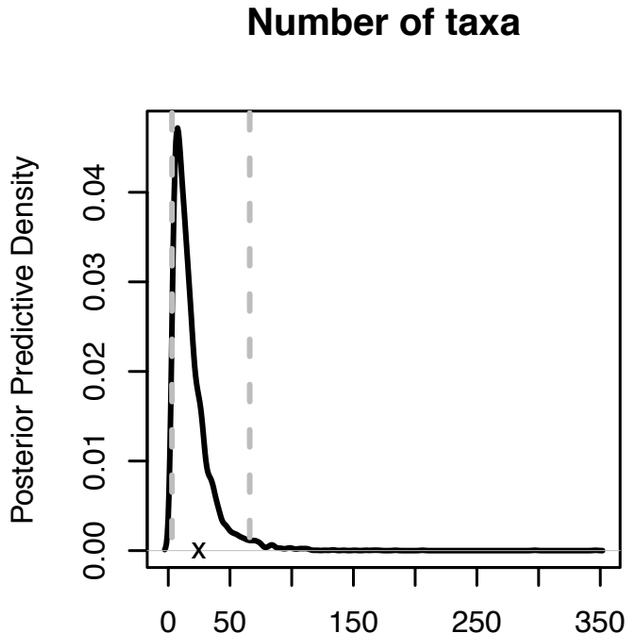


What is the probability of the data under the model? Is the data an outlier and improbable under this model?





Consider that our data are the number of species and the model is a birth-death process for species diversification. This example shows that the data is not unexpected under the model.



In simple cases, like the number of species under a birth-death model, we can compute the distribution of the data analytically.

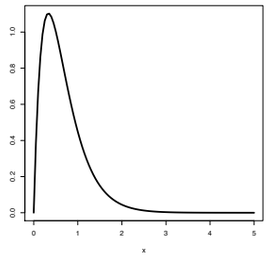
But what should we do in complex situations, like multiple sequence alignments, to compare the data?

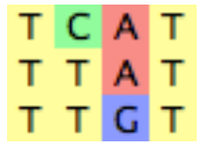
We use simulations and summary statistics!

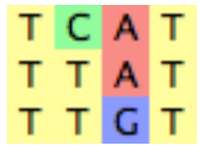
Simulation

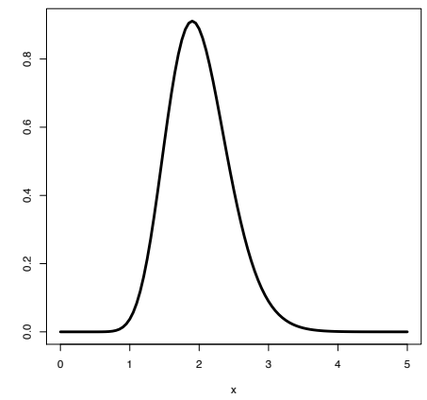
Posterior Estimates

Summary









Σ

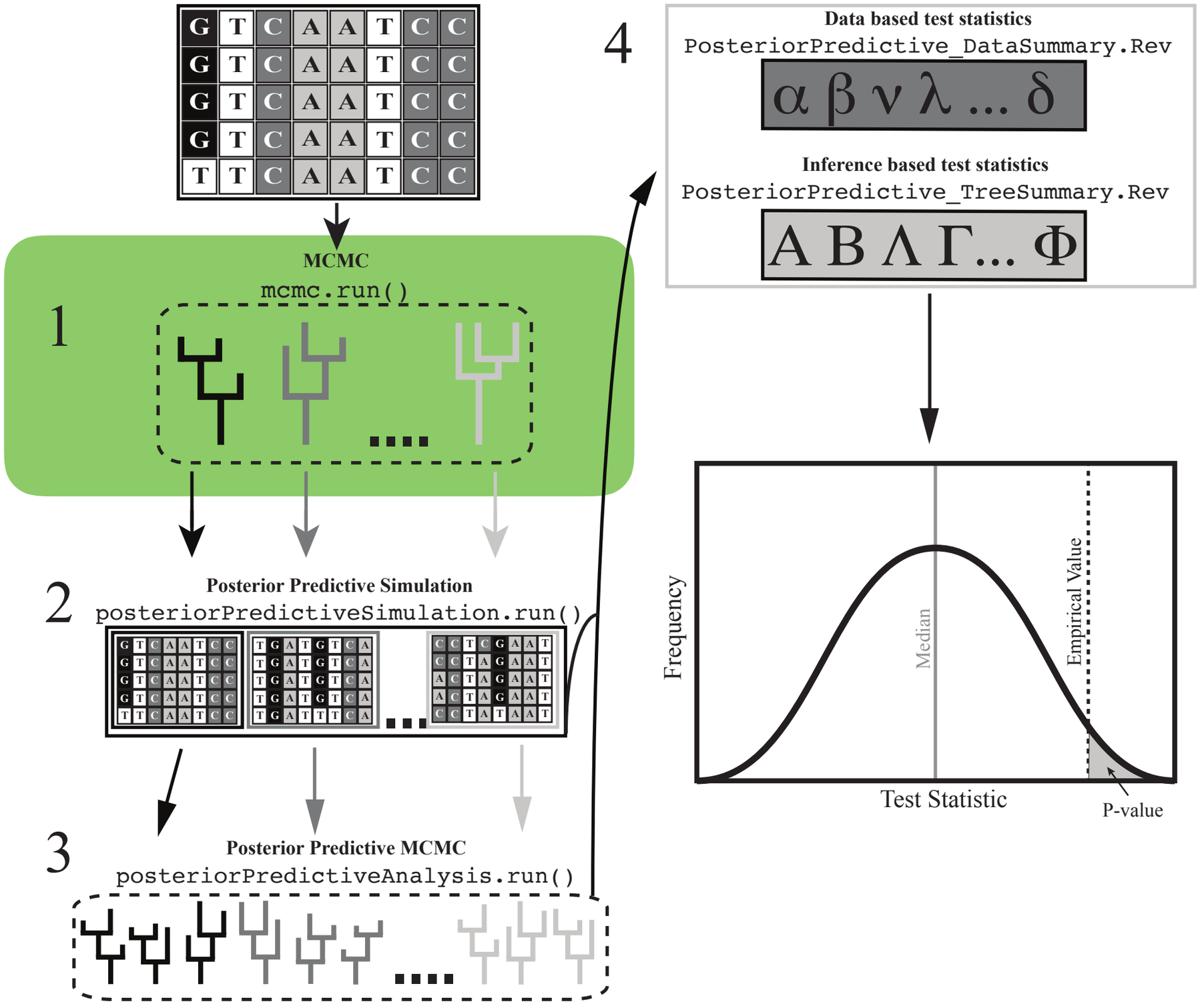
Σ

Σ

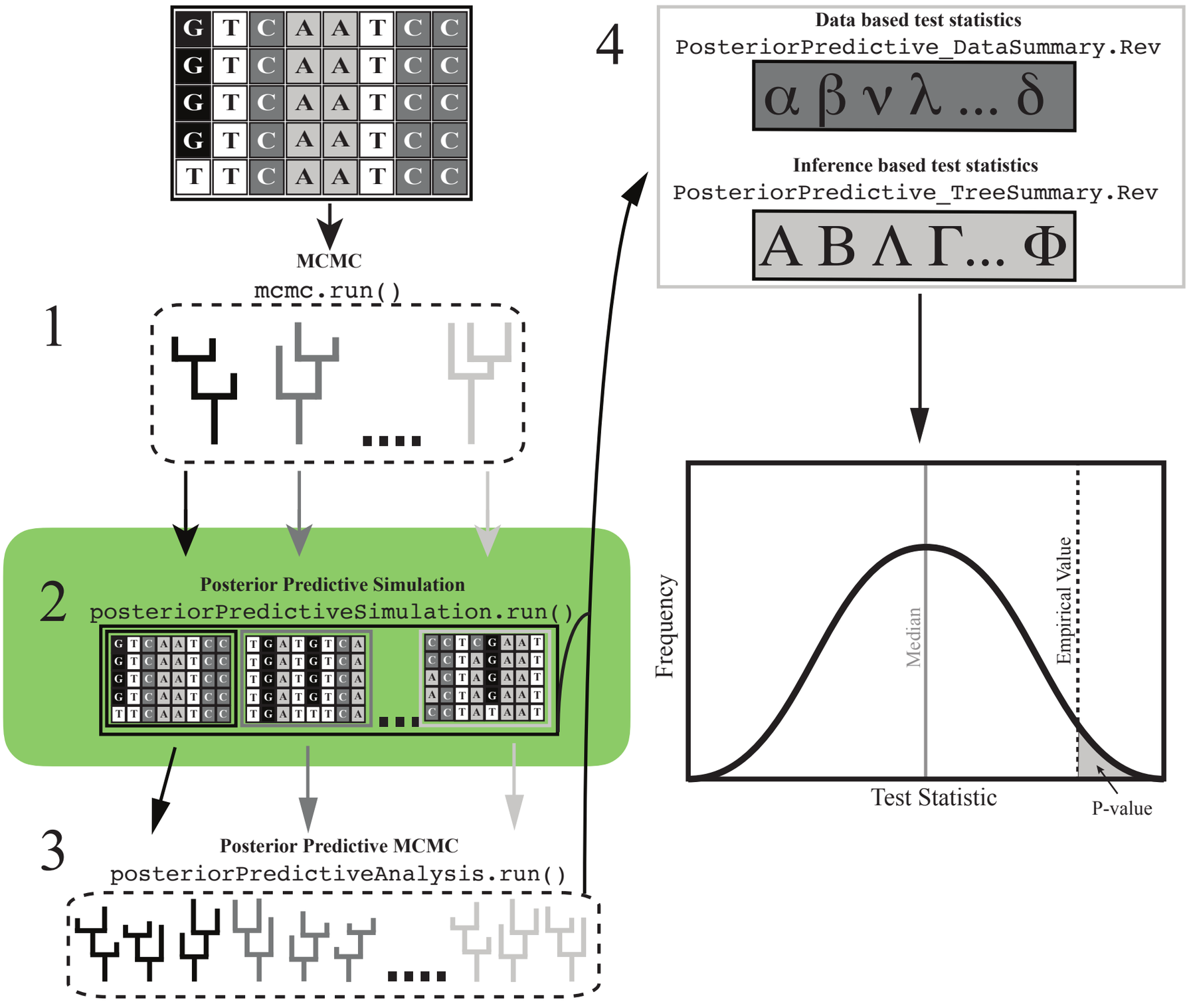
Σ

Σ

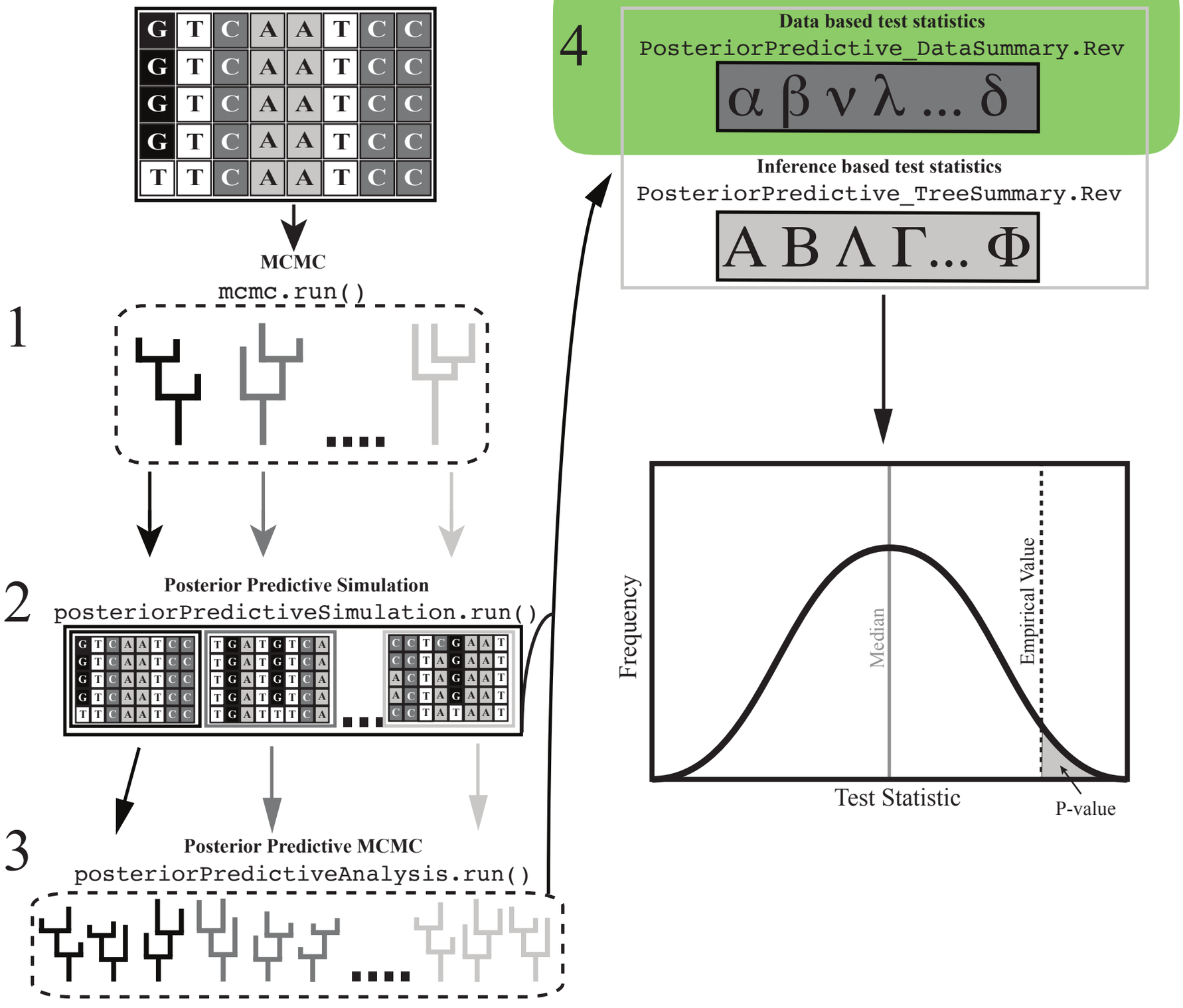
P^3: Phylogenetic Posterior Prediction



P^3: Phylogenetic Posterior Prediction



P^3: Phylogenetic Posterior Prediction



Example data-based summary statistics

Number of invariant sites:

Tests the rate of evolutions.

Maximum , Minimum & Mean GC content:

Tests for base composition and outliers in base composition

Variance of GC content:

Variation on base composition and heterogeneity of substitution process

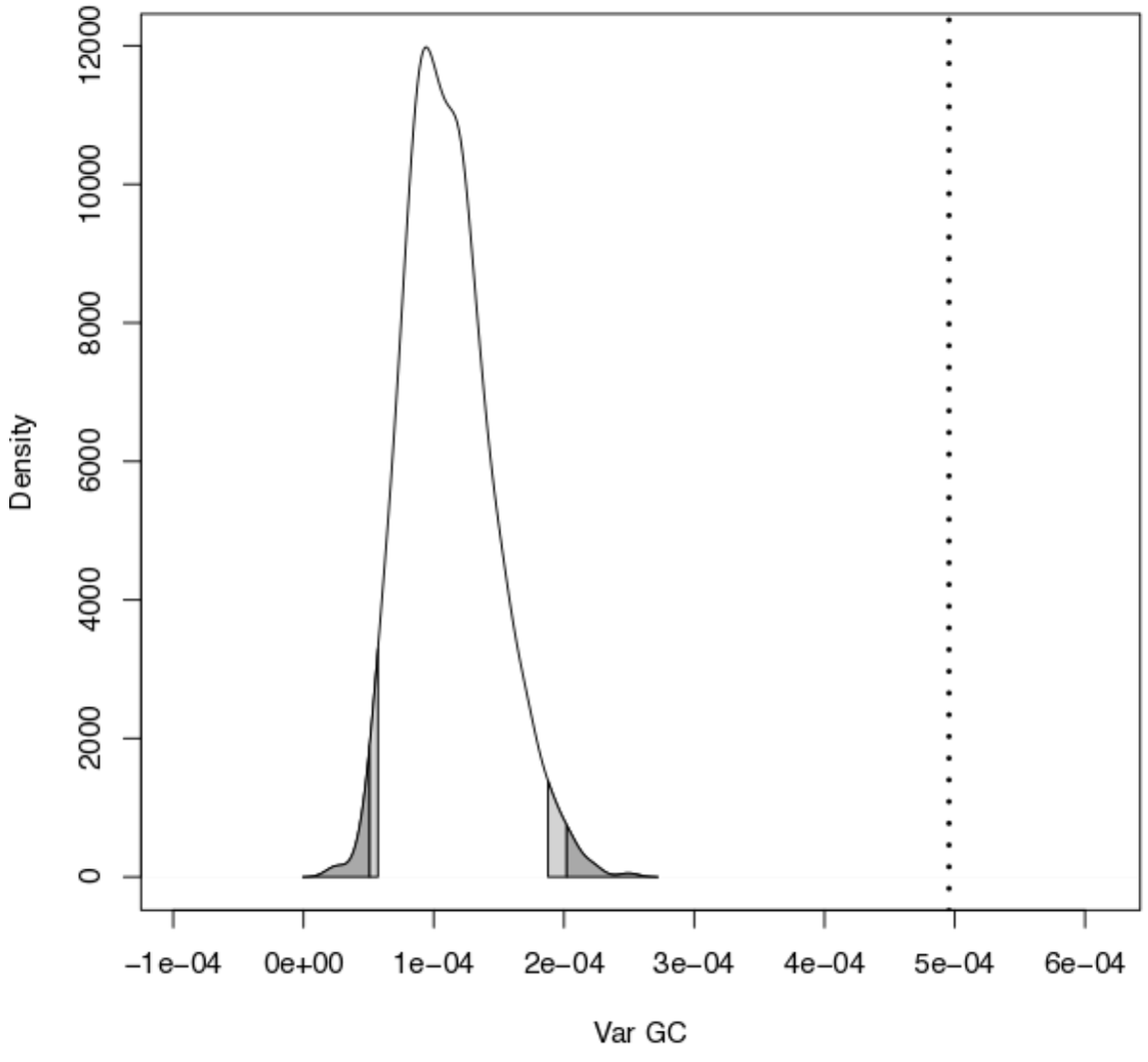
Maximum pairwise distance:

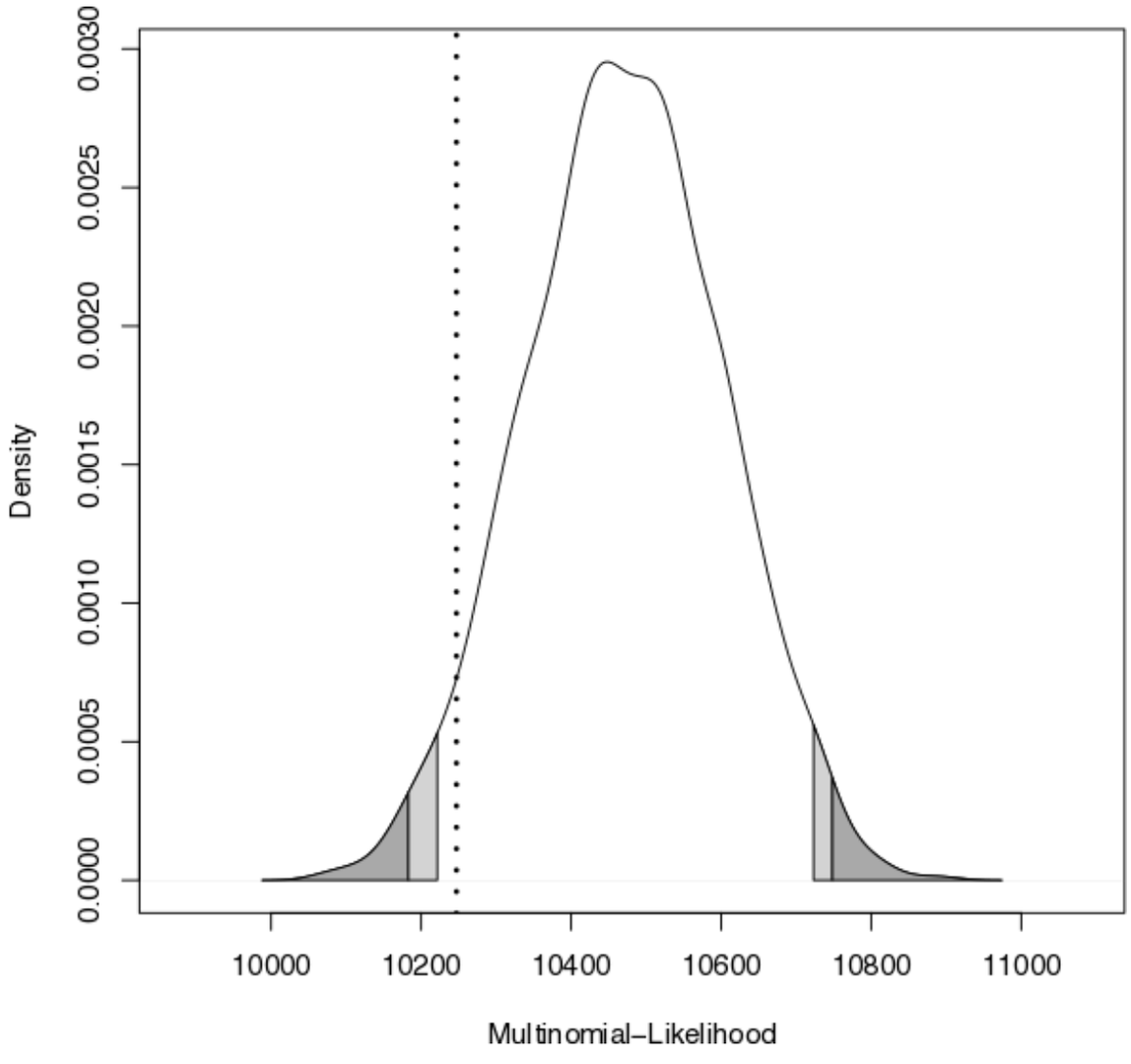
Tests for the rate of evolution along branches.

Multinomial-likelihood:

General test of model fit.

Example Analysis





Computing P-Values

Lower one-tailed p-value:

$$p_L = P(s(d) \leq s(e)) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}_{s(d_i) \leq s(e)}$$

Upper one-tailed p-value:

$$p_U = P(s(d) \geq s(e)) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}_{s(d_i) \geq s(e)}$$

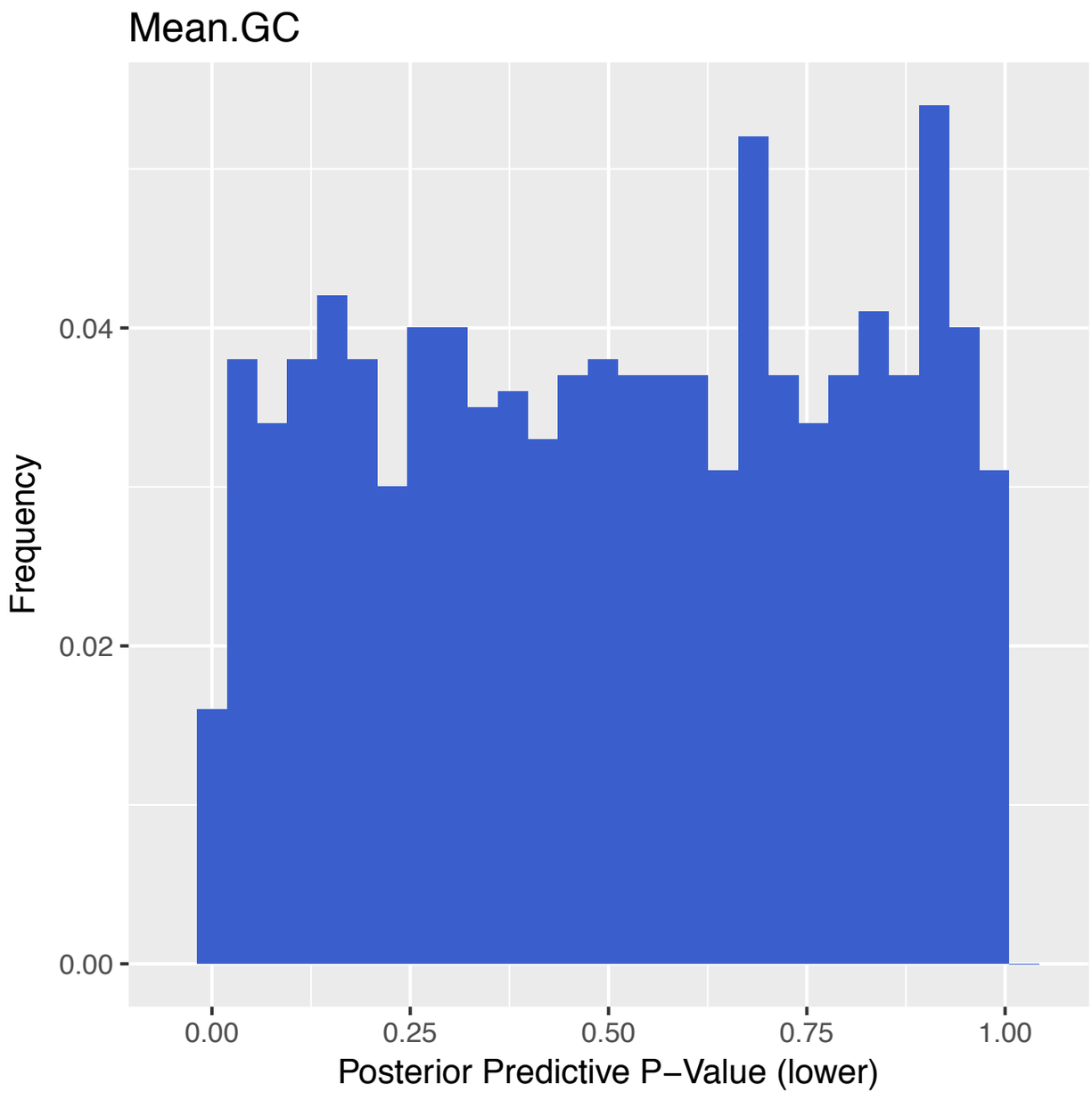
Two-tailed p-value:

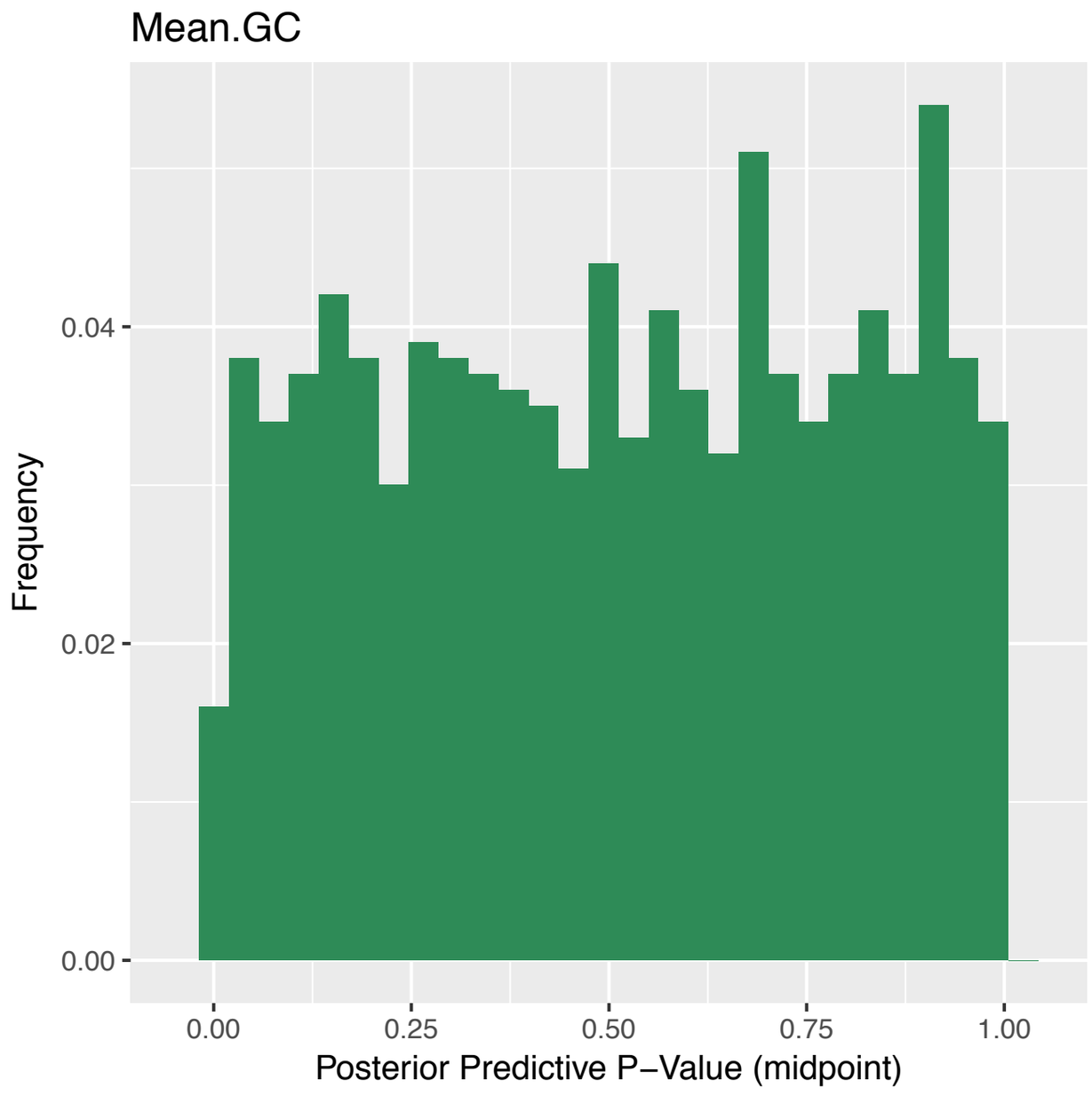
$$p_T = 2 \times \min(p_L, p_U)$$

Midpoint lower one-tailed p-value:

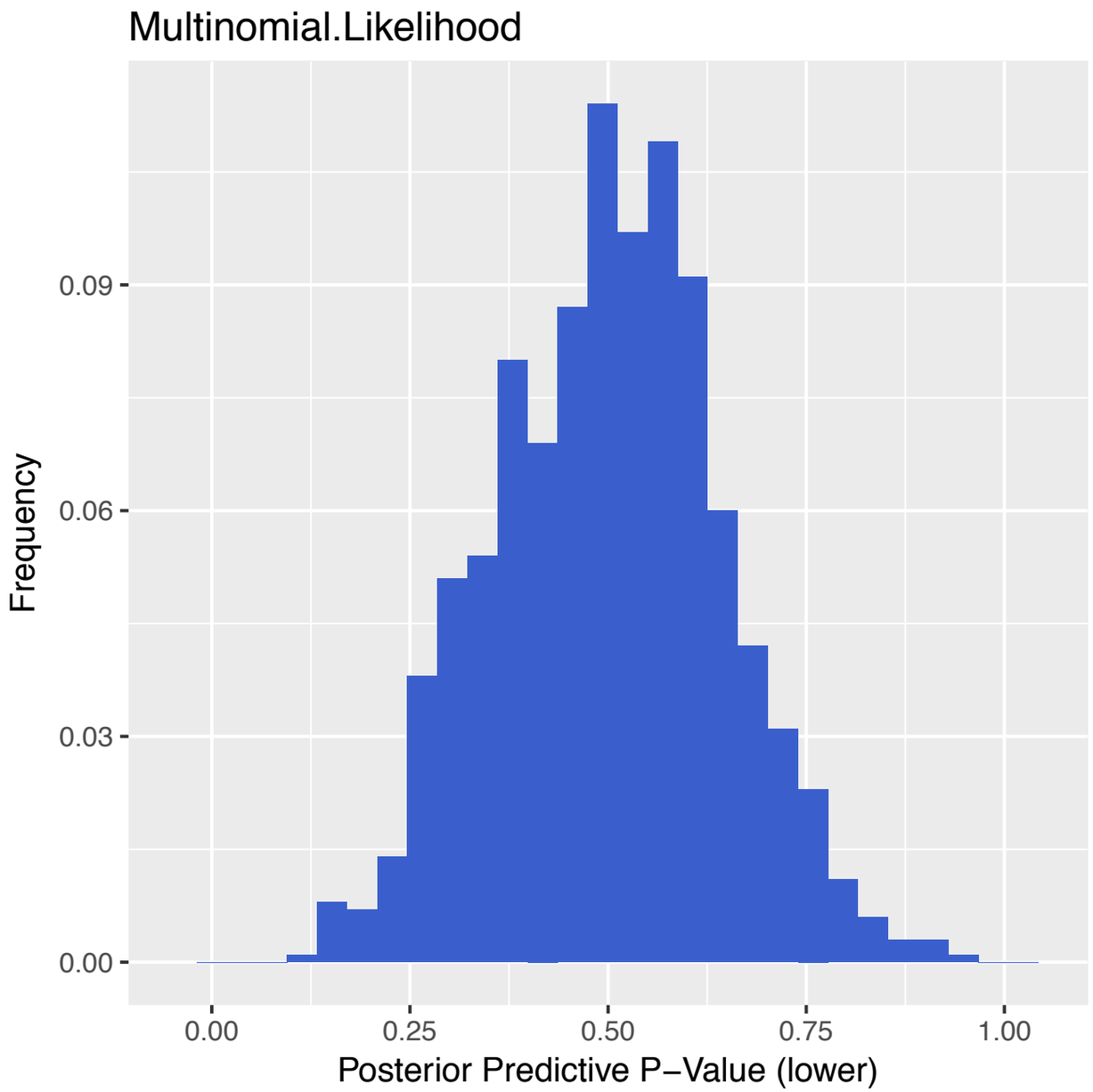
$$p_M = P((d) \lesssim s(e)) = \frac{1}{M} \sum_{i=1}^{M} \mathbf{1}_{s(d_i) < s(e)} + \frac{1}{2M} \sum_{i=1}^{M} \mathbf{1}_{s(d_i) = s(e)}$$

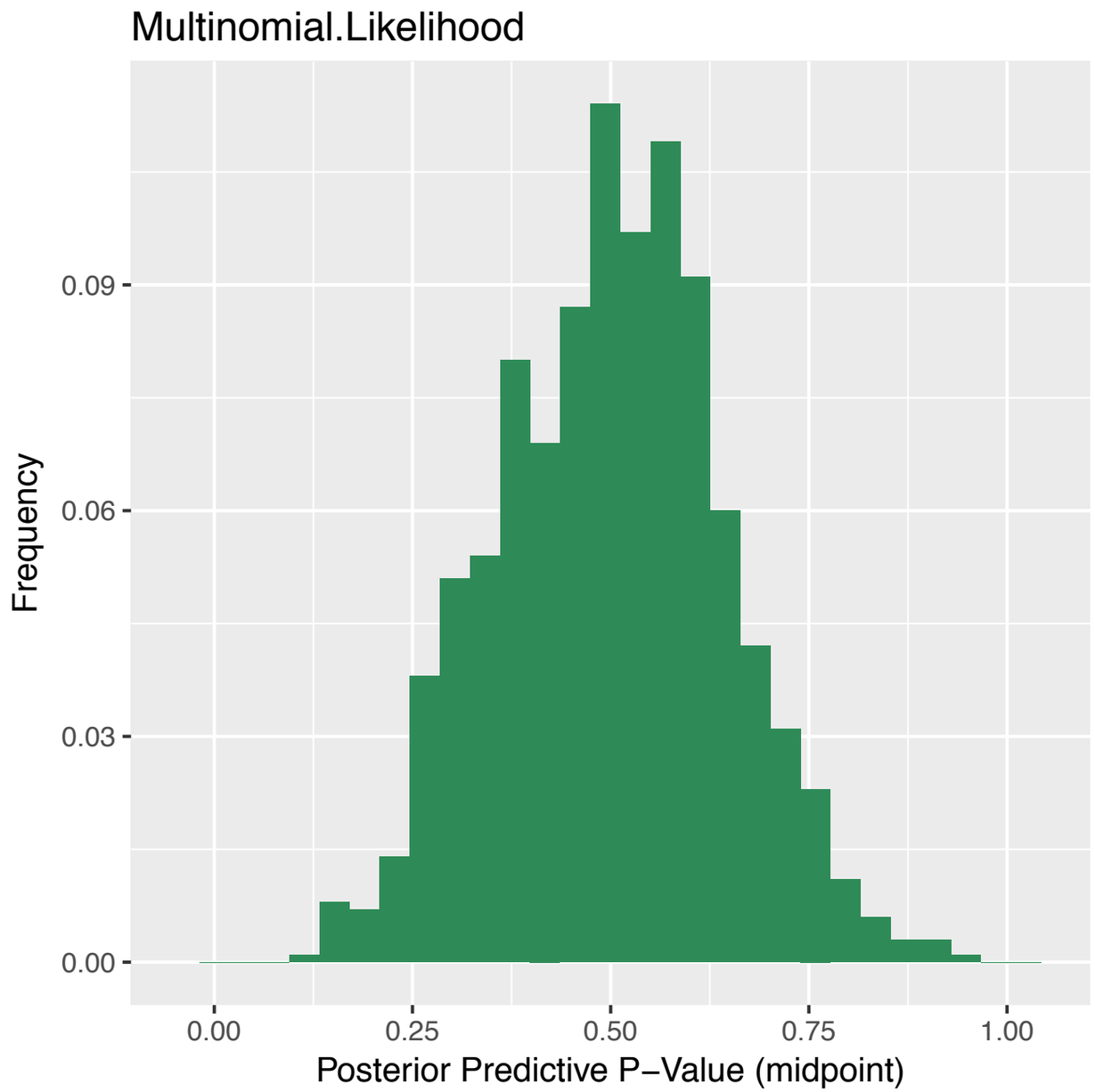
Behavior of p-values under simulation



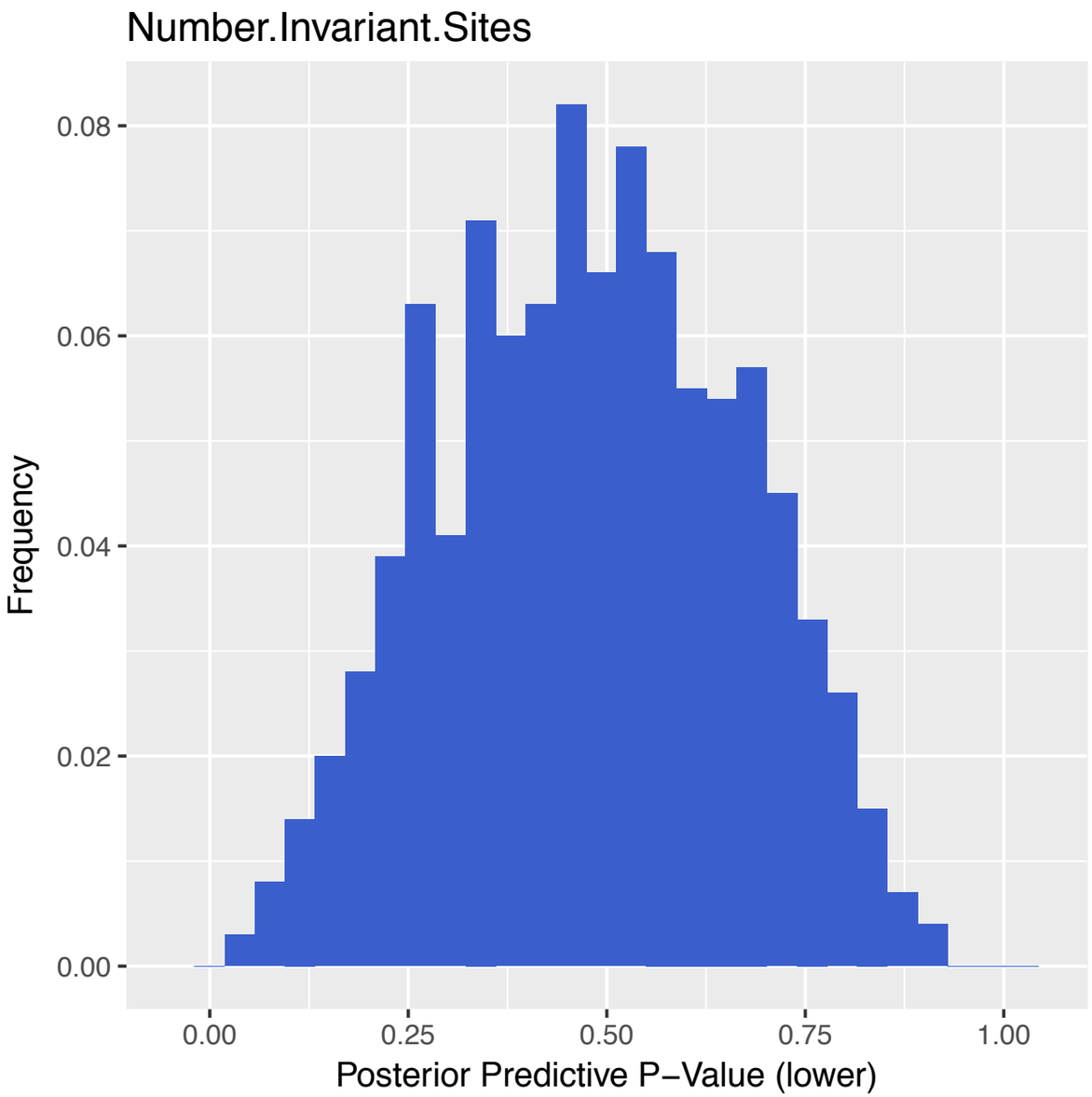


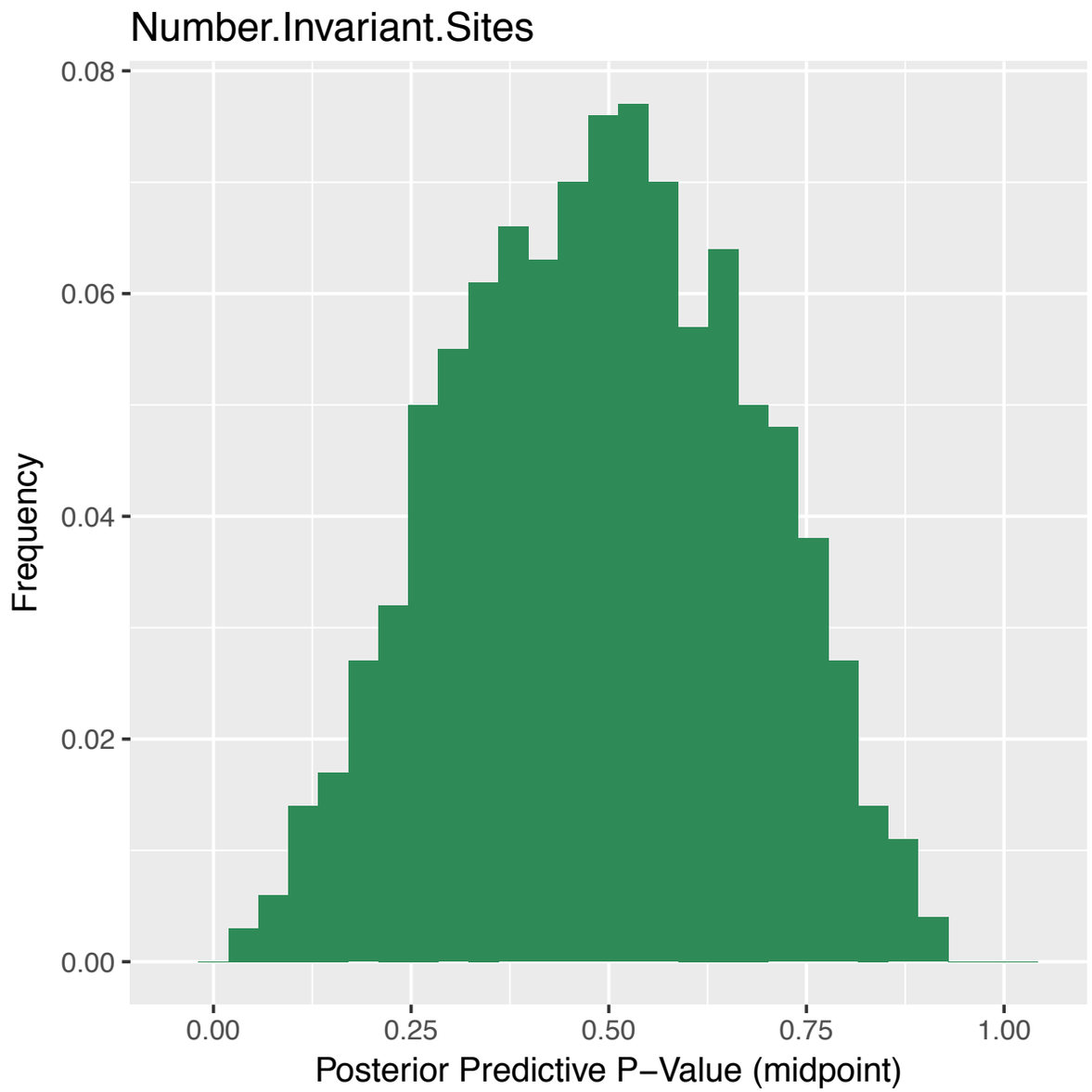
Behavior of p-values under simulation



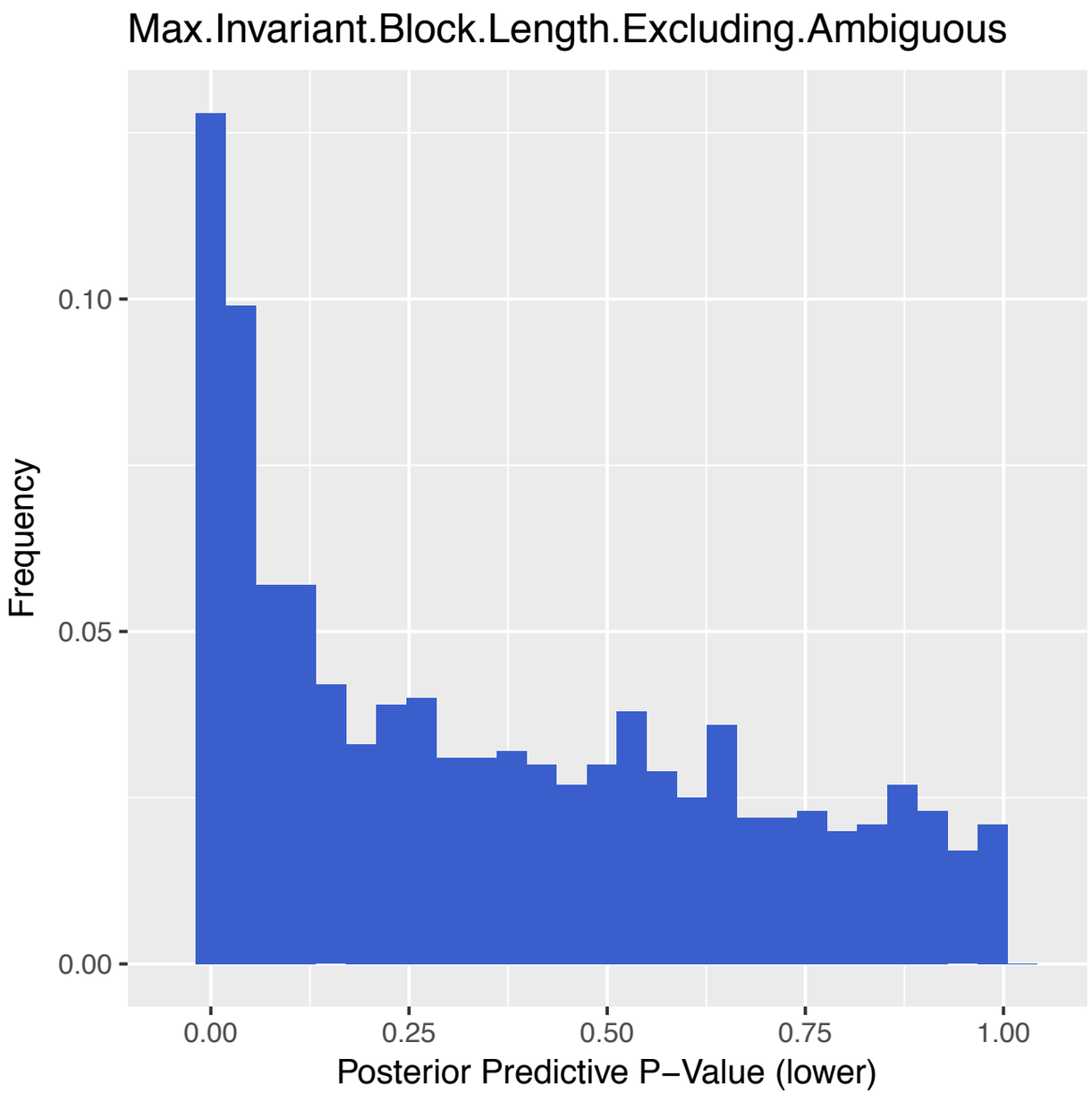


Behavior of p-values under simulation





Behavior of p-values under simulation



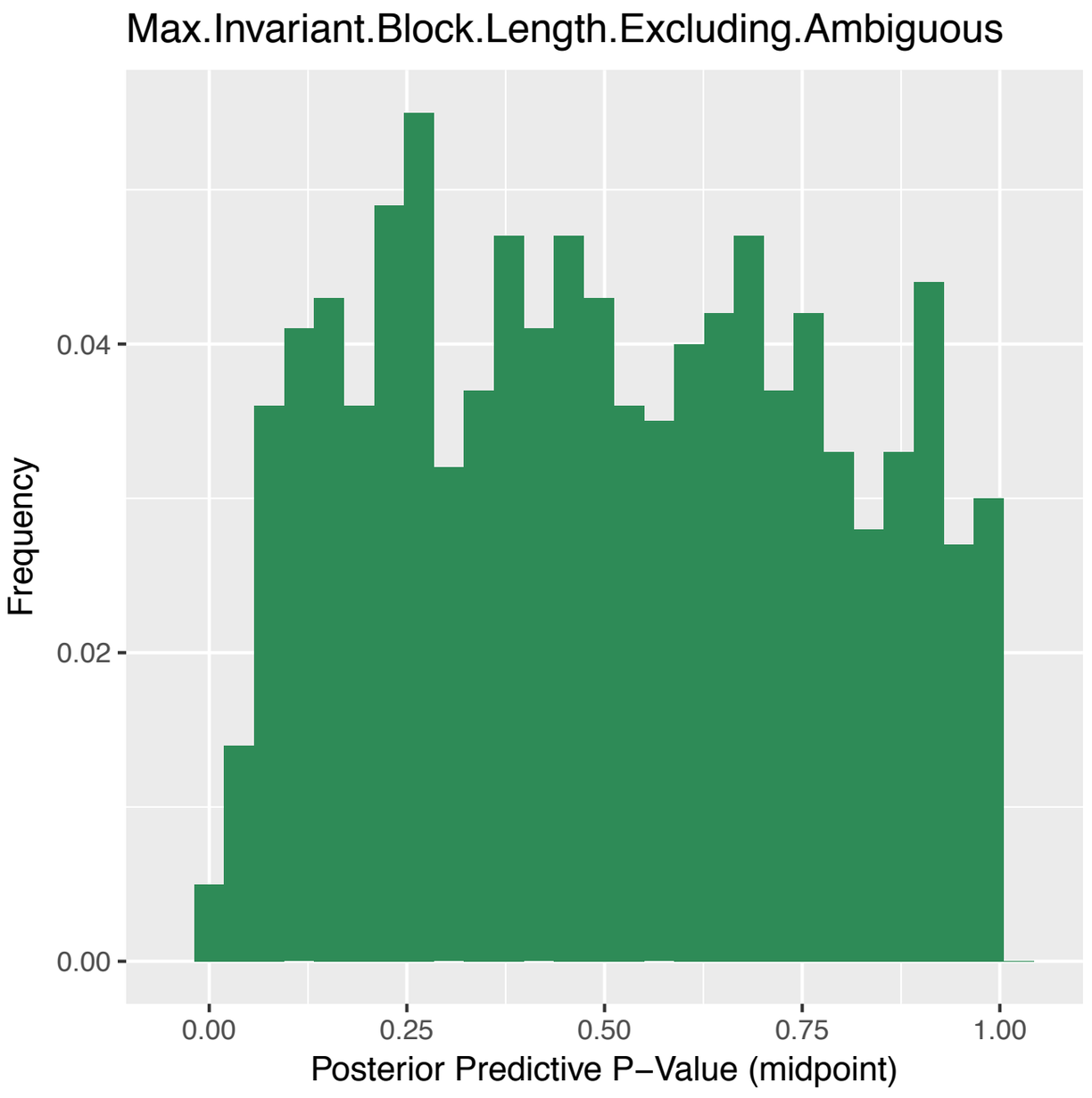
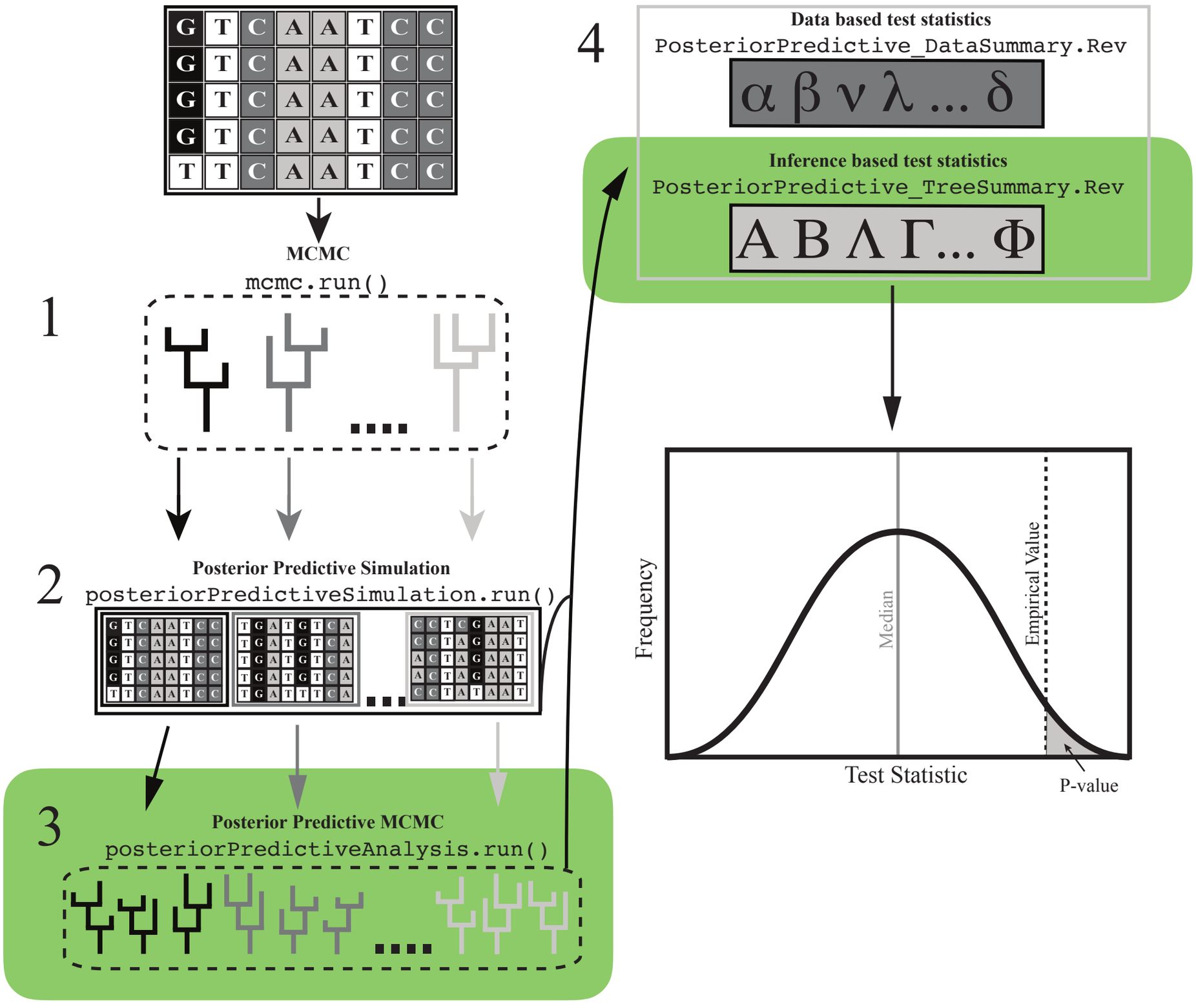
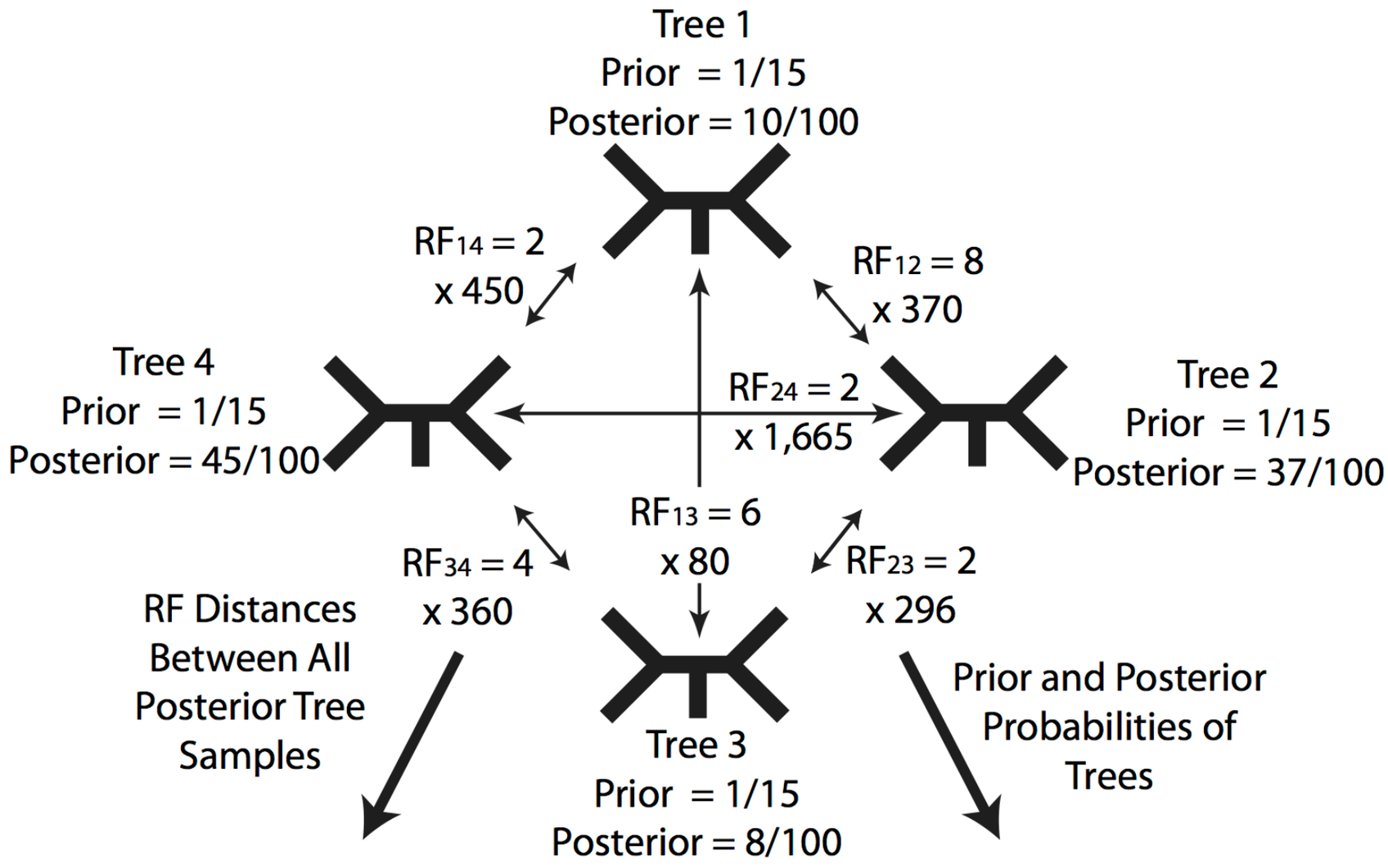


Table 2: Lower 1-tailed midpoint p-values for data based test statistics.

Statistic\Model	JC	GTR	GTR + Inv	GTR + Gamma	GTR + Gamma + Inv
#Invariant Sites	**1**	**1**	0.903	0.8535	0.224
Max(GC)	**0**	**1**	**0.9815**	0.683	0.9355
Max(PD)	**0**	**0**	**0**	0.0885	0.3075
Min(GC)	**0**	**0.984**	0.5455	0.088	0.341
Min(PD)	0.3685	0.8665	0.554	0.7205	0.544
Mean(GC)	**0**	**0.999**	0.7485	0.158	0.5385
Var(GC)	**1**	**1**	**1**	**1**	**1**
Multinomial-Likelihood	**1**	**1**	0.879	0.535	0.043

P^3: Phylogenetic Posterior Prediction





x1,729 x2,411 x360 x80 x370

0 . . . 0 2 . . . 2 4 . . . 4 6 . . . 6 8 . . . 8

Median = 2

1st Quartile = 0

3rd Quartile = 2

99th Percentile = 8

Interquartile Range = 2 - 0 = 2

$$\text{Information Gain} = \ln(N) + \sum_{i=1}^{N} p(\tau_i|\mathbf{X}) * \ln[p(\tau_i|\mathbf{X})]$$

$$= \ln(15) + 0.1\ln(0.1) + 0.37\ln(0.37) + 0.08\ln(0.08) + 0.45\ln(0.45)$$

$$= 1.548$$

Example inference-based summary statistics

Mean and Quantiles Robinson-Foulds-Distance

Captures the spread of the posterior distribution on trees.

Mean and variance tree-length:

Tests for inferred branch lengths and the expected number of substitutions

Entropy:

comparing the prior to the posterior distribution of phylogenetic tree topologies

Table 1: Lower 1-tailed midpoint p-values for inference based test statistics.

Statistic\Model	JC	GTR	GTR + Inv	GTR + Gamma	GTR + Gamma + Inv
mean(RF)	**1**	0.974	**0.981**	0.931	0.772
q(RF,0.25)	0.5	0.499	0.9335	0.755	0.4905
q(RF,0.5)	**0.999**	0.4795	0.9625	0.829	0.5795
q(RF,0.75)	**1**	0.868	**0.9835**	0.8885	0.6685
q(RF,0.99)	**1**	**0.9995**	**0.995**	**0.9925**	0.755
q(RF.0.999)	**1**	1	**0.9985**	0.969	0.823
mean(TL)	0.426	0.3855	0.51	0.945	0.887
var(TL)	0.391	**0.998**	0.911	**1**	**1**
Entropy	**0.999**	**1**	**0.99**	**0.992**	0.972

P^3: Phylogenetic Posterior Prediction