CAGEE: computational analysis of gene expression evolution

Jason Bertram[1,2,*], Ben Fulton[1,3], Jason P. Tourigny[1,4], Yadira Peña-Garcia[1], Leonie C. Moyle[1], and Matthew W. Hahn[1,4,*]

[1]Department of Biology, Indiana University, Bloomington, IN 47405, USA

[2]Department of Mathematics, Western University, London, ON N6A 5B7, Canada

[3]University Information Technology Services, Indiana University, Bloomington, IN 47405, USA

[4]Department of Computer Science, Indiana University, Bloomington, IN 47405, USA

*Corresponding authors: E-mail: jason.bertram@uwo.ca, mwh@indiana.edu

Abstract

Despite the increasing abundance of whole transcriptome data, few methods are available to analyze global gene expression across phylogenies. Here, we present a new software package (CAGEE) for inferring patterns of increases and decreases in gene expression across a phylogenetic tree, as well as the rate at which these changes occur. In contrast to previous methods that treat each gene independently, CAGEE can calculate genome-wide rates of gene expression, along with ancestral states for each gene. The statistical approach developed here makes it possible to infer lineage-specific shifts in rates of evolution across the genome, in addition to possible differences in rates among multiple tissues sampled from the same species. We demonstrate the accuracy and robustness of our method on simulated data, and apply it to a dataset of ovule gene expression collected from multiple self-compatible and self-incompatible species in the genus *Solanum* to test hypotheses about the evolutionary forces acting during mating system shifts. These comparisons allow us to highlight the power of CAGEE, demonstrating its utility for use in any empirical system and for the analysis of most morphological traits. Our software is available at https://github.com/hahnlab/CAGEE/.

Key words: RNA-seq; phylogenetic comparative methods; Brownian motion; Solanum

Introduction

Early studies of gene expression in single genes revealed widespread and frequent changes in the levels, timing, and breadth of expression across species (reviewed in Wray et al. 2003; Fay and Wittkopp 2008; Hill et al. 2021). Such changes in gene expression have been shown to be responsible for many differences between species, and may be a major driver of evolution (King and Wilson 1975). Advances in sequencing technologies (i.e. RNA-seq) have transformed research into gene expression, allowing researchers to cheaply and accurately measure transcript levels for every gene in a genome, in multiple tissues, and across several timepoints or conditions (Wang et al. 2009). There is now a flood of interest in applying RNA-seq to whole clades of organisms in order to identify the genetic changes and evolutionary forces driving species differences (e.g. Brawand et al. 2011; Meisel et al. 2012; Coolon et al. 2014; Harrison et al. 2015; Berthelot et al. 2018; Catalan et al. 2019; Blake et al. 2020; El Taher et al. 2021).

To better understand the importance of changes in gene expression, researchers must be able to characterize the mechanisms and modes by which gene expression evolves. Such work entails understanding the role of natural selection in driving species differences, the stages of development or the tissues that evolve most rapidly, as well as the environments most likely to generate changes in gene expression (Dunn et al. 2013; Hill et al. 2021; Price et al. 2022). Phylogenetic comparative methods enable the rigorous study of traits like gene expression across a species tree (Revell and Harmon 2022). These methods can be used for testing hypotheses about natural selection, the inference of ancestral states (allowing us to polarize the direction of changes), and the estimation of evolutionary rates. Multiple software packages are available that implement a wide variety of comparative methods (e.g. Pennell et al. 2014), including models specifically intended for studying gene expression across a tree (Bedford and Hartl 2009; Rohlfs et al. 2014; Rohlfs and Nielsen 2015; Catalán et al. 2019; Chen et al. 2019; Yang et al. 2019).

However, as far as we are aware, all existing comparative methods for analyzing gene expression implement fundamentally single-gene analyses. Each gene is

65 considered a separate trait, such the evolutionary parameters for each gene are
66 estimated separately. Single-gene analyses can be used to identify tissue-specific or
67 lineage-specific shifts in evolutionary rates, but their power is quite low (Beaulieu et al.
68 2012). As a result, identifying trends in evolution must be carried out *post hoc* by
69 summing the number of genes found to be individually significant (e.g. Harrison et al.
70 2015; El Taher et al. 2021). This approach is less than ideal, especially when carrying
71 out comparisons between branches of different lengths or between tissues with different
72 average expression levels (both of which can result in differential statistical power).

73 Therefore, to better characterize the forces affecting gene expression evolution,
74 we must be able to model effects shared along a lineage, experienced by many genes
75 in the same tissue, or experienced by all genes found in the same environment. In this
76 article, we present a genome-scale platform for the analysis of gene expression data
77 that allows for such shared factors. Our software, CAGEE (Computational Analysis of
78 Gene Expression Evolution), provides a robust set of methods for analyzing expression
79 data across a species tree. CAGEE estimates ancestral states and rates, with rates
80 shared by all or subsets of genes (single-gene analyses can also be carried out). We
81 show that lineage-specific and tissue-specific (or condition-specific) rates can be
82 accurately inferred, and provide principled statistical approaches for model selection.
83 Our current implementation uses a bounded Brownian motion model and assumes
84 expression data are accurate, but the architecture and codebase will easily allow for
85 future extensions that relax these and other assumptions.

86

87 **New Approaches**

88 We model gene expression evolution as a bounded Brownian motion (BBM)
89 process on a known species tree (cf. Boucher and Démery 2016). Our model has a
90 single bound: trait values must be greater than or equal to zero; there is no upper bound
91 (Figure 1). Previous researchers have often modeled gene expression using an
92 Ornstein-Uhlenbeck (OU) process (e.g. Bedford and Hartl 2009; Rohlfs et al. 2014;
93 Rohlfs and Nielsen 2015; Chen et al. 2019), a model that includes a force constraining
94 traits about the mean. However, to our knowledge, the OU model has only been

compared against an unbounded Brownian motion model (i.e. one that allows negative expression values), making fair comparisons difficult. In addition, OU models may be frequently and incorrectly favored over simpler models due to several biases (e.g. measurement error), especially when the number of tips in a tree is small (Pennell et al. 2015; Silvestro et al. 2015; Boucher and Démery 2016; Cooper et al. 2016; Catalán et al. 2019). Therefore, the initial version of our software models gene expression with the BBM process, which naturally bounds possible values without invoking an additional constraining force.

Let $E_{ij} \geq 0$ be the expression level of gene i in species j. We assume that log-transformed expression $X_{ij} = \ln(E_{ij} + e_{\min})$ evolves as a Brownian motion process with variance σ^2 per unit time, where $e_{\min}$ is a small offset (constant across genes and species) that prevents X_{ij} from taking infinite values if measured values of E_{ij} are zero. We log-transform before assuming Brownian motion because we expect the variance in the evolutionary process to scale with expression level. Assuming that E_{ij} is itself Brownian would unrealistically assume that the rate of evolution is constant across expression levels, even though expression levels vary by many orders of magnitude. We impose a reflecting lower boundary at $x_{\min} = \ln(e_{\min})$, meaning that the Brownian walk immediately bounces back if it reaches $x_{\min}$. Expression can therefore effectively never reach zero, our theoretical lower bound (Figure 1).

The second major feature of our model (as implemented in CAGEE) is that many genes can share the evolutionary rate parameter, σ^2. This rate may be shared among genes expressed in the same tissue or sample, among genes located on the same chromosome, or among genes evolving along the same lineage of the phylogenetic tree. The simplest model allows σ^2 to be shared among all genes, providing an average rate of evolution across the genome and over time; this average may include genes that vary in their individual rates of evolution. We explain this model briefly here, with more detail provided in the Materials and Methods.

CAGEE infers the most likely value(s) of σ^2 consistent with an ultrametric tree, T, and a set $E_{\{ij\}}$ of measured expression values at the tips of the tree; i.e. it maximizes

124 the likelihood $L(\sigma^2|E_{\{ij\}},T)$. Each gene is assumed to evolve independently, and so the
125 likelihood for each gene $L_i(\sigma^2|E_{i\{j\}},T)$ is computed independently. The overall likelihood
126 is obtained as the product $L\left(\sigma^2\middle|E_{\{ij\}},T\right) = \Pi_i L_i(\sigma^2|E_{i\{j\}},T)$ across genes. The likelihood
127 for each gene $L_i(\sigma^2|E_{i\{j\}},T)$ is computed using the pruning algorithm (Felsenstein
128 1973). The key ingredient needed to apply the pruning algorithm is the transition
129 probability density $p\left(x_t\middle|x_{t_0}\right) = \Pr\left[X(t) = x_t|X(t_0) = x_{t_0}\right]$ for log-expression at time t
130 conditional on having log-expression x_{t_0} at time t_0 along a lineage. CAGEE computes
131 the transition density by solving the standard Brownian diffusion equation with reflecting
132 boundary conditions (Materials and Methods). The transition density is used to
133 propagate expression probabilities along the tree: if the probability density of log-
134 expression at time t_0 is $f(x_{t_0})$, then the probability density at time t on the same lineage
135 is $f(x_t) = \int p\left(x_t\middle|x_{t_0}\right)f\left(x_{t_0}\right)dx_{t_0}$. At each tip the probability density $f(x_{t_0})$ is a delta
136 function centered at the corresponding measured value of X_{ij}.

137 Starting with the known tip distributions, the pruning algorithm propagates back
138 to the tips' parent nodes. The distribution at the parent node is then the product of the
139 two backward-propagated child node distributions. Proceeding iteratively across the
140 tree, we ultimately obtain the gene-specific probability density for expression value at
141 the root $f_i(x_R)$. Viewed as a likelihood for σ^2, $f_i(x_R)$ is the gene-specific likelihood
142 conditional on the unknown ancestral root value; i.e. $f_i(x_R) = L_i(\sigma^2|E_{i\{j\}},T,x_R)$.
143 Therefore, we integrate over all possible x_R to obtain,

144 $$L_i\left(\sigma^2\middle|E_{i\{j\}},T\right) = \int L_i\left(\sigma^2\middle|E_{i\{j\}},T,x_R\right)\rho(x_R)dx_R, \quad (1)$$

145 where $\rho(x_R)$ is the prior distribution for the root value of a randomly selected gene.

146 The default prior $\rho(x_R)$ is assumed to be a gamma distribution with $k = 0.375$ and $\theta =$
147 1600, though this distribution can also be set by the user in CAGEE. This choice is
148 based on estimated expression distributions across genes in individual species, which
149 we take as our baseline for the ancestral distribution. CAGEE uses the Nelder-Mead
150 simplex method to find the optimal value(s) of σ^2.

151

Results

Using CAGEE

The required inputs for CAGEE are a Newick-formatted, rooted, ultrametric tree (with branch lengths) and a tab-delimited data file containing the expression levels of all species or taxa being studied. The data file can consist of data on one gene/transcript or thousands of different genes. The first line of the data file should contain the species' names (matching those used in the Newick tree). In addition, headers for gene names, gene descriptions, and sample IDs (see next section for an explanation of "samples" in CAGEE) can be used. Subsequent lines each correspond to a single gene and contain expression levels for each species. Missing data can be denoted using multiple characters (-/?/N). Examples of Newick trees and corresponding data files can be found in the online user manual (https://github.com/hahnlab/CAGEE/docs/manual/cagee_manual.md).

We expect that CAGEE will most often be used to calculate the following outputs: one or more σ^2 values, ancestral states at each internal node (including 95% credible intervals around these states), and the final likelihood associated with a model. However, users do not have to search for σ^2: if a value for this parameter is specified, then the output of CAGEE will just be the ancestral states and a likelihood. In addition to the raw outputs provided in multiple formats (both tab-delimited files and NEXUS-formatted files), CAGEE computes basic statistics about changes in expression levels by comparing values at parent and child nodes. Summaries of these inferred changes for every gene and for every branch of the tree are output, so that the evolutionary history of gene expression changes in every gene are accessible to users. To avoid over-interpretation of small changes in inferred expression levels—especially when there is uncertainty in ancestral states—CAGEE will also compare the credible intervals at parent and child nodes to note if a change is "credible" (i.e. the intervals do not overlap). Credible intervals are calculated by summing the probabilities across possible ancestral states at each node, so that 95% of the probability density is included. Credible changes on each branch are annotated as such in the output.

181 We most often expect that an ultrametric species tree will be used as the input
182 topology, but this is not required by CAGEE. If users wish to specify a gene tree, or
183 some other bifurcating tree, as input, those can be used in CAGEE as well. However,
184 the major advantage of CAGEE—incorporating information from multiple genes to
185 accurately estimate genome-wide rates—will rapidly diminish for trees that represent
186 the history of only a minority of the genome. Trees that include duplication events
187 should provide suitable estimates for any genes that follow this topology, but CAGEE
188 does not have a way to further combine disparate gene trees.

189 There are multiple options available for running CAGEE. Users who can take
190 advantage of multiple threads can specify the number to use on the command line.
191 Complex models can also take a long time to converge; by default, CAGEE runs a
192 maximum of 300 iterations of the Nelder-Mead search, but users can increase this
193 number in subsequent runs if the likelihood is still improving when the limit is hit. As
194 mentioned above, the default prior distribution for the root state is a gamma distribution
195 with $k = 0.375$ and $\theta = 1600$. This distribution can also be specified by the user if
196 desired. Information on how to run more complex evolutionary models, beyond a single
197 σ^2, is given in the next section.

198 *Estimating evolutionary rates in CAGEE*

199 We tested CAGEE's ability to accurately estimate σ^2 by varying this rate
200 parameter and the number of genes used for inference, as well as the amount of
201 missing data in each dataset. We simulated different single values of σ^2 across a tree
202 with constant branch lengths (Supplementary Figure 1) using the simulation tool
203 available within CAGEE. (Note that the total amount of evolution in a tree is determined
204 by the product $\sigma^2 \cdot t$, such that changes in branch lengths will have an effect
205 commensurate with changes in σ^2.) Figure 2 shows the average error associated with
206 estimates of different σ^2 values and using different numbers of genes within each
207 dataset. As can be seen, the error across all parameter values and dataset sizes is
208 quite small (generally less than 2.5%), and is less variable for larger dataset sizes.
209 Fortunately, we expect that most empirical datasets will contain closer to 10,000 genes

210 than 1,000 genes. The results in Figure 2 are for an ancestral state vector of length
211 *N*=200 (the default setting in CAGEE; Materials and Methods); we also estimated σ^2
212 when allowing the ancestral state vector to have length *N*=500 (Supplementary Figure
213 2A). There appears to be minimal gain from increasing the resolution in this vector,
214 though the computational time is greatly increased (similar to results in Boucher and
215 Démery 2016). We evaluated the accuracy of CAGEE when different amounts of data
216 were randomly missing: from 0% to 75% for a dataset of 1,000 genes. As shown in
217 Supplementary Figure 2B, CAGEE has high accuracy even when large amounts of data
218 are missing (at random) from a dataset.

219 One major advantage of using CAGEE is that it combines information from
220 multiple genes to infer a rate of evolution: this is why it can return estimates with high
221 accuracy even when a large fraction of the data are missing. To further demonstrate this
222 advantage, we simulated evolution in 1,000 genes using the same parameter value
223 (σ^2=1) and then estimated σ^2 for each of the 1,000 genes individually. Supplementary
224 Figure 2C shows that these individual estimates of σ^2 are quite error-prone: although
225 the mean of all genes is close to the true value, individual estimates can be 7-8X higher
226 or lower and there is a large amount of variance. Although we have not shown it here,
227 we do expect that the accuracy of σ^2 will be greater for trees with larger numbers of
228 tips, even for estimates derived from single genes (cf. O'Meara et al. 2006). On the
229 other hand, CAGEE is combining information from multiple genes to infer an *average*
230 rate of evolution, even when the underlying rate may be quite variable. To explore any
231 effect of underlying rate variation, we carried out further simulations that combined three
232 simulations of 1,000 genes each with σ^2 equal to 0.5, 3, and 9, respectively (we
233 repeated these simulations 10 times). When analyzed as single datasets with 3,000
234 genes total, the average σ^2 inferred was 3.76, approximately 9% lower than the
235 arithmetic mean rate (Supplementary Figure 2D). It is well-known that single-rate
236 phylogenetic likelihood models tend to underestimate rates of evolution when there is
237 underlying variation (Golding 1983; Gillespie 1986; Yang 1996; Mendes et al. 2020),
238 and we see this effect here. Fortunately, the bias is small, and can be corrected in the
239 future by including gamma-distributed rate variation into CAGEE. Overall, inferences of

240 σ^2 should be quite accurate when a single rate parameter is shared across the tree and
241 across all genes and lineages.

242 Variation in the rate of expression can currently be accommodated by CAGEE in
243 a number of ways, using multi-rate σ^2 models. One type of model allows users to
244 specify that their data come from different "samples": these samples can represent
245 tissues, conditions, timepoints, and even subsets of the genome (e.g. the X
246 chromosome, or a specific functional class of genes). In the input data file, the
247 "SAMPLETYPE" column is used to indicate which sample each gene is a member of; a
248 separate σ^2 value will be calculated for each sample or set of samples (these values
249 are assumed to be shared among all lineages in the tree). Specifying more than one
250 sample means that an individual gene or transcript name can be used more than once
251 (i.e. once for each sample), but there is no requirement that genes are measured in
252 each sample. For instance, assigning all autosomal genes to sample 1 and all X-linked
253 genes to sample 2 would not permit for any overlap in gene assignment, but is perfectly
254 allowable in CAGEE.

255 Each additional sample requires another σ^2 parameter to be estimated, and often
256 researchers would like to know if fitting this extra parameter is justified by the data.
257 Under standard information-theoretic criteria (Burnham and Anderson 2002), twice the
258 difference in log-likelihoods between nested models should be χ^2-distributed with
259 degrees of freedom equal to the difference in the number of parameters between
260 models. To test this expectation, we simulated 1000 datasets with a single σ^2 value, but
261 fit models with two σ^2 values (assigning 1000 genes to two equal-sized samples at
262 random; the relative size of the samples should not affect the false positive rate). As
263 anticipated, the results fit a χ^2 distribution with one degree of freedom, with 4.4% of
264 datasets having a difference in 2*log-likelihood greater than 3.84 (5% are expected by
265 chance). This indicates that standard statistical procedures should adequately control
266 the false positive rate when fitting multi-sample σ^2 models.

267 CAGEE also allows models in which σ^2 varies across branches of the species
268 tree. It does so by fitting separate σ^2 parameters for different parts of the tree. On the

269 command line, CAGEE enables users to specify how multiple σ^2 parameters should be
270 assigned to branches. For *n* taxa, from 1 to 2*n*-2 parameters can be specified, and
271 branches can be grouped together in any way. For instance, a two-parameter model
272 can have all branches that share a rate adjacent to one another in the tree
273 (Supplementary Figure 3A) or spread out across the tree (Supplementary Figure 3B).
274 Similar to the analyses carried out above for the false positive rate associated with
275 multiple samples, we simulated data with a single σ^2 value and then fit models with
276 multiple σ^2 parameters. Regardless of how we distributed the two rate classes across
277 the tree we observed good control of the false positive rate: 4.5% and 5.4% of 1000
278 simulated datasets were significant at the *P*=0.05 level (for the trees shown in
279 Supplementary Figures 3A and 3B, respectively). More limited simulations also showed
280 that we could accurately estimate multiple σ^2 parameters when the data were simulated
281 with multiple rates (Supplementary Table 1). Together, our results suggest that we can
282 estimate multiple types of multi-rate models, and can accurately control the false
283 positive rate when doing so.

284 *Analysis of wild tomato transcriptome data*

285 To demonstrate the utility of CAGEE in an empirical system, we analyzed data
286 from a clade that includes domesticated tomato, *Solanum lycopersicum*. This dataset
287 contains gene expression levels in unfertilized ovules from the flowers of six species,
288 one of which (*S. pennellii*) has two different populations represented (Figure 3). There
289 are 14,556 genes with expression levels measured in all seven accessions. RNA-seq
290 data for five of the seven accessions have been published previously (Moyle et al. 2021;
291 Hibbins and Hahn 2021), while two others are presented here for the first time
292 (Materials and Methods). Note, however, that all data were collected from all samples at
293 the same time (Materials and Methods).

294 Most species within the tomato clade are self-incompatible (SI), the ancestral
295 state in the family Solanaceae (Igić et al. 2006). Self-incompatibility means that plants
296 must outcross in order to successfully fertilize ovules. However, self-compatibility (SC)
297 has evolved multiple times both within the Solanaceae and within the genus *Solanum*

298 (Goldberg et al. 2010; Bedinger et al. 2011). Self-compatible individuals are able to
299 successfully fertilize ovules using their own pollen, though many also still outcross
300 (Whitehead et al 2018; including in *Solanum*: Vosters et al. 2014 and references
301 therein). Importantly, we have *a priori* expectations about the rate at which reproductive
302 traits—including ovule gene expression—might evolve between groups with different
303 mating systems. Due to conflict within and between the sexes, it is generally expected
304 that reproductive traits in species that outcross more (i.e. SI taxa) should evolve more
305 rapidly than in species that inbreed more (i.e. SC taxa; Clark et al. 2006). Such patterns
306 are found in some analyses of the rate of protein evolution (e.g. Gossmann et al. 2016;
307 Harrison et al. 2019), but are equivocal in other comparisons (e.g. Gossmann et al.
308 2014, Moyle et al. 2021). These complex patterns might reflect additional effects that
309 also accompany mating system shifts; for instance, such shifts often lead to reductions
310 in effective population size in more selfing lineages (Charlesworth and Wright 2001).
311 Mating system shifts could also alter global patterns of molecular evolution (including
312 gene expression) by changing the strength and pattern of purifying selection, as
313 morphological changes often accompany mating system changes. The exact effect of
314 shifts in mating system on molecular evolution remains an open question.

315 The *Solanum* species sampled here represent two independent transitions from
316 SI to SC, with one of the transitions (in accession *S. pennellii* LA0716) occurring
317 recently enough that different populations within this species have different
318 incompatibility systems (Figure 3). We therefore fit a series of nested models within
319 CAGEE to test two related hypotheses about ovule gene expression evolution. First, we
320 would like to know whether the rate of evolution of ovule gene expression is different in
321 SI species than in SC species. Second, given the recent transition to SC within
322 accession *S. pennellii* LA0716, we wanted to know if it shows a pattern of evolution
323 more similar to SI or to SC species. In total, we fit four separate evolutionary models
324 (Table 1; Figure 3). Model A has a single rate parameter for the entire tree. Model B has
325 two rate parameters, one for SI species and one for SC species. This model assigns the
326 branch leading to *S. pennellii* LA0716 as SC. Model C also has two rate parameters,
327 one for SI and one for SC, but assigns *S. pennellii* LA0716 as SI. Model D has three

328 rate parameters: one for SI species, one for longer-term SC species, and one for *S.*
329 *pennellii* LA0716.

330 Estimated results from the different models are shown in Table 1. Model A has a
331 worse fit than any other model, with a single σ^2 value of 0.102. For context, this value
332 means that the bounded Brownian motion process the data are fit to has a variance of
333 0.102 per million years (of log-transformed expression values). This is the average rate
334 across all 14,556 genes and across all branches of the tree. In contrast to a single-rate
335 model, both models B and C are significantly better fits to the data. Contrary to some
336 hypotheses, both models find that SI lineages (σ_1^2) have a lower rate of evolution than
337 SC lineages (σ_2^2; Table 1). There is also a difference between the models, with model C
338 (the one in which *S. pennellii* LA0716 shares a rate with SI species) fitting significantly
339 better. To further examine the evolution of *S. pennellii* LA0716, model D fits a three-
340 parameter model, with this lineage assigned its own rate of evolution. This model is a
341 significantly better fit than model C (P<0.00001; χ^2 test with 1 degree of freedom), and
342 demonstrates that *S. pennellii* LA0716 has a rate of evolution (σ_3^2 in Table 1) that is
343 slightly *lower* than SI species. This highly similar rate to SI species implies that it has
344 only recently transitioned to self-compatibility, which is consistent with previous
345 inferences about the timing of transition to SC in this particular accession (e.g. Rick and
346 Tanksley 1981).

347 CAGEE also allows users to infer the number and direction of changes in gene
348 expression levels along each branch of the tree. Figure 3 reports the number of genes
349 that had "credible" increases and decreases in expression level under model D.
350 Credible changes require that the credible intervals around states at parent and
351 daughter nodes do not overlap, in order to account for uncertainty in our inferences.
352 However, because of this, fewer credible changes will be inferred deeper in the tree,
353 where credible intervals get wider. Therefore, while inferences about the identity of the
354 genes changing along each branch is greatly strengthened by using credible changes
355 (these genes are noted in the raw output from CAGEE), the absolute numbers of
356 credible changes cannot be compared across branches, except for sister branches of

357 equal length. For completeness, we show the total numbers of increases and decreases
358 of gene expression in Supplementary Figure 4; as expected, these total numbers are
359 more uniformly distributed across older and younger branches.

360 We assessed whether the genes identified as having credible increases or
361 decreases in expression specifically on any SC branch (solid red branches in Figure 3)
362 were significantly enriched for any biological process or molecular function gene
363 ontology (GO) categories compared to genes with credible changes on any SI branch
364 (black branches in Figure 3). This comparison specifically assesses gene expression
365 evolution associated with a transition to SC, over and above "background" rates of
366 expression evolution across the rest of the clade. Although fold enrichment was modest
367 1.20-1.36X; Supplementary Table 2), there were 11 terms significantly enriched
368 (FDR<0.05) specifically on SC branches; these terms primarily focused on regulation of
369 transcription, metabolic processes, and biosynthesis (Supplementary Table 2). Among
370 the genes in these over-represented categories, a large fraction are transcription factors
371 associated with development (e.g. WRKY and MADS Box), hormonal responses
372 (including ethylene- and auxin-responsive transcription factors), and regulation of cell
373 cycle (e.g. cyclins), in addition to protein kinases (Supplementary Table 2). This
374 enrichment is consistent with increased expression changes in genes involved in cell
375 division, differentiation, and development, that could follow transitions to SC.

376

377

Discussion

Here, we have developed a new software package that enables the estimation of rates of gene expression evolution across a tree, CAGEE. Gene expression levels are much like many other continuous traits, and multiple papers have introduced phylogenetic comparative methods for studying gene expression (Bedford and Hartl 2009; Rohlfs et al. 2014; Rohlfs and Nielsen 2015; Catalán et al. 2019; Chen et al. 2019). However, as far as we are aware none of these methods allows genes to share evolutionary parameters, which precludes the analysis of genome-wide trends, either along the branches of a tree or between tissues/samples/conditions. To overcome this limitation, CAGEE calculates the likelihood of the data using the pruning algorithm (Felsenstein 1973) to facilitate the sharing of evolutionary parameters along branches of the species tree, providing more statistical power to test evolutionary hypotheses. Fortunately, we were able to take advantage of much of the codebase of our existing software, CAFE (Hahn et al. 2005; De Bie et al. 2006; Hahn et al. 2007; Han et al. 2013; Mendes et al. 2020), which implements the pruning algorithm for the analysis of gene family sizes across a tree. While gene expression levels and gene family sizes differ in the type of data they represent (continuous vs. discrete) and their underlying evolutionary models (bounded Brownian motion vs. birth-death), many of the required likelihood calculations and software components are the same.

An important thing to consider for the input to CAGEE is the normalization used to make gene expression levels comparable across species. The data from wild tomatoes used here was normalized using TPM (transcripts per million; Wagner et al. 2012); other published datasets also use this normalization (Berthelot et al. 2018; Chen et al. 2019; El Taher et al. 2021). However, multiple other normalizations have also been used in comparative analyses, including RPKM (Brawand et al. 2011), FPKM (Catalán et al. 2019), and both TMM and CPM (Blake et al. 2020). Each normalization approach has its advantages and disadvantages, and we cannot yet strongly recommend one specific approach as input to CAGEE. The normalization method used will likely depend on the conditions under which samples are collected: if all species can be raised simultaneously in a greenhouse, vivarium, or growth chamber, we expect

408 many fewer batch effects than in samples collected from the field, which will therefore
409 necessitate different normalizations. However, even animals raised in a common
410 environment—but fed different diets—can show many differences in gene expression
411 not due to heritable change (e.g. Somel et al. 2008). Conversely, many between-sample
412 normalization approaches (e.g. TMM, trimmed mean of M values; Robinson and
413 Oshlack 2010) make the assumption that differences in gene expression between
414 samples are rare. While such normalization is sensible in the context of testing for
415 differential expression between samples from the same species, for a set of species
416 that have been evolving independently for millions of years this is likely not an
417 appropriate assumption.

418 CAGEE currently has multiple limitations, both in the available models that can
419 be applied and in the types of data that can be analyzed. As mentioned earlier, many
420 researchers have modeled gene expression using an OU process (Bedford and Hartl
421 2009; Rohlfs et al. 2014; Chen et al. 2019; Yang et al. 2019). Although OU models may
422 be artifactually preferred over unbounded Brownian motion models due to a number of
423 non-biological factors (see discussion in "New Approaches" above), it would still be
424 helpful to be able to compare such a model to the bounded Brownian motion model
425 used here. However, fitting such a model to genome-wide data is non-trivial: each gene
426 must have its own mean expression value (μ), but possibly shared constraint
427 parameters (α) across genes. We have the goal of implementing such a model in the
428 near future, as well as other models commonly used in comparative methods research
429 (e.g. Landis and Schraiber 2017; Boucher et al. 2018).

430 Beyond the evolutionary model applied to any dataset, there are multiple
431 additional sources of variation that could be modeled. For instance, we have previously
432 accounted for measurement error in a likelihood framework, using an empirically
433 parameterized error model (Han et al. 2013). We can imagine both applying a similar
434 model here to RNA-seq data, as well as extending CAGEE to more error-prone data
435 such as single-cell sequencing. Such an extension would treat the level of expression in
436 each cell within a cell type as an error-prone draw from an underlying distribution; one
437 would then be able to infer the rate of evolution within and across cell-types across

438 multiple species. The biggest obstacle to this approach may be in identifying
439 homologous cell types across species (e.g. Tarashansky et al. 2021). In addition, not all
440 genes necessarily share the same average rate of evolution; gamma-distributed rate
441 categories can be used to model this variation among genes (cf. Ames et al. 2012;
442 Mendes et al. 2020). As shown above, not accounting for this rate variation leads to a
443 slight underestimate of σ^2, but also obscures interesting patterns of evolution among
444 genes. Finally, the gene tree discordance found in many phylogenomic datasets implies
445 that complex traits (such as expression levels) will also be controlled by discordant gene
446 trees (Hahn and Nakhleh 2016; Hibbins and Hahn 2021). This underlying discordance
447 can cause evolutionary rates to be overestimated (Mendes et al. 2018), and should be
448 taken into account when seeking accurate parameter estimates (see discussion of wild
449 tomato data below). Our goal is to include methods for dealing with all these sources of
450 variation in future versions of CAGEE.

451 In terms of the types of data that can be analyzed, at present CAGEE is limited to
452 positive, continuously varying traits (i.e. the BBM model). However, we also envision
453 different ways to represent and model gene expression data, including as a ratio (e.g.
454 male/female expression). Such a ratio, after log2-transformation, would be most
455 appropriately modeled by an unbounded Brownian motion model since both negative
456 and positive values are possible. This and other data types will be supported in future
457 releases. Moreover, CAGEE does not have to analyze whole-genome or even
458 molecular data: it can be applied to any single trait for which the BBM model is
459 appropriate, even morphological traits. One intriguing application of CAGEE could be to
460 suites of morphological traits that are hypothesized to share a common evolutionary
461 rate parameter. If, for instance, there is a shift in body plan along some lineages, then
462 multiple traits may all increase or decrease their rate of evolution at once, and CAGEE
463 can be used to estimate these shared parameters. Even in the context of single-trait
464 analyses, the pruning algorithm has been hailed as a solution for large-scale
465 comparative analyses (Freckleton 2012). Importantly, the number of branches in a
466 rooted, bifurcating tree with n tips is $2n$-2, so that the number of calculations scales
467 linearly with the number of species. This makes the pruning algorithm ideal for

468 comparative datasets with large numbers of taxa (e.g. Hahn et al. 2005; FitzJohn 2012;
469 Hiscott et al. 2016; Caetano and Harmon 2018; Mitov et al. 2020).

470 The analysis of data from a clade of wild tomatoes revealed a possibly
471 unexpected result: the rate of ovule gene expression evolution among self-compatible
472 (SC) species is twice as high as the rate among self-incompatible (SI) species (Table
473 1). This finding is contrary to some prior expectations—informed by research focused
474 on male-female interactions, especially between interacting proteins in the reproductive
475 tract (e.g. Swanson and Vacquier 2002; Clark et al. 2006)— that suggest that lineages
476 might experience slower evolution after transitioning to self-compatibility. However, it is
477 possible that global gene expression levels do not evolve in the same sort of tit-for-tat
478 manner as interacting protein sequences, such that increases/decreases in male-
479 expressed genes are not matched by increases/decreases in interacting female-
480 expressed genes (or vice versa). Alternatively, only a very small subset of genes may
481 evolve in this manner. Indeed, even prior studies comparing protein evolution have
482 failed to find clear evidence of slower global evolutionary rates in more inbreeding
483 species (e.g. Wong 2011). One caveat to the observed rate differences in our data is
484 that underlying gene tree discordance, whether due to incomplete lineage sorting or
485 introgression, can lead to artifactually higher rate estimates (Mendes et al. 2018;
486 Hibbins and Hahn 2021). However, there is in fact less discordance among the SC
487 lineages sampled here (Pease et al. 2016), which is the reverse of the pattern that
488 would be required to explain our results.

489 If not due to underlying bias in our estimates, these findings still raise the
490 question: why is ovule gene expression evolving more rapidly in SC than SI species?
491 One possibility is that this increased rate is due to a relaxation of selection in SC
492 species, possibly because genes involved in male-female interactions are no longer
493 needed. If this were the case, we might expect to see a general decrease in expression
494 levels in SC species; however, there appears to be no consistent directionality to the
495 changes along SC branches (Figure 3, Supplementary Figure 4). Instead, an alternative
496 hypothesis is that transitions to SC involve adaptation to new optima of ovule gene
497 expression, compared to SI species that tend to maintain ancestral optima. For
498 example, transitions to SC might favor greater investment in fewer ovules, because self-

compatibility decreases the probability that each ovule within a flower will go unfertilized—an otherwise wasted investment under conditions (like SI) where receiving sufficient compatible pollen to fertilize each ovule is less predictable (Burd et al. 2009). The nature of these new optima might be even more complex, as traits like ovule size and number can vary with multiple reproductive and ecological conditions, and often trade-off with each other (Greenway and Harder 2007). Of the species examine here, for example, two SC lineages (*S. pimpinellifolium*, and *S. lycopersicon*—domesticated tomato) have significantly larger seeds than most of the SI lineages and SC *S. pennellii* (unpubl. data). Indeed, individual genes identified in our GO analysis are known to directly influence ovule and/or seed size in *Solanum* (e.g. *NOR-like1* [SOLYC07G063420.3.1; Han et al, 2014], *GRAS2* [SOLYC07G063940.2.1; Li et al. 2018], and *CRY2* [SOLYC09G090100.3.1; Fantini et al. 2019]). Some of our hypotheses could be evaluated with matching gene expression data from other (non-ovule) reproductive tissues. Analyses including pollen in the same SI and SC lineages, and/or data addressing alternative constraints and conditions shaping ovule evolution including ovule size and number (e.g. Mione and Anderson 1992), would be useful in teasing apart these hypotheses.

Material and Methods

Bounded Brownian motion model of expression evolution

The probability density of expression, $p(x,t)$, at time t for evolutionary trajectories following a Brownian motion process starting at value x_{t_0} at time t_0 is governed by the diffusion equation

$$\frac{\partial p(x,t)}{\partial t} = \frac{\sigma^2}{2}\frac{\partial^2 p(x,t)}{\partial x^2}, \qquad (2)$$

with initial condition $p(x,t_0) = \delta(x - x_{t_0})$ where δ is the Dirac delta function. The reflective boundary condition at $x = x_{\min}$ implies that the probability fluxes into and out of the boundary are balanced, imposing the boundary condition

$$\frac{\partial p(x{=}x_{\min},t)}{\partial x} = 0. \qquad (3)$$

Note that $p(x,t)$ is identical to the transition density $p(x_t|x_{t_0})$.

Without the reflecting boundary, $p(x,t) \propto e^{-(x-x_{t_0})^2/2\sigma^2(t-t_0)}$ is a normal distribution with variance $\sigma^2(t-t_0)$. The variance therefore scales linearly with elapsed time, $t - t_0$. With the reflecting boundary, $p(x,t)$ is the sum of this spreading normal and its mirror image centered at $2x_{\min} - x_{t_0}$. The analytical solution to this bounded process is helpful for understanding the behavior of $p(x,t)$, but is not used in CAGEE. In anticipation of implementing additional (and possibly more complicated) processes into CAGEE, we instead solve Eq. (2) numerically using the approach described in Boucher and Démery (2016). Briefly, the continuous diffusion equation is converted into a matrix equation by discretizing expression values into N equal bins of width $\delta = \frac{x_{\max}-x_{\min}}{N-1}$. Following Boucher and Démery (2016), we have used a default *N*=200, but this number can be set by the user (see Results). This approach gives

$$\frac{\partial P(t)}{\partial t} = \frac{\sigma^2}{2\delta^2} M \cdot P(t), \qquad (4)$$

where $P(t)$ is the vector obtained by discretizing $p(x,t)$ and $x_{\max}$ is the largest expression value accounted for. The matrix M is tridiagonal with -2 on the diagonal

543 except at the first and last diagonal entries which are -1. The sub- and supra-diagonal
544 entries are 1. This equation has the matrix exponential solution

545 $$P(t) = \exp\left(\frac{\sigma^2(t-t_0)}{2\delta^2} M\right), \quad (5)$$

546 which is evaluated by diagonalizing M.

547 *Implementation of CAGEE*

548 CAGEE is written in C++ and is compatible with the C++11 standard. A
549 comprehensive manual and extensive unit tests facilitate further code development and
550 maintenance. CAGEE is organized into modular components. A *clade* class, with
551 references to a parent clade and any number of descendant clades, represents a tree
552 structure, and a *gene_transcript* class represents the expression levels observed in the
553 various species. These two classes comprise the fundamental data structures upon
554 which CAGEE performs its analysis (Supplementary Figure 5).

555 Calculations are carried out by additional classes. The *optimizer* class has the
556 responsibility of determining the σ^2 value with the highest likelihood, by comparing the
557 likelihood of candidate values and searching the likelihood surface using the Nelder-
558 Mead optimization algorithm. The work of computing the likelihood of a given σ^2 value is
559 performed by a subclass of the *model* class, which for now is limited to a single *Base*
560 model (allowing for further development in the future). After appropriate estimated
561 values are found, the *transcript_reconstructor* class builds a possible set of transcript
562 values for the entire tree (Supplementary Figure 5).

563 Performing the likelihood calculations requires extensive matrix operations; it is
564 recommended (though not required) that these be passed off to a specialized library
565 such as Intel's MKL or Nvidia's CUBLAS. If no external library is available, CAGEE will
566 carry out these calculations (slowly) by itself. Creating the diffusion matrix (M) requires
567 calculation of eigenvalues and eigenvectors, and is computationally expensive. This
568 work is performed by the Eigen linear algebra library (https://eigen.tuxfamily.org);
569 various internal data structures also take advantage of Eigen classes. To enable faster
570 searching, the matrix for an ancestral state vector of length 200 (the default in CAGEE)

571 has been pre-computed and is included with CAGEE. Users who wish to use vectors of
572 different lengths can specify this as an option.

573 Unit-testing is performed using the Doctest testing framework
574 (https://github.com/doctest/doctest). At the time of writing more than 200 unit tests had
575 been created, comprising more than 1200 individual assertions. For complex logging
576 and debugging cases, CAGEE uses the EasyLogging framework
577 (https://github.com/amrayn/easyloggingpp). C++ development is always made easier by
578 using the Boost C++ libraries (https://www.boost.org/), so we include them as well in
579 CAGEE.

580 *RNA-seq data from wild tomatoes*

581 We briefly describe here the data collected from seven accessions of wild
582 tomatoes (*S. lycopersicum* LA3475, *S. chmielewskii* LA1316, *S. pimpinellifolium*
583 LA1589, *S. habrochaites* LA1777, *S. chilense* LA4117A, *S. pennellii* LA3778, and *S.*
584 *pennellii* LA0716; all accession ID numbers from tgrc.ucdavis.edu). Further details are
585 given in Moyle et al. (2021). Ovule RNA-seq was performed on between one to four
586 (usually three) biological replicates (individual plants) from each accession. Plants were
587 germinated from seed, and cultivated until flowering. For each replicate individual,
588 ovules were dissected from mature, unpollinated flowers, flash frozen, and maintained
589 at -80C until extraction. For each individual, all ovule collections were pooled into a
590 single sample prior to library construction and sequencing on an Illumina HiSeq 2000.
591 Reads were mapped against the tomato reference genome (ITAG 2.4) and the number
592 of reads mapped onto genic regions were estimated with featureCounts (Liao et al.,
593 2014). We normalized the read counts from each library by calculating TPM (transcripts
594 per million; Wagner et al. 2012) and then calculated the mean normalized read counts
595 across all samples (individuals) within each accession. These means per accession
596 were used as input to CAGEE.

597 To construct a species tree for use with CAGEE, we started with the topology
598 given in Pease et al. (2016). Specifically, we used the tree found in the supplementary
599 file Pease_etal_TomatoPhylo_RAxMLConcatTree_no1360_Fig2A.nwk, and pruned it to
600 include only the accessions in our study using the software ETE (Huerta-Cepas et al.

601 2016). Using the “extend” method found in ETE, we converted this tree to ultrametric
602 (same root-to-tip distance for all taxa). Setting the root age to 2.48 million years ago
603 (following Pease et al. 2016), we were able to express all branches in millions of years.
604 Analyses of GO enrichment were carried out using ShinyGO (Ge et al. 2020) with a
605 false discovery rate of 0.05.

606

607

608 **Supplementary Material**

609 Supplementary data are available at .

610

611 **Acknowledgements**

612 We thank Mark Hibbins for assistance with the tomato phylogeny, Matthew Gibson for
613 putting together the tomato gene expression data, and especially Dan Vanderpool for
614 invaluable help in the initial development of CAGEE. Two reviewers provided helpful
615 comments, and Scott Edwards pointed out relevant work that we had previously missed.
616 This work was supported by National Science Foundation grants DEB-1856469 to
617 L.C.M. and DBI-2146866 to M.W.H.

618

619 **Data Availability**

620 Raw reads for each sample library are available at NCBI BioProject PRJNA714065. The
621 CAGEE software is available at https://github.com/hahnlab/CAGEE.

622 **References**

623 Ames RM, Money D, Ghatge VP, Whelan S, Lovell SC. 2012. Determining the
624 evolutionary history of gene families. Bioinformatics 28:48-55.

625 Beaulieu JM, Jhwueng DC, Boettiger C, O'Meara BC. 2012. Modeling stabilizing
626 selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution
627 66:2369-2383.

628 Bedford T, Hartl DL. 2009. Optimization of gene expression by natural selection.
629 Proceedings of the National Academy of Sciences 106:1133-1138.

630 Bedinger PA, Chetelat RT, McClure B, Moyle LC, Rose JK, Stack SM, van der Knaap E,
631 Baek YS, Lopez-Casado G, Covey PA. 2011. Interspecific reproductive barriers in
632 the tomato clade: opportunities to decipher mechanisms of reproductive isolation.
633 Sexual Plant Reproduction 24:171-187.

634 Berthelot C, Villar D, Horvath JE, Odom DT, Flicek P. 2018. Complexity and
635 conservation of regulatory landscapes underlie evolutionary resilience of mammalian
636 gene expression. Nature Ecology & Evolution 2:152-163.

637 Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, Eres I,
638 Cuevas C, Marques-Bonet T, Gilad Y. 2020. A comparison of gene expression and
639 DNA methylation patterns across tissues and species. Genome Research 30:250-
640 262.

641 Boucher FC, Démery V. 2016. Inferring bounded evolution in phenotypic characters
642 from phylogenetic comparative data. Systematic Biology 65:651-661.

643 Boucher FC, Démery V, Conti E, Harmon LJ, Uyeda J. 2018. A general model for
644 estimating macroevolutionary landscapes. Systematic Biology 67:304-319.

645 Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti
646 A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in
647 mammalian organs. Nature 478:343-348.

648 Burd M, Ashman T-L, Campbell DR, Dudash MR, Johnston MO, Knight TM, Mazer SJ,
649 Mitchell RJ, Steets JA, Vamosi JC. 2009. Ovule number per flower in a world of
650 unpredictable pollination. American Journal of Botany 96:1159-1167.

651 Burnham KP, Anderson DR. 2002. Model selection and multimodel inference: A
652 practical information-theoretic approach. New York: Springer.

653 Caetano DS, Harmon LJ. 2018. Estimating correlated rates of trait evolution with uncertainty.
654 Systematic Biology 68:412-429.

655 Catalán A, Briscoe AD, Höhna S. 2019. Drift and directional selection are the
656 evolutionary forces driving gene expression divergence in eye and brain tissue of
657 *Heliconius* butterflies. Genetics 213:581-594.

658 Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. Current
659 Opinion in Genetics & Development 11:685-690.

660 Chen J, Swofford R, Johnson J, Cummings BB, Rogel N, Lindblad-Toh K, Haerty W, Di
661 Palma F, Regev A. 2019. A quantitative framework for characterizing the
662 evolutionary history of mammalian gene expression. Genome Research 29:53-63.

663 Clark NL, Aagaard JE, Swanson WJ. 2006. Evolution of reproductive proteins from
664 animals and plants. Reproduction 131:11-22.

665 Coolon JD, McManus CJ, Stevenson KR, Graveley BR, Wittkopp PJ. 2014. Tempo and
666 mode of regulatory evolution in *Drosophila*. Genome Research 24:797-808.

667 Cooper N, Thomas GH, Venditti C, Meade A, Freckleton RP. 2016. A cautionary note
668 on the use of Ornstein Uhlenbeck models in macroevolutionary studies. Biological
669 Journal of the Linnean Society 118:64-77.

670 De Bie T, Demuth JP, Cristianini N, Hahn MW. 2006. CAFE: a computational tool for the
671 study of gene family evolution. Bioinformatics 22:1269-1271.

672 Dunn CW, Luo X, Wu Z. 2013. Phylogenetic analysis of gene expression. Integrative
673 and Comparative Biology 53:847-856.

674 El Taher A, Böhne A, Boileau N, Ronco F, Indermaur A, Widmer L, Salzburger W. 2021.
675 Gene expression dynamics during rapid organismal diversification in African cichlid
676 fishes. Nature Ecology & Evolution 5:243-250.

677 Fantini E, Sulli M, Zhang L, Aprea G, Jiménez-Gómez JM, Bendahmane A, Perrotta G,
678 Giuliano G, Facella P. 2018. Pivotal roles of cryptochromes 1a and 2 in tomato
679 development and physiology. Plant Physiology 179:732-748.

680 Fay JC, Wittkopp PJ. 2008. Evaluating the role of natural selection in the evolution of
681 gene regulation. Heredity 100:191-199.

682 Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating
683 evolutionary trees from data on discrete characters. Systematic Biology 22:240-249.

684 FitzJohn RG. 2012. Diversitree: comparative phylogenetic analyses of diversification in
685 R. Methods in Ecology and Evolution 3:1084-1092.

686 Freckleton RP. 2012. Fast likelihood calculations for comparative analyses. Methods in
687 Ecology and Evolution 3:940-947.

688 Ge SX, Jung D, Yao R. 2020. ShinyGO: a graphical gene-set enrichment tool for
689 animals and plants. Bioinformatics 36:2628-2629.

690 Gillespie JH. 1986. Variability of evolutionary rates of DNA. Genetics 113:1077-1091.

691 Goldberg EE, Kohn JR, Lande R, Robertson KA, Smith SA, Igić B. 2010. Species
692 selection maintains self-incompatibility. Science 330:493-495.

693 Golding G. 1983. Estimates of DNA and protein sequence divergence: an examination
694 of some assumptions. Molecular Biology and Evolution 1:125-142.

695 Gossmann TI, Saleh D, Schmid MW, Spence MA, Schmid KJ. 2016. Transcriptomes of
696 plant gametophytes have a higher proportion of rapidly evolving and young genes
697 than sporophytes. Molecular Biology and Evolution 33:1669-1678.

698 Gossmann TI, Schmid MW, Grossniklaus U, Schmid KJ. 2014. Selection-driven
699 evolution of sex-biased genes is consistent with sexual selection in *Arabidopsis*
700 *thaliana*. Molecular Biology and Evolution 31:574-583.

701 Greenway CA, Harder LD. 2007. Variation in ovule and seed size and associated size–
702 number trade-offs in angiosperms. American Journal of Botany 94:840-846.

703 Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo
704 and mode of gene family evolution from comparative genomic data. Genome
705 Research 15:1153-1160.

706 Hahn MW, Demuth JP, Han S-G. 2007. Accelerated rate of gene gain and loss in
707 primates. Genetics 177:1941-1949.

708 Hahn MW, Nakhleh L. 2016. Irrational exuberance for resolved species trees. Evolution
709 70:7-17.

710 Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. 2013. Estimating gene gain and
711 loss rates in the presence of error in genome assembly and annotation using CAFE
712 3. Molecular Biology and Evolution 30:1987-1997.

713 Han QQ, Song YZ, Zhang JY, Liu LF. 2014. Studies on the role of the *SlNAC3* gene in
714 regulating seed development in tomato (*Solanum lycopersicum*). The Journal of
715 Horticultural Science and Biotechnology 89:423-429.

716 Harrison MC, Mallon EB, Twell D, Hammond RL. 2019. Deleterious mutation
717 accumulation in *Arabidopsis thaliana* pollen genes: a role for a recent relaxation of
718 selection. Genome Biology and Evolution 11:1939-1951.

719 Harrison PW, Wright AE, Zimmer F, Dean R, Montgomery SH, Pointer MA, Mank JE.
720 2015. Sexual selection drives evolution and rapid turnover of male gene expression.
721 Proceedings of the National Academy of Sciences 112:4393-4398.

722 Hibbins MS, Hahn MW. 2021. The effects of introgression across thousands of
723 quantitative traits revealed by gene expression in wild tomatoes. PLoS Genetics
724 17:e1009892.

725 Hill MS, Zande PV, Wittkopp PJ. 2021. Molecular and evolutionary processes
726 generating variation in gene expression. Nature Reviews Genetics 22:203-215.

727 Hiscott G, Fox C, Parry M, Bryant D. 2016. Efficient recycled algorithms for quantitative
728 trait models on phylogenies. Genome Biology and Evolution 8:1338-1350.

729 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and
730 visualization of phylogenomic data. Molecular Biology and Evolution 33:1635-1638.

731 Igić B, Bohs L, Kohn JR. 2006. Ancient polymorphism reveals unidirectional breeding
732 system shifts. Proceedings of the National Academy of Sciences 103:1359-1363.

733 King M-C, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees.
734 Science 188:107-116.

735 Landis MJ, Schraiber JG. 2017. Pulsed evolution shaped modern vertebrate body sizes.
736 Proceedings of the National Academy of Sciences 114:13224-13229.

737 Li M, Wang X, Li C, Li H, Zhang J, Ye Z. 2018. Silencing *GRAS2* reduces fruit weight in
738 tomato. Journal of Integrative Plant Biology 60:498-513.

739 Liao Y, Smyth GK, Shi W. 2013. featureCounts: an efficient general purpose program
740 for assigning sequence reads to genomic features. Bioinformatics 30:923-930.

741 Meisel RP, Malone JH, Clark AG. 2012. Disentangling the relationship between sex-
742 biased gene expression and X-linkage. Genome Research 22:1255-1265.

743 Mendes FK, Fuentes-González JA, Schraiber JG, Hahn MW. 2018. A multispecies
744 coalescent model for quantitative traits. eLife 7:e36482.

745 Mendes FK, Vanderpool D, Fulton B, Hahn MW. 2020. CAFE 5 models variation in
746 evolutionary rates among gene families. Bioinformatics 36:5516-5518.

747 Mione T, Anderson GJ. 1992. Pollen-ovule ratios and breeding system evolution in
748 *Solanum* section *Basarthrum* (Solanaceae). American Journal of Botany 79:279-
749 287.

750 Mitov V, Bartoszek K, Asimomitis G, Stadler T. 2020. Fast likelihood calculation for
751 multivariate Gaussian phylogenetic models with shifts. Theoretical Population
752 Biology 131:66-78.

753 Moyle LC, Wu M, Gibson MJ. 2021. Reproductive proteins evolve faster than non-
754 reproductive proteins among *Solanum* species. Frontiers in Plant Science
755 12:635990.

756 O'Meara BC, Ané C, Sanderson MJ, Wainwright PC. 2006. Testing for different rates of
757 continuous trait evolution using likelihood. Evolution 60:922-933.

758 Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics reveals three sources
759 of adaptive variation during a rapid radiation. PLoS Biology 14:e1002379.

760 Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, Alfaro ME,
761 Harmon LJ. 2014. geiger v2. 0: an expanded suite of methods for fitting
762 macroevolutionary models to phylogenetic trees. Bioinformatics 30:2216-2218.

763 Pennell MW, FitzJohn RG, Cornwell WK, Harmon LJ. 2015. Model adequacy and the
764 macroevolution of angiosperm functional traits. The American Naturalist 186:E33-
765 E50.

766 Price PD, Palmer Droguett DH, Taylor JA, Kim DW, Place ES, Rogers TF, Mank JE,
767 Cooney CR, Wright AE. 2022. Detecting signatures of selection on gene expression.
768 Nature Ecology & Evolution 6:1035-1045.

769 Revell LJ, Harmon LJ. 2022. Phylogenetic Comparative Methods in R: Princeton
770 University Press.

771 Rick CM, Tanksley SD. 1981. Genetic variation in *Solanum pennellii*: Comparisons with
772 two other sympatric tomato species. Plant Systematics and Evolution 139:11-45.

773 Robinson MD, Oshlack A. 2010. A scaling normalization method for differential
774 expression analysis of RNA-seq data. Genome Biology 11:R25.

775 Rohlfs RV, Harrigan P, Nielsen R. 2014. Modeling gene expression evolution with an
776 extended Ornstein–Uhlenbeck process accounting for within-species variation.
777 Molecular Biology and Evolution 31:201-211.

778 Rohlfs RV, Nielsen R. 2015. Phylogenetic ANOVA: the expression variance and
779 evolution model for quantitative trait evolution. Systematic Biology 64:695-708.

780 Silvestro D, Kostikova A, Litsios G, Pearman PB, Salamin N. 2015. Measurement errors
781 should always be incorporated in phylogenetic comparative analysis. Methods in
782 Ecology and Evolution 6:340-346.

783 Somel M, Creely H, Franz H, Mueller U, Lachmann M, Khaitovich P, Pääbo S. 2008.
784 Human and chimpanzee gene expression differences replicated in mice fed different
785 diets. PLoS ONE 3:e1504.

786 Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nature
787 Reviews Genetics 3:137-144.

788 Szövényi P, Ricca M, Hock Z, Shaw JA, Shimizu KK, Wagner A. 2013. Selection is no
789 more efficient in haploid than in diploid life stages of an angiosperm and a moss.
790 Molecular Biology and Evolution 30:1929-1939.

791 Tarashansky AJ, Musser JM, Khariton M, Li P, Arendt D, Quake SR, Wang B. 2021.
792 Mapping single-cell atlases throughout Metazoa unravels cell type evolution. eLife
793 10:e66747.

794 Vosters SL, Jewell CP, Sherman NA, Einterz F, Blackman BK, Moyle LC. 2014. The
795 timing of molecular and morphological changes underlying reproductive transitions in
796 wild tomatoes (*Solanum* sect. *Lycopersicon*). Molecular Ecology 23:1965-1978.

797 Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq
798 data: RPKM measure is inconsistent among samples. Theory in Biosciences
799 131:281-285.

800 Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for
801 transcriptomics. Nature Reviews Genetics 10:57-63.

802 Whitehead MR, Lanfear R, Mitchell RJ, Karron JD. 2018. Plant mating systems often
803 vary widely among populations. Frontiers in Ecology and Evolution 6:38.

804 Wong A. 2011. The molecular evolution of animal reproductive tract proteins: What
805 have we learned from mating-system comparisons? International Journal of
806 Evolutionary Biology 2011:908735.

807 Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA.
808 2003. The evolution of transcriptional regulation in eukaryotes. Molecular Biology
809 and Evolution 20:1377-1419.

810 Yang J, Ruan H, Xu W, Gu X. 2019. TreeExp2: an integrated framework for
811 phylogenetic transcriptome analysis. Genome Biology and Evolution 11:3276-3282.

Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution 11:367-372.

Figures and Tables

Figure 1. Bounded Brownian motion model. An example trait is shown in the bottom graph, evolving along the tree shown above. Although the data input to CAGEE are linear expression levels, internally it logs expression to ensure higher variance among more highly expressed genes. There is also a minimum value, x_{min}, added to all tips.

Figure 2. Accuracy of CAGEE. For five different values of σ^2 we simulated 1000 datasets, with each dataset comprised of either 1000 genes or 10000 genes. All genes in a dataset shared the same σ^2, but their values at the root were drawn independently from the prior. We then provided each simulated dataset to CAGEE in order to infer σ^2. Each box-and-whisker plot shows the mean (horizontal line), 50% interquartile range (box), 1.5X the interquartile range (vertical lines), and outliers (dots).

Figure 3. Changes in gene expression along the tomato phylogeny. Given the set of relationships among the seven *Solanum* accessions used here, we tested multiple models that had branches assigned as different σ^2 parameters (Table 1). In model A, all branches share σ_1^2. In model B, all black branches share σ_1^2, while all red branches share σ_2^2. In model C, all black branches and the dashed red branch share σ_1^2, while all solid red branches share σ_2^2. In model D, all black branches share σ_1^2, all solid red branches share σ_2^2, and the dashed red branch is assigned σ_3^2. Using the results from model D, we inferred the number of genes that had credible increases or decreases in expression level along each branch (results for all changes are shown in Supplementary Figure 4). Numbers are reported as +increases/-decreases for each branch.

840 **Table 1.** Model parameters estimated from the tomato data.
841

Model	Number of rates	-lnL	σ^2_1	σ^2_2	σ^2_3
A	1	67252.4	0.102		
B	2	65883.9	0.074	0.134	
C	2	65124.5	0.075	0.152	
D	3	65108.6	0.077	0.152	0.067

842

843

Supplementary Figures

Supplementary Figure 1. The tree used for simulations. The Newick-formatted tree string with branch lengths is:
((((sp1:1,sp2:1):1,sp3:2):1,sp4:3):1,((sp5:2,sp6:2):2)):1,sp7:5):1,sp8:6)

Supplementary Figure 2. Accuracy of CAGEE. A) This figure is the same as Figure 2 in the main text, but the ancestral state vector has length *N*=500 (Figure 2 uses *N*=200). B) For each of three different simulated values of σ^2, we randomly removed different amounts of data from an input dataset with 1,000 genes (the tree is the same as in all other simulations). C) For 1,000 genes simulated with σ^2=1 (dashed vertical line), we ran CAGEE independently on each one to estimate σ^2. D) We combined three datasets of 1,000 genes each simulated with three different values of σ^2 (we repeated these simulations 10 times). The 10 estimates of σ^2 on the combined datasets were slightly downwardly biased compared to the expected value (dashed horizontal line). Each dot represents each of the 10 estimates, with jitter added for clarity,

Supplementary Figure 3. Trees used for simulations with lineage-specific values of σ^2.

A) All black branches share a rate parameter (σ_1^2), and all red branches share a rate parameter (σ_2^2). This "sigma_tree" is specified in CAGEE with the Newick string:
((((sp1:2,sp2:2):2,sp3:2):2,sp4:2):2,((sp5:1,sp6:1):1)):1,sp7:1):1,sp8:1)

B) All black branches share a rate parameter (σ_1^2), and all red branches share a rate parameter (σ_2^2). This "sigma_tree" is specified in CAGEE with the Newick string:
((((sp1:2,sp2:1):1,sp3:2):1,sp4:1):1,((sp5:1,sp6:2):1)):1,sp7:1):1,sp8:1)

Supplementary Figure 4. Changes in gene expression along the tomato phylogeny. This figure is the same as Figure 3 in the main text, but all increases and decreases are reported, regardless of whether they are "credible".

872 **Supplementary Figure 5.** Component diagram for the CAGEE software.

Figure 1

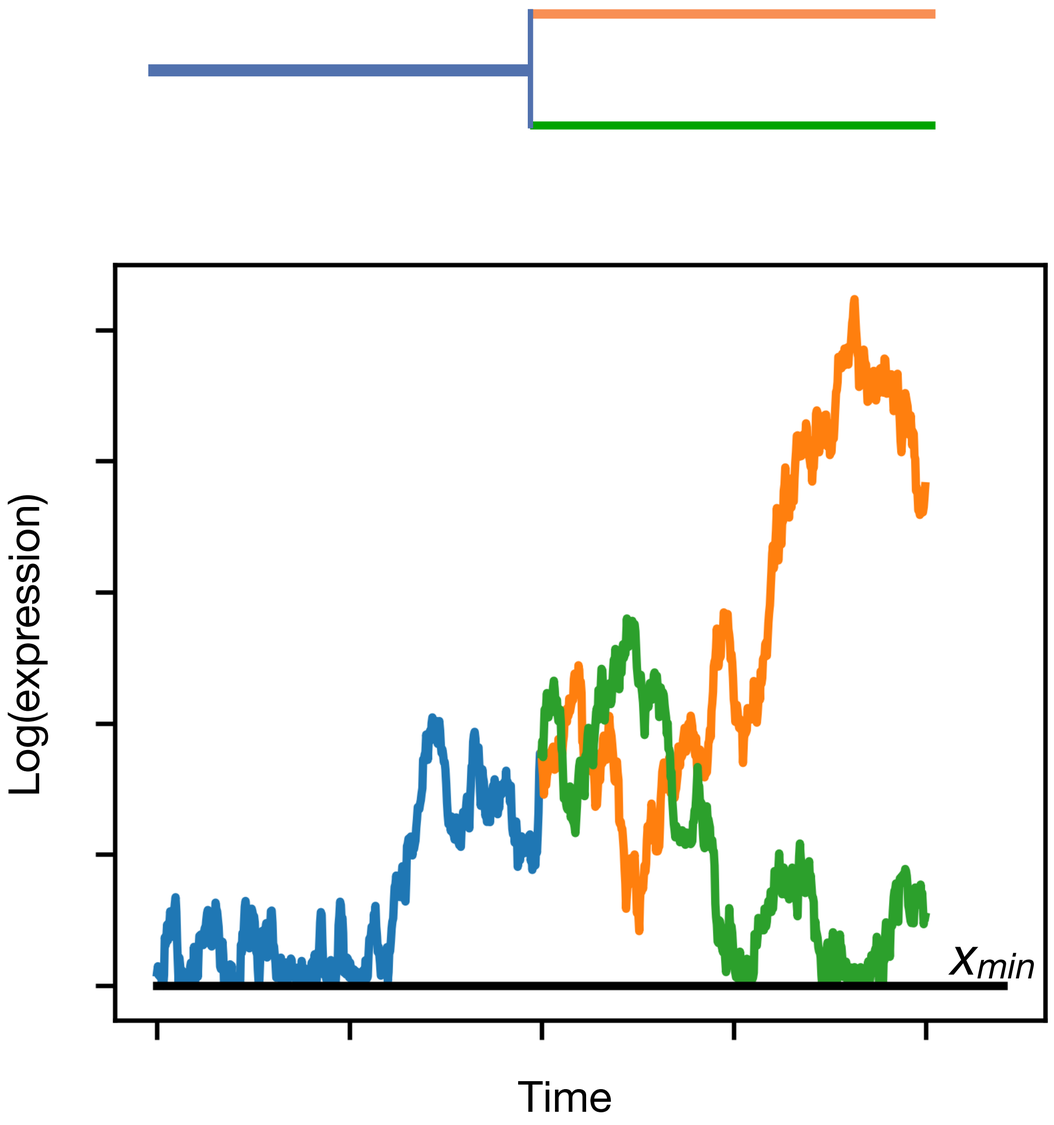


Figure 2

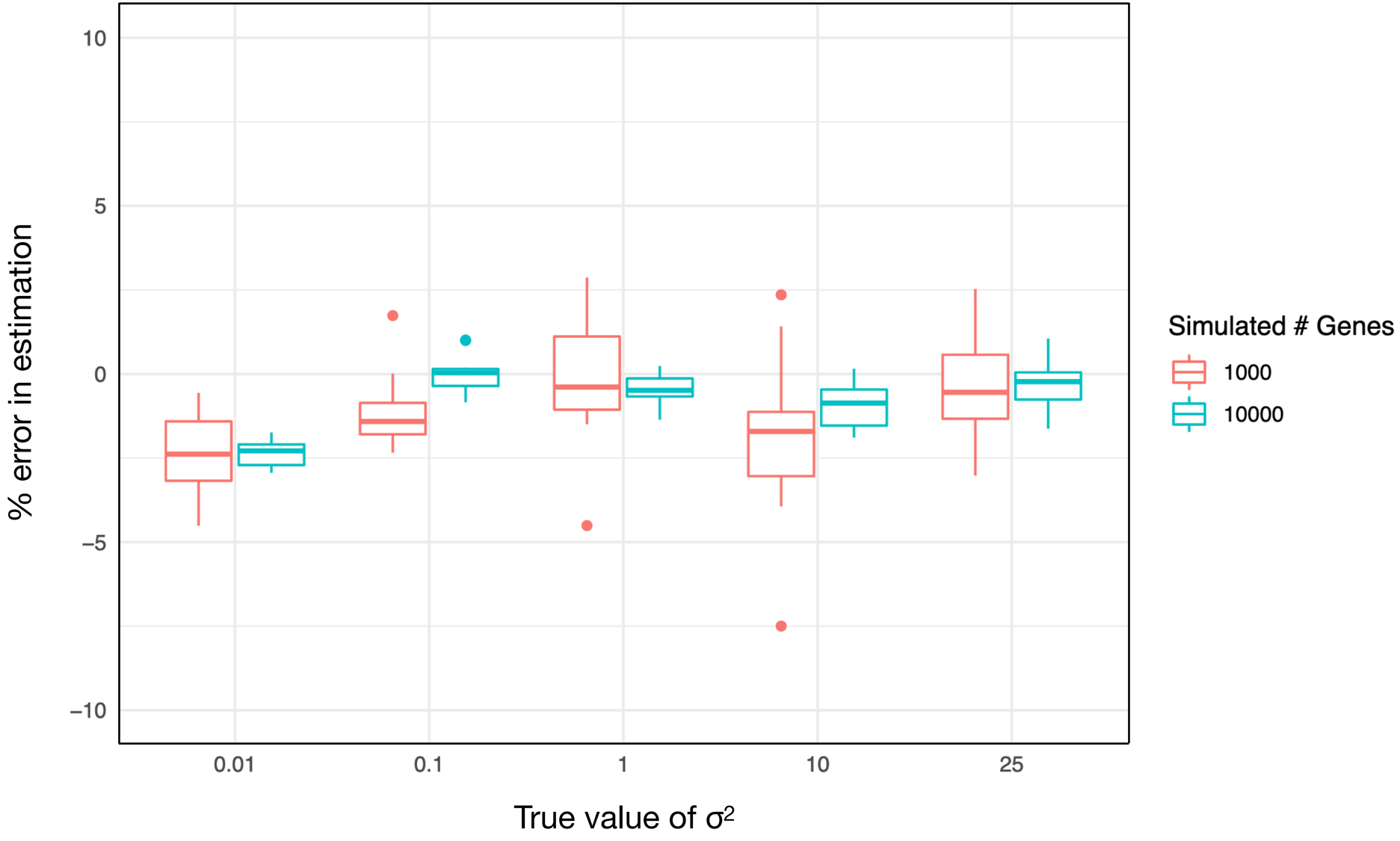

10
5
0
−5
−10
% error in estimation
0.01
0.1
1
10
25
True value of σ2
Simulated # Genes
1000
10000

Figure 3

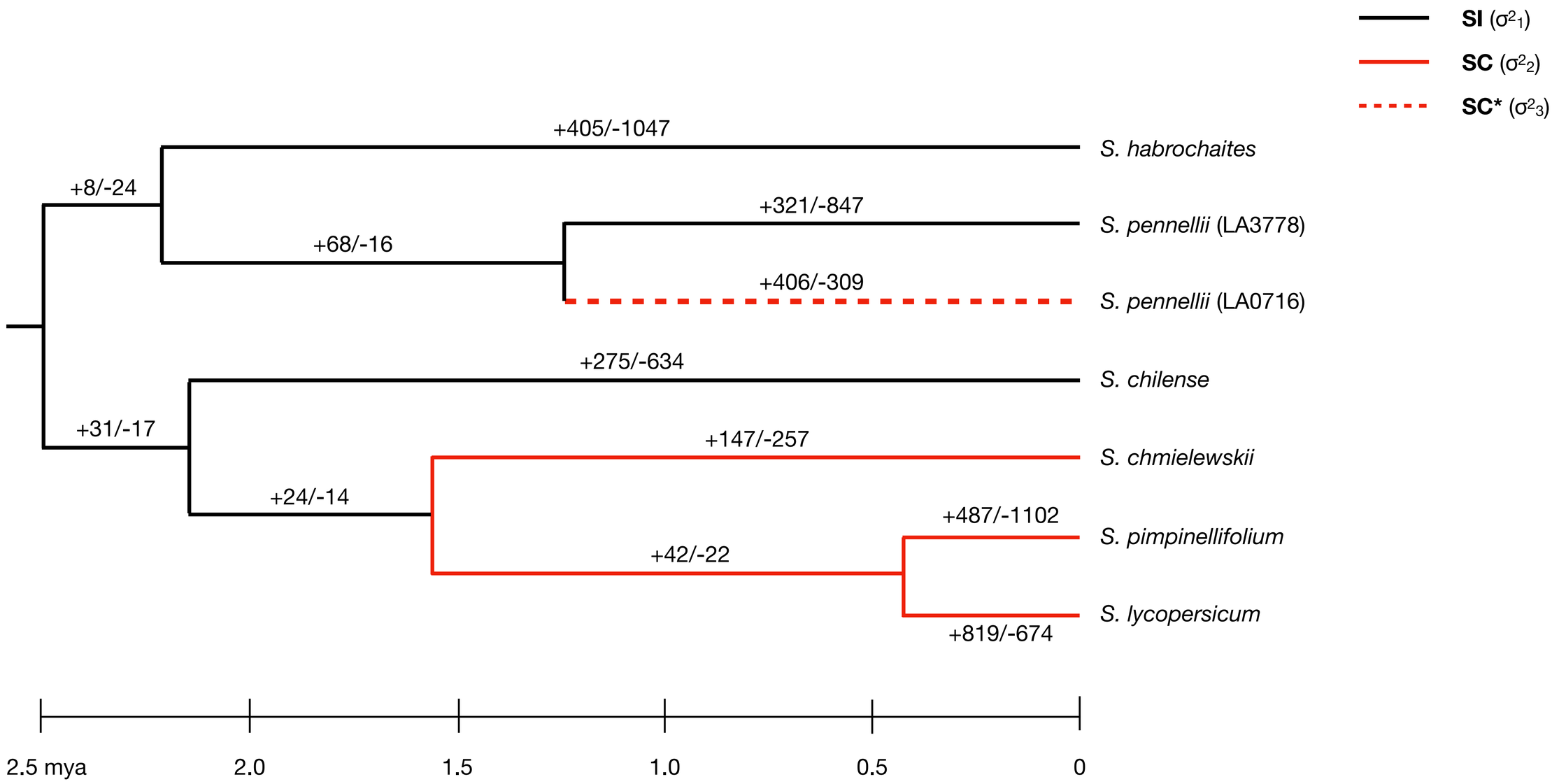
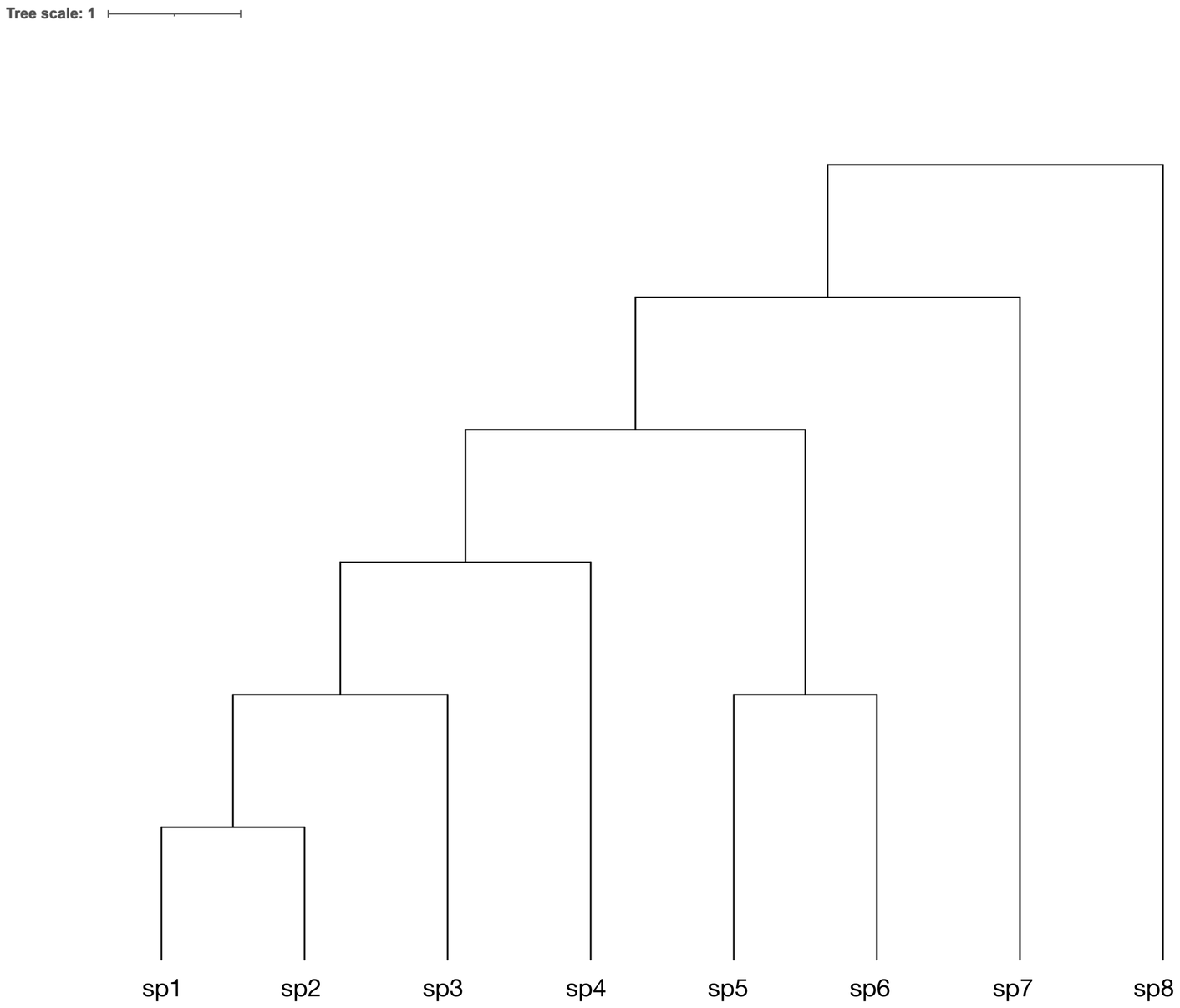

SI (σ^2_1)
SC (σ^2_2)
SC* (σ^2_3)
+405/-1047
+8/-24
+321/-847
+68/-16
+406/-309
+275/-634
+31/-17
+147/-257
+24/-14
+487/-1102
+42/-22
+819/-674
S. habrochaites
S. pennellii (LA3778)
S. pennellii (LA0716)
S. chilense
S. chmielewskii
S. pimpinellifolium
S. lycopersicum
2.5 mya
2.0
1.5
1.0
0.5
0

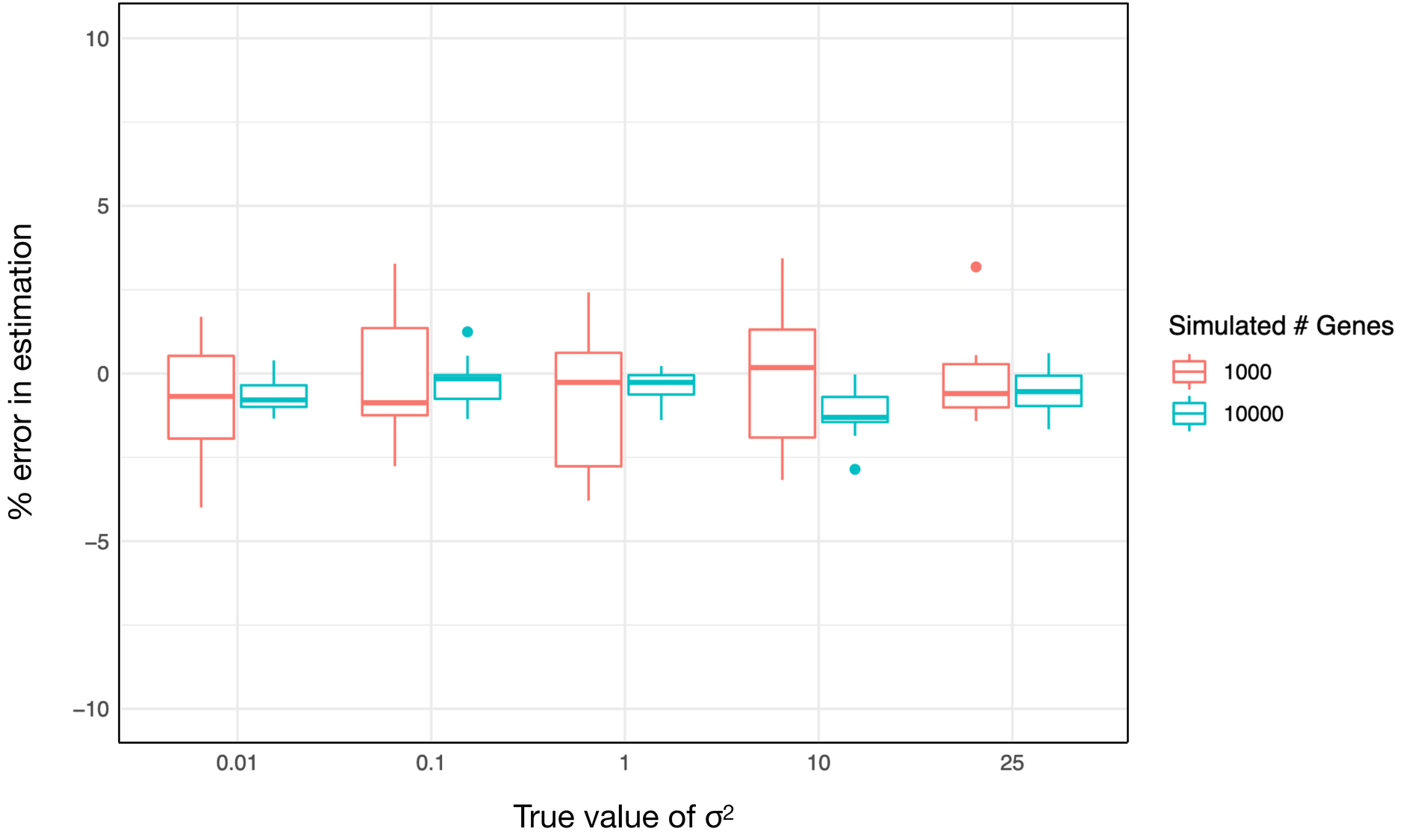
Table 1

Model	Number of rates	-lnL	σ^2_1	σ^2_2	σ^2_3
A	1	67252.4	0.102		
B	2	65883.9	0.074	0.134	
C	2	65124.5	0.075	0.152	
D	3	65108.6	0.077	0.152	0.067

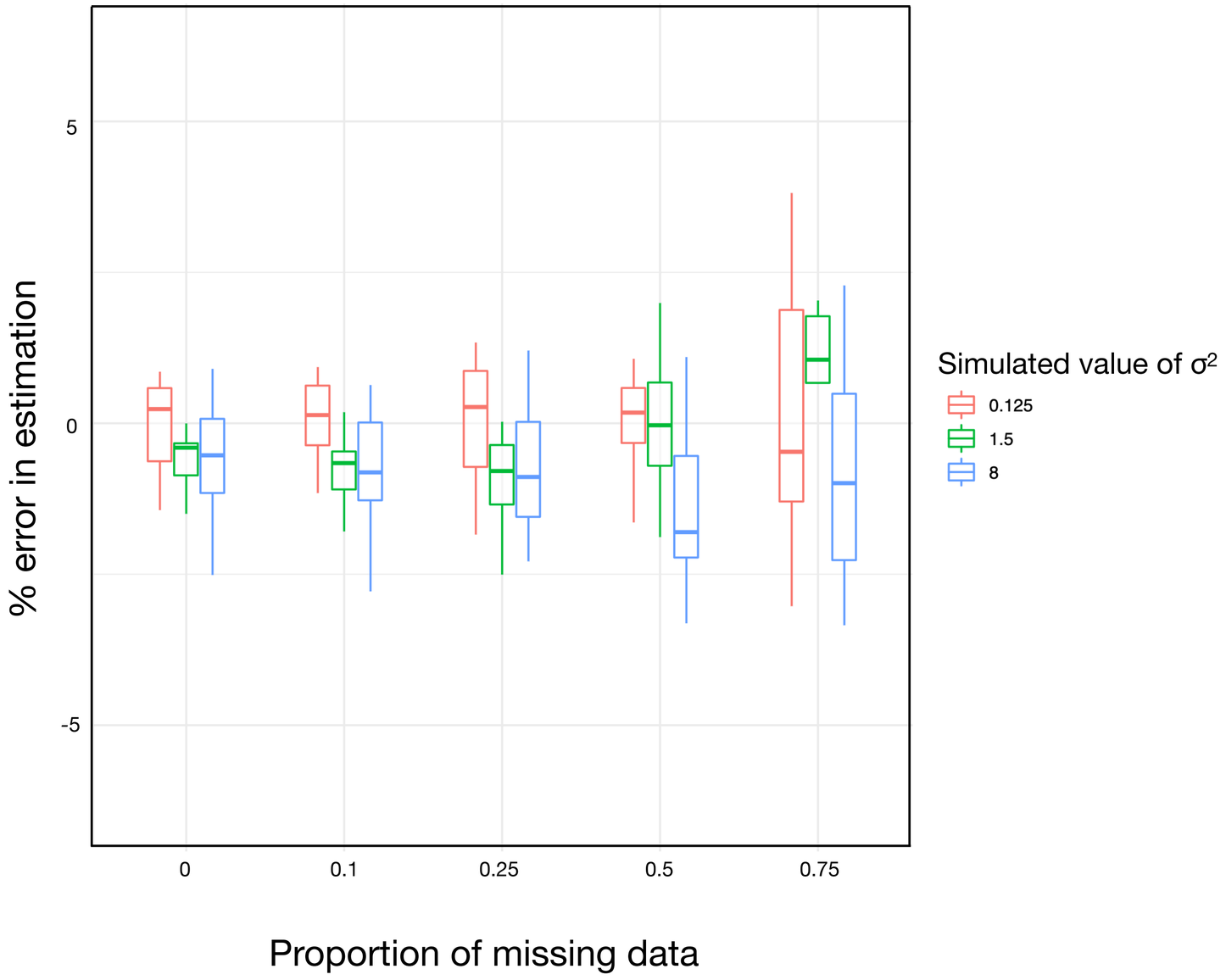
Supplementary Figure 1



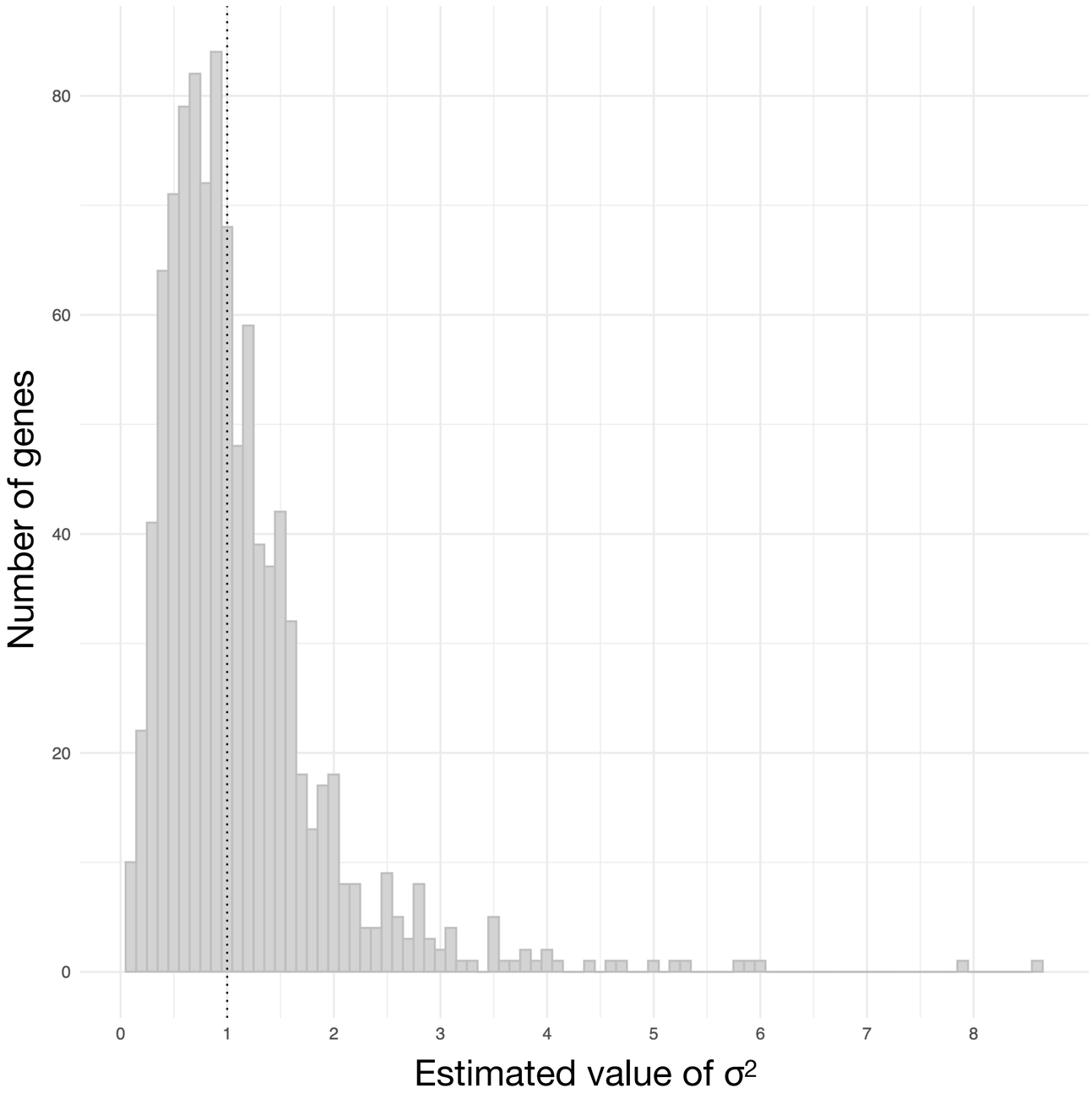
Supplementary Figure 2A



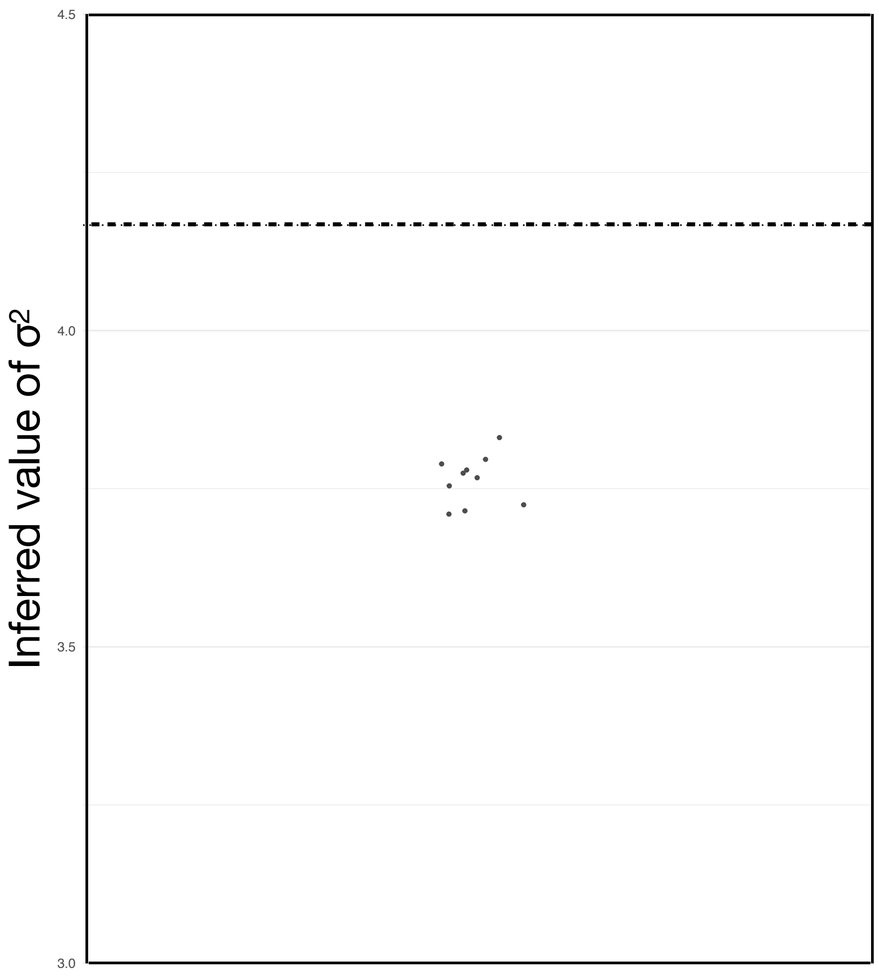
Supplementary Figure 2B



Supplementary Figure 2C

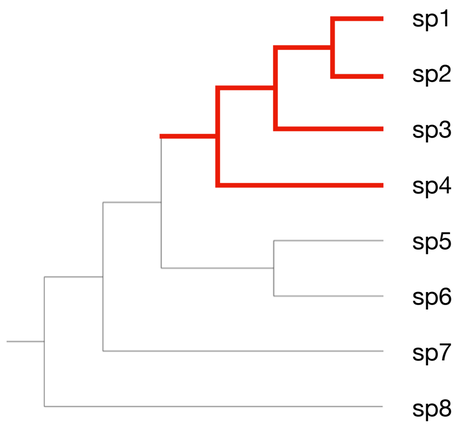


Supplementary Figure 2D

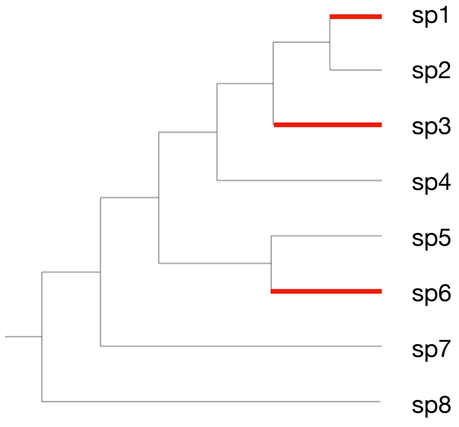


Supplementary Figure 3

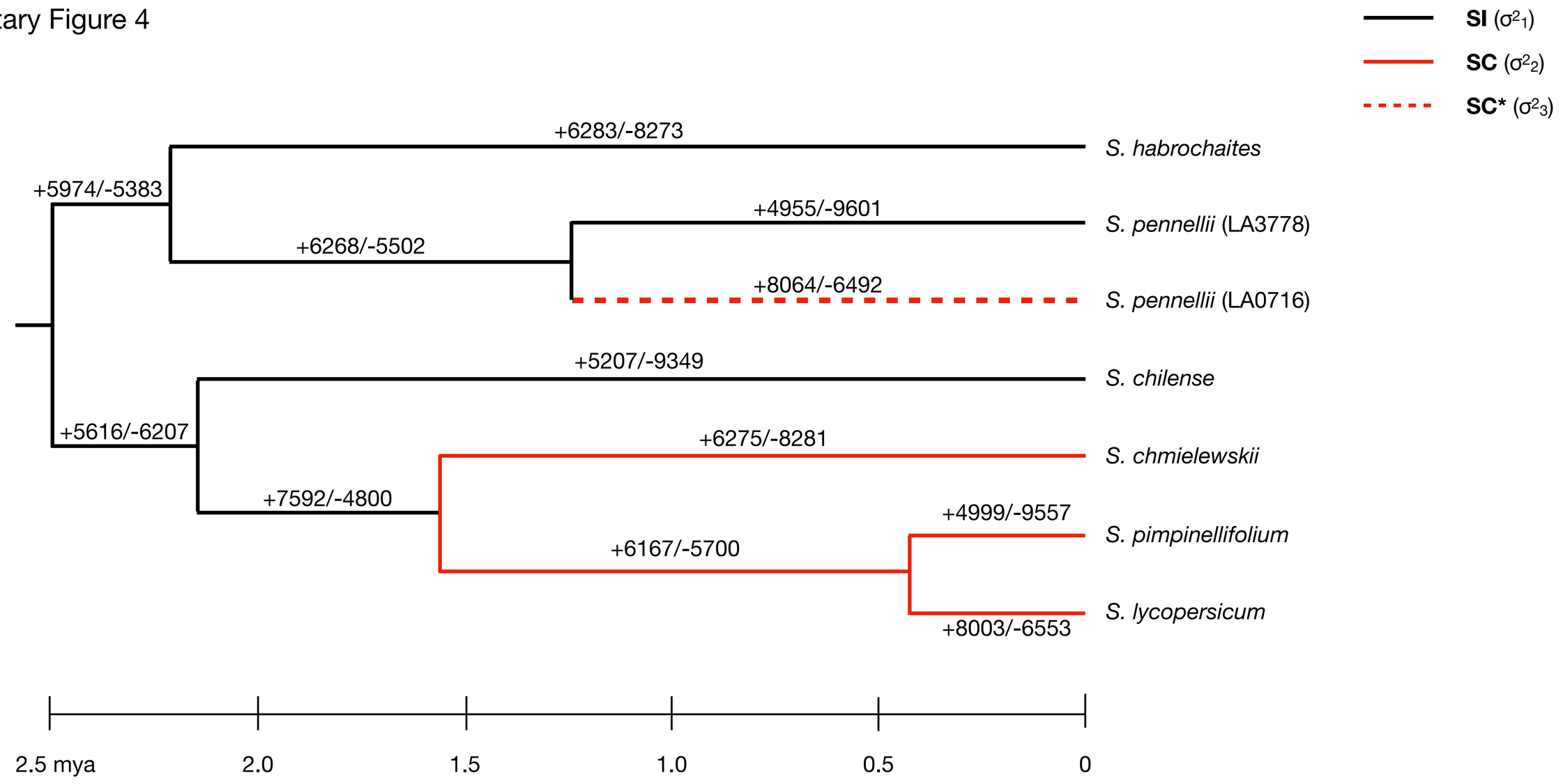
A)



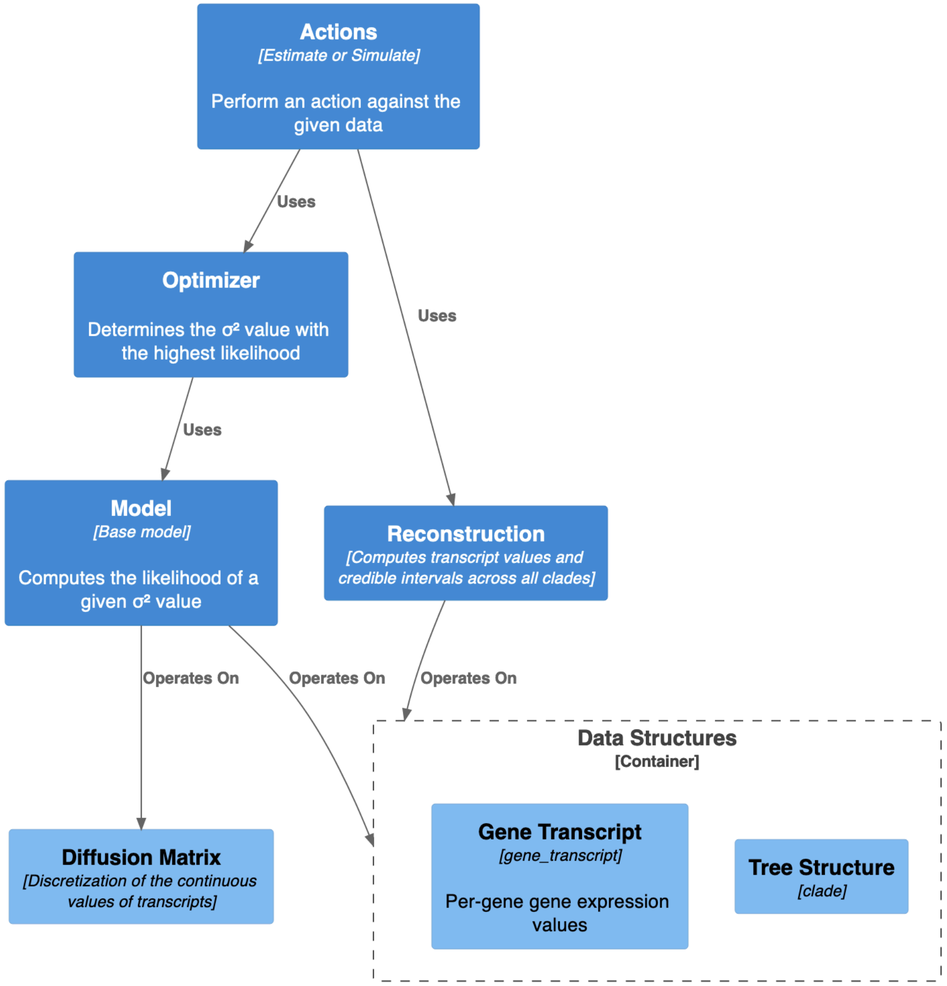
B)



Supplementary Figure 4



Supplementary Figure 5



Supplementary Table 1

Set of σ^2 values	Replicate	Simulated σ^2	Inferred σ^2	% error
0.2, 1.0	1	0.2	0.1926803	-3.66%
		1.0	1.0024905	0.249%
	2	0.2	0.1943218	-2.84%
		1.0	1.0054018	0.54%
	3	0.2	0.2031335	1.57%
		1.0	0.9577028	-4.23%
	4	0.2	0.1967767	-1.61%
		1.0	0.9996745	-0.0325%
	5	0.2	0.1924732	-3.76%
		1.0	1.0075440	0.754%
	6	0.2	0.2004951	0.248%
		1.0	1.0242412	2.42%
	7	0.2	0.1944284	-2.79%
		1.0	0.9733314	-2.67%
	8	0.2	0.2024781	1.24%
		1.0	0.9884935	-1.15%
	9	0.2	0.2018622	0.931%
		1.0	1.0382544	3.83%
	10	0.2	0.1956765	-2.16%
		1.0	1.0065474	0.655%
1.0, 3.0	1	1.0	0.9451556	-5.48%
		3.0	3.0287977	0.96%
	2	1.0	0.9769290	-2.31%
		3.0	3.1000229	3.33%
	3	1.0	0.9861053	-1.39%
		3.0	2.9452423	-1.83%
	4	1.0	0.9724728	-2.75%
		3.0	2.9203197	-2.66%
	5	1.0	0.9226307	-7.74%
		3.0	2.9930805	-0.231%
	6	1.0	0.9939074	-0.609%
		3.0	2.8856497	-3.81%
	7	1.0	0.9871113	-1.29%
		3.0	2.9705234	-0.983%
	8	1.0	0.9684494	-3.16%
		3.0	2.9187219	-2.71%
	9	1.0	0.9667607	-3.32%
		3.0	3.1521085	5.07%
	10	1.0	1.0059515	0.595%
		3.0	3.0232185	0.774%
5.0, 10.0	1	5.0	4.7464719	-5.07%
		10.0	9.9814701	-0.185%
	2	5.0	4.8651485	-2.7%
		10.0	10.0907926	0.908%
	3	5.0	4.7598429	-4.8%
		10.0	9.8058231	-1.94%
	4	5.0	4.9265027	-1.47%
		10.0	10.0593675	0.594%
	5	5.0	4.9239001	-1.52%
		10.0	9.9818898	-0.181%
	6	5.0	4.8684289	-2.63%
		10.0	9.8020147	-1.98%
	7	5.0	4.8752468	-2.5%
		10.0	10.0378953	0.379%
	8	5.0	4.9042695	-1.91%
		10.0	9.8541603	-1.46%
	9	5.0	4.6019327	-7.96%
		10.0	9.9286600	-0.713%
	10	5.0	4.8589572	-2.82%
		10.0	9.8313460	-1.69%