Working Notes: Single-celled bottlenecks, germlines and the evolution of complex multi-cellularity

Contents

1 Data Collection

1.1 Search strategy

Searches were conducted broadly for literature focussed on reproductive mode and germline development across the tree of life. This included chapters reviews and chapters within textbooks.

As Fisher paper was used for estimates of individual complexity (which in turn used Bell paper as a foundation), narrower searches were conducted for each species/genus from Bell paper on Web of Knowledge and on Google Scholar.

```
(ALL = (reproduct* OR sex* OR asex* OR vegetat* OR fissi* OR clonal* OR regenerat* OR
rhizo* OR germ-line* OR germline* OR germ line* OR bud* OR fragment* OR parthenogen* OR
stolon*)) AND (ALL = TAXON)
```

We recorded whether sexual, parthenogenetic/clonal and agametic have been observed as binary values. We did not attempt to capture the relative frequency of different reproductive strategies, as these data do not exist for the majority of species.

Research into reproduction and development is heterogeneous across organisms, and this is necessarily reflected in the required searching effort for different groups: the reproductive biology and development of model organisms such as C. elegans, M. musculus, D. melanogaster, A. thaliana are well known, but in many other groups reproduction may never have been observed. We therefore conducted searches for each species, but if no literature discussing reproductive strategy was observed, then we conducted an additional search at the genus level. This assumes that genera will tend to be relatively similar in their reproductive strategies: this is not always the case. The planarian S. mediterranea, for example, has strictly sexual and strictly asexual strains within even the same species. However, we are focussed on patterns through longer spans of evolution than between individual genera.

Caveats:

- Sexual organisms contain more cells because they have gonads that asexual organisms lack...

- hard to fit algae with gametophyte/sporophyte stages: they have life cycles where some stages can fragment, where some stages can reproduce by spores, parthenogenesis or by sex. If an algae can stay in one loop and reproduce exclusively through parthenogenesis/fragmentation, then counts– similar to organisms that can fission, but don't always.

1.2 Production of phylogenetic tree

We used a phylogenetic tree to control for non-independence of species based on shared evolutionary history. The tree was constructed within R using the latest evolutionary classifications found on the Tree of Life, AlgaeBase.org, and the World Register of Marine species. The relationships among species were reconstructed by ordering the taxa from Kingdom through to species, and grouping according to these names.

As a comparison, we also constructed a tree using the 'R Tree of Life Project.' These two trees were largely congruent: some larger groups had switched places, but within these groups relationships were predominantly the same. As the Rtol tree dropped X data points from the tree, we used the tree based on the taxa names. Multichotomies within the tree were randomly resolved, before branch lengths were generated as described by [@grafen1990]. Branches smaller than 10^{-25} were deleted, and the dichotomies here collapsed to multichotomies. Figure () shows a cophylogeny based on each tree.

2 Statistical Analyses

2.1 MCMCglmm parameters

All analyses were conducted in R [@R-base] using the package MCMCglmm [@MCMCglmm], while documents were produced using [RMarkdown]. All data and code are accessible at github.

Model parameters were optimised using the first model described in our results, for which we ran a total of 38 MCMCglmm chains of varying lengths (500000 - 10000000 iterations), with varying warm-ups (100000 - 1000000, and with thinning of either 100 or 1000 fold, see Figure S1. All subsequent models were then fit using the combination of these parameters where the autocorrelation of successive sampled mean and variance were minimal: 8×10^6 iterations, a warm-up of 10^6 iterations and thinning by a factor of 100. In all fitted models, the autocorrelation was well below the suggested tolerable maximum of 0.1 [@hadfield?]. For

each model, 6 chains were run which were visually inspected for chain convergence. Convergence was also supported by the Gelman-Rubin [@Gelman-Rubin] convergence diagnostic, which approximated 1 (<1.05) in all cases–these are reported in the summary of each model below.

The four models described in section @ref(Without Phylogeny) are phylogenetically naive, and treat each species as independent data points. The default priors used for fixed effects, and residual variance prior of V = 1 and nu = 0.002.

The differences between each level for the fixed effects were calculated at each MCMC iteration to produce a posterior distribution for the difference. Levels are considered statistically significant if the 95% credible interval of this difference distribution did not overlap with 0, and if the proportion of MCMC iterations that were greater or less than 0 was less than 0.05.

The entire analysis, including the parameter optimisation step and creating all output documents, runs in approximately 3hrs 15 mins using a 2020 MacBook Pro running 4 chains in parallel.

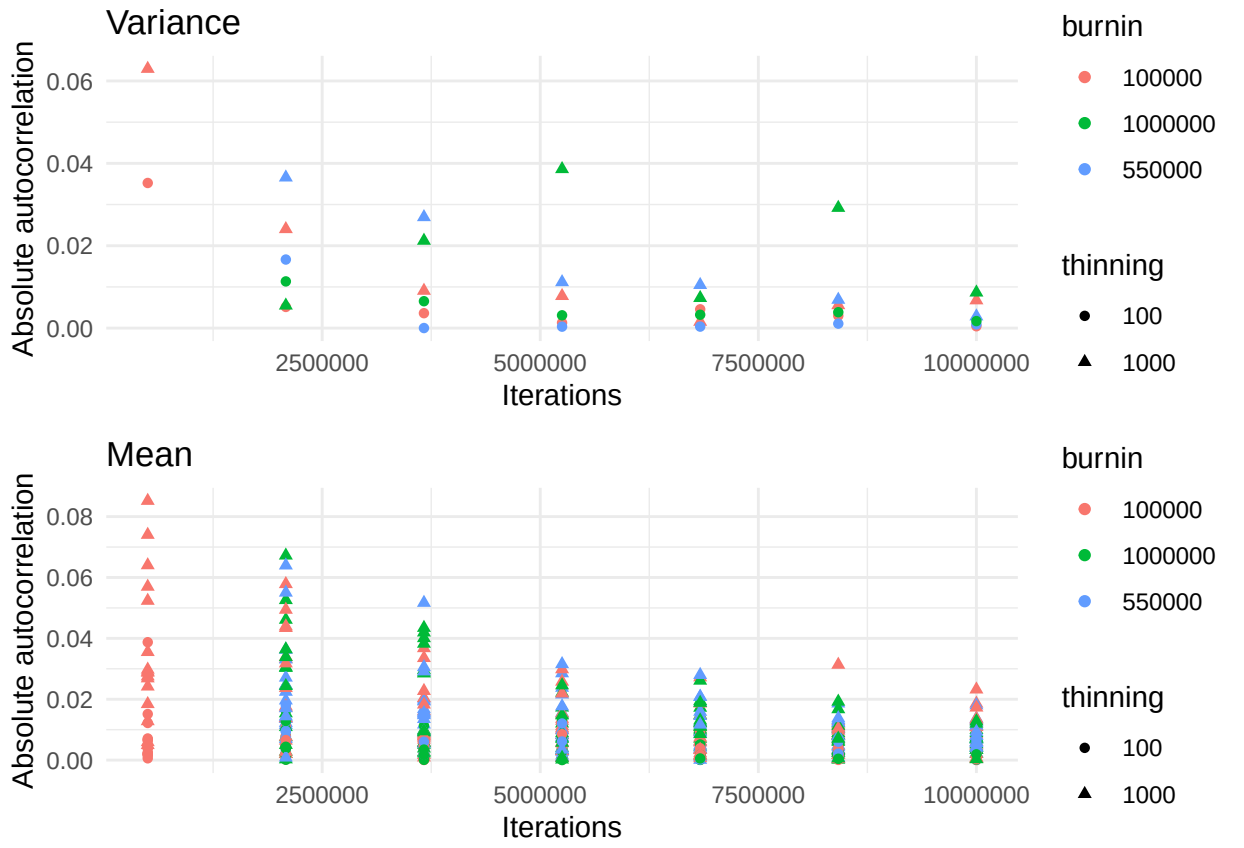

Figure 1: Autocorrelation of successively sampled mean and variance values from posterior distribution

2.2 Without Phylogeny

2.2.1 Model 1: Fission vs Cell Number

Do organisms that reproduce by fission have more cells? Fissiparous organisms appear to be larger (makes sense, trees, fungi, algae, etc)

priors: p1=list(R = list(V = 1, nu=0.002)) #sets prior for residual variance, the defaults are used as priors for fixed effects (see MCMCglmm course notes)

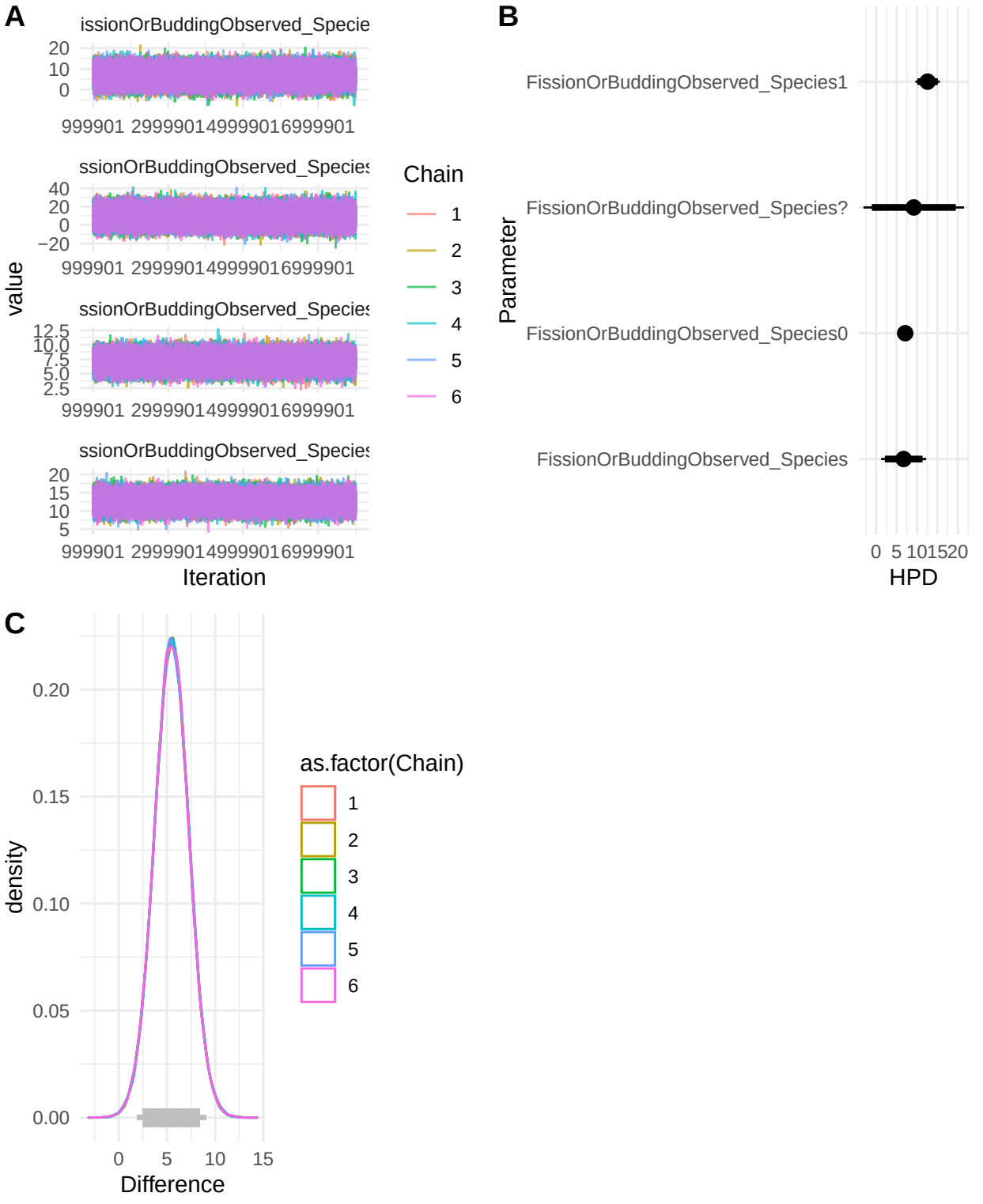

Figure 2: **Model 1: Cell Numbers vs Fission** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively. *C* Density plot of estimated differences, bar represents 90% and 95% credible intervals.

2.2.2 Model 2: Fission vs Cell Types

priors p1=list(R = list(V = 1, nu=0.002))

Do organisms that reproduce by fission have more cell types? HCI overlaps with zero, so doesn't seem likely.

2.2.3 Model 3: Germline vs Cell Numbers

Should we subset to only those organisms that have sterile cells for the germline models?

Do organisms with early segregating germline have more cells? HCI just about overlaps with 0, so maayyyybe, but not clear.

priors p1=list(R = list(V = 1, nu=0.002))

2.2.4 Model 4: Germline vs Cell Types

Do organisms that segregate germline early have more cell types? Again, just about overlaps with 0, so not clear.

p1=list(R = list(V = 1, nu=0.002))

2.3 Phylogenetically Informed Models

Models below here use inverse covariance matrix describing the relationships among species to control for phylogeny.

2.3.1 Model 5: Fission vs Cell Number

Again, just about overlaps with 0, so not clear.

p2=list(R = list(V = 1, nu=0.002), G = list(G1=list(V=1, nu=0.002)))

2.3.2 Model 6: Fission vs Cell Types

Overlaps with 0, no difference

p2=list(R = list(V = 1, nu=0.002), G = list(G1=list(V=1, nu=0.002)))

2.3.3 Model 7: Germline vs Cell Number

Things with a germline might be smaller: the 95% CI is *just* below 0 (-0.34)

p2=list(R = list(V = 1, nu=0.002), G = list(G1=list(V=1, nu=0.002)))

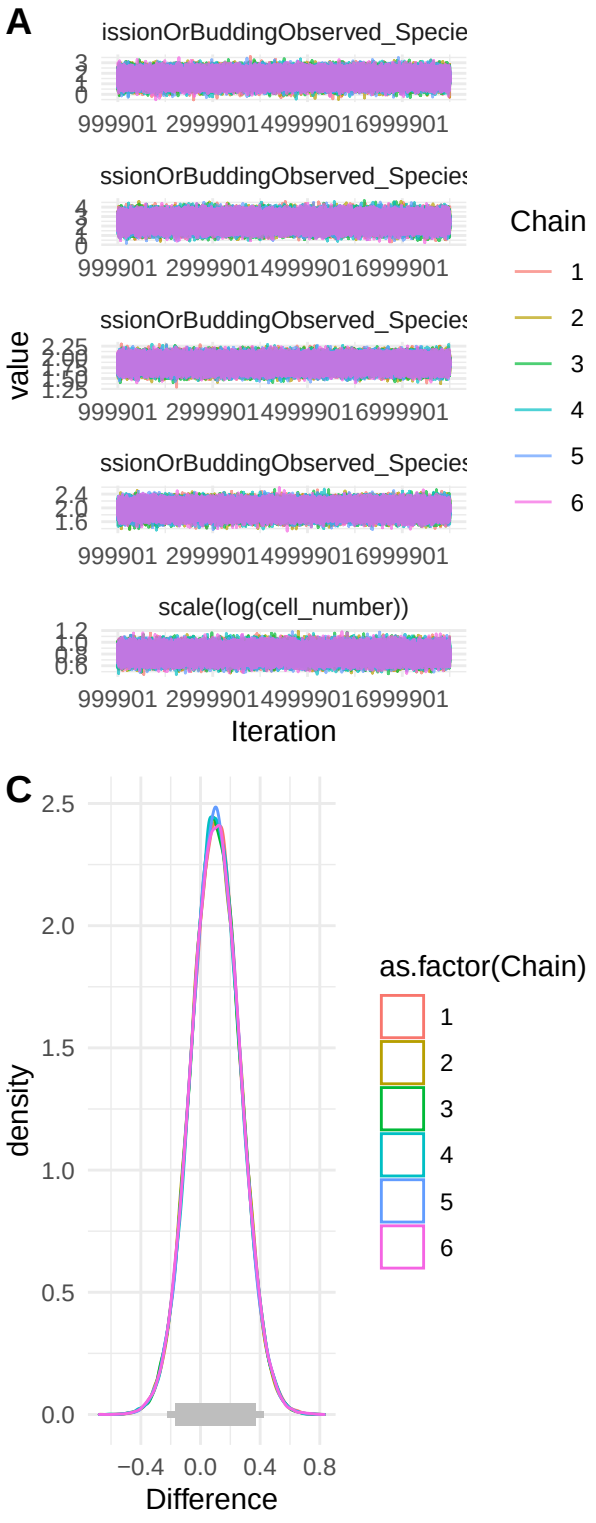

Figure 3: **Model 2: Cell Types vs Fission** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively. *C* Density plot of estimated differences, bar represents 90% and 95% credible intervals.

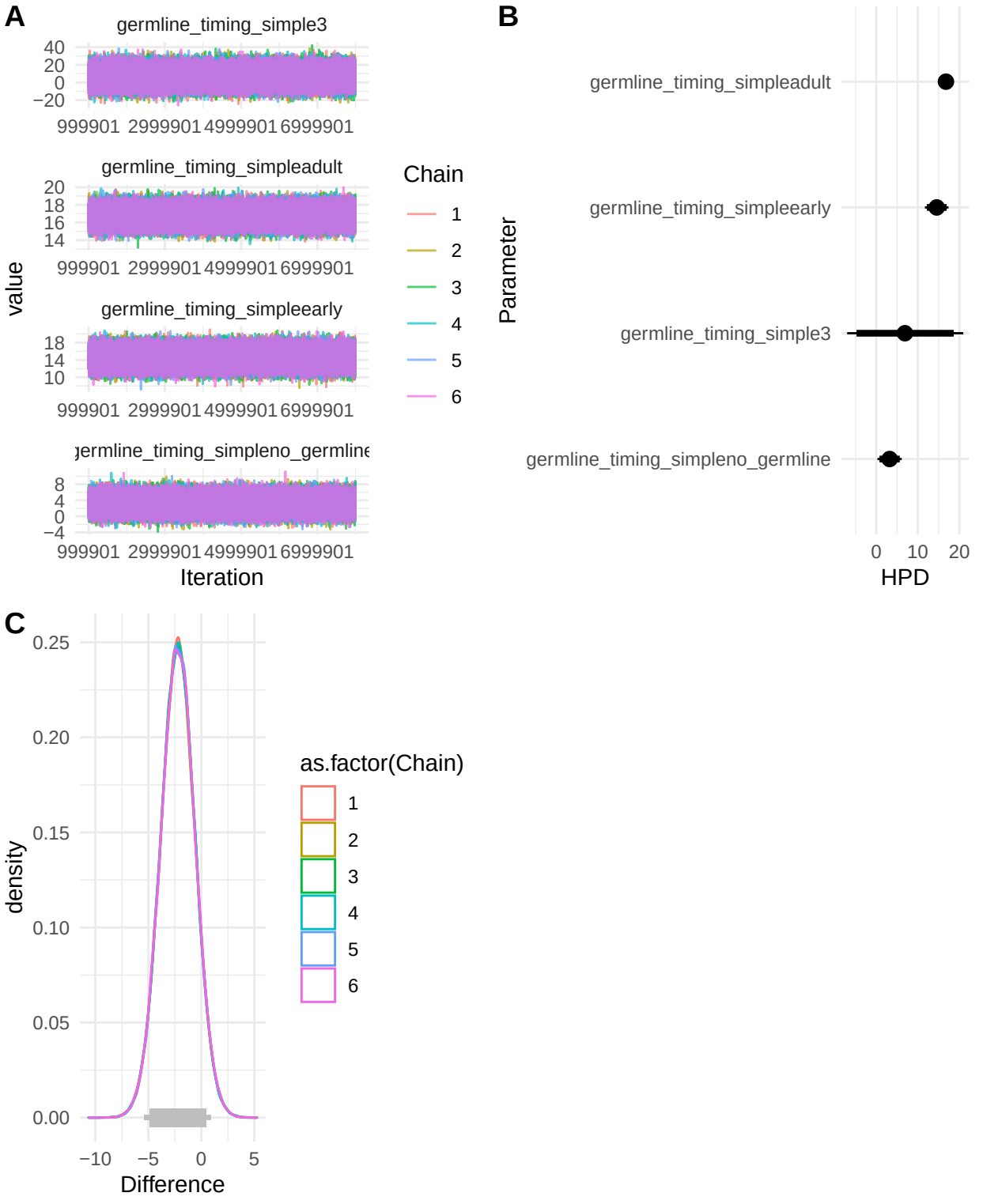

Figure 4: **Model 3: Cell number vs Germline** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively. *C* Density plot of estimated differences, bar represents 90% and 95% credible intervals.

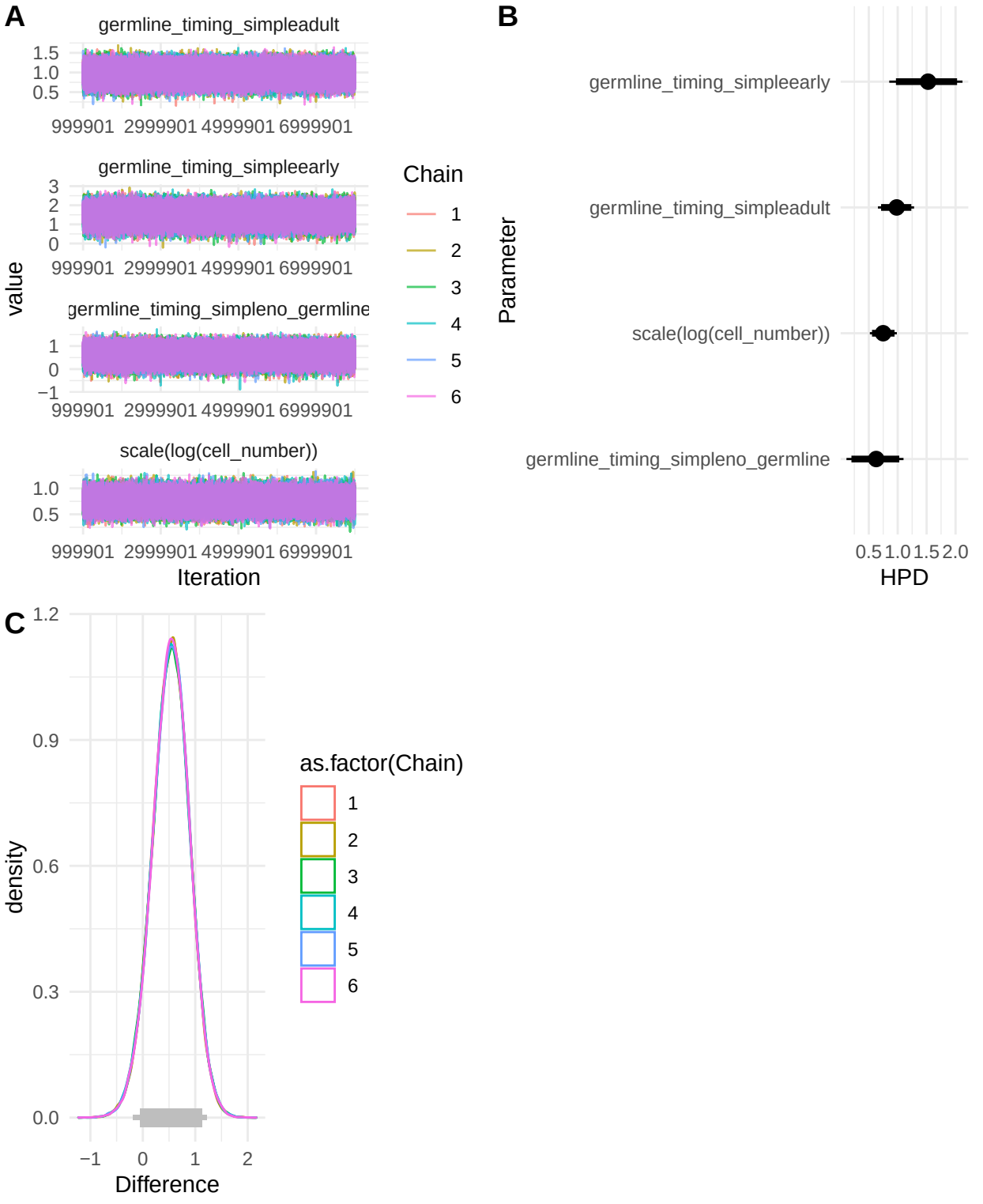

Figure 5: **Model 4: Cell Types vs Germline** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively. *C* Density plot of estimated differences, bar represents 90% and 95% credible intervals.

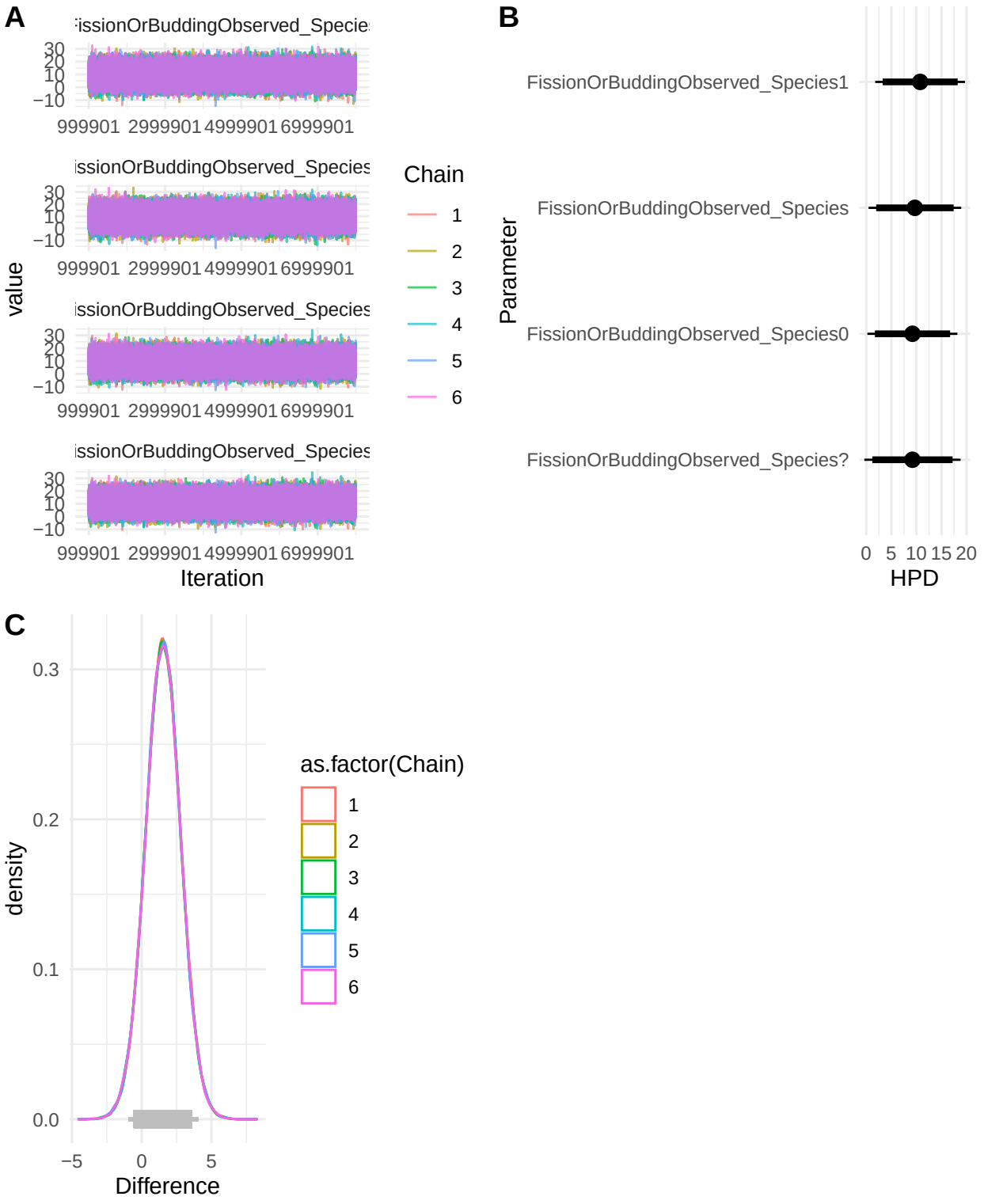

Figure 6: **Model 5: Cell Number vs Fission with phylogeny** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively. *C* Density plot of estimated differences, bar represents 90% and 95% credible intervals.

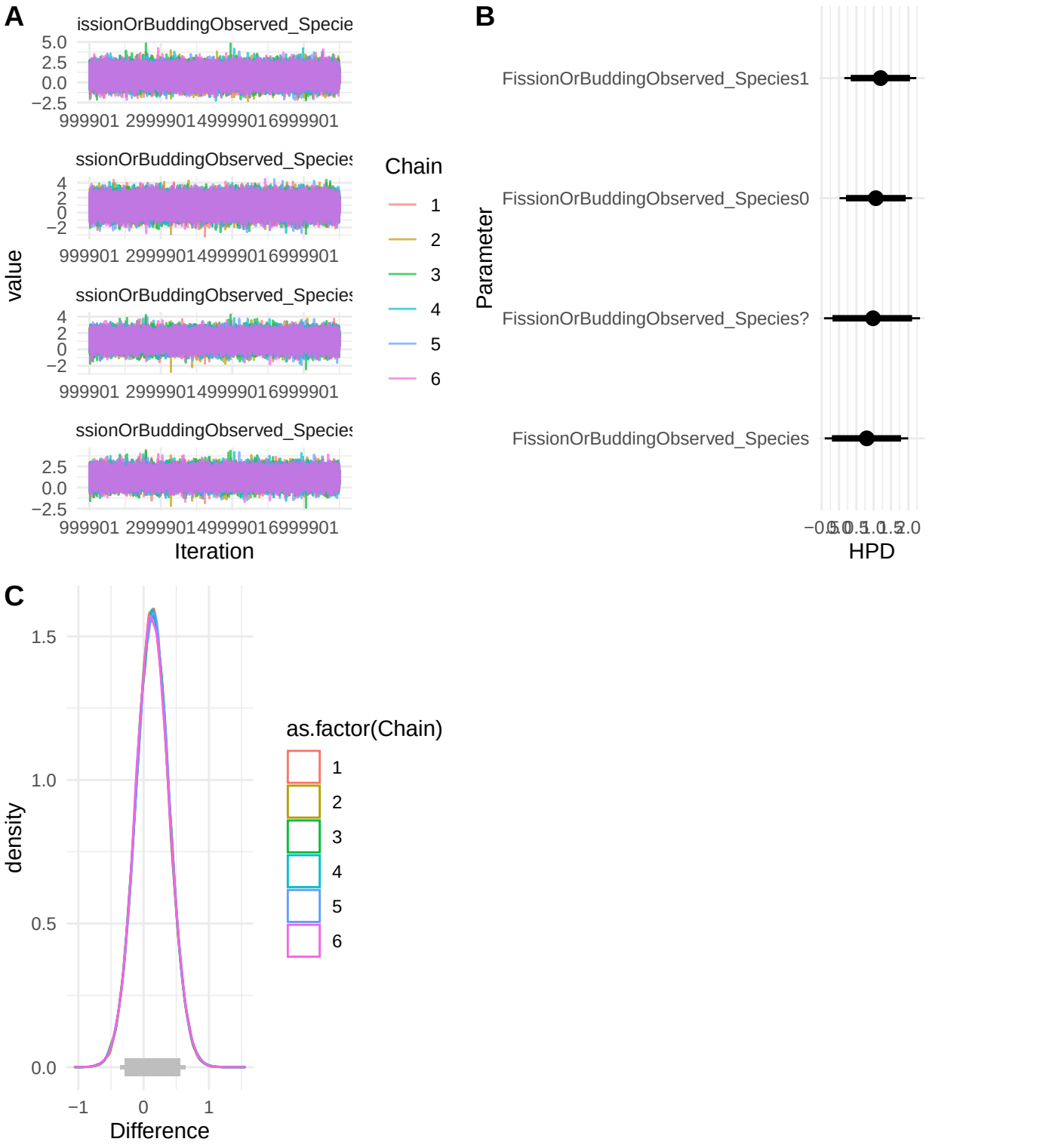

Figure 7: **Model 6: Cell Number vs Fission with phylogeny** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively. *C* Density plot of estimated differences, bar represents 90% and 95% credible intervals.

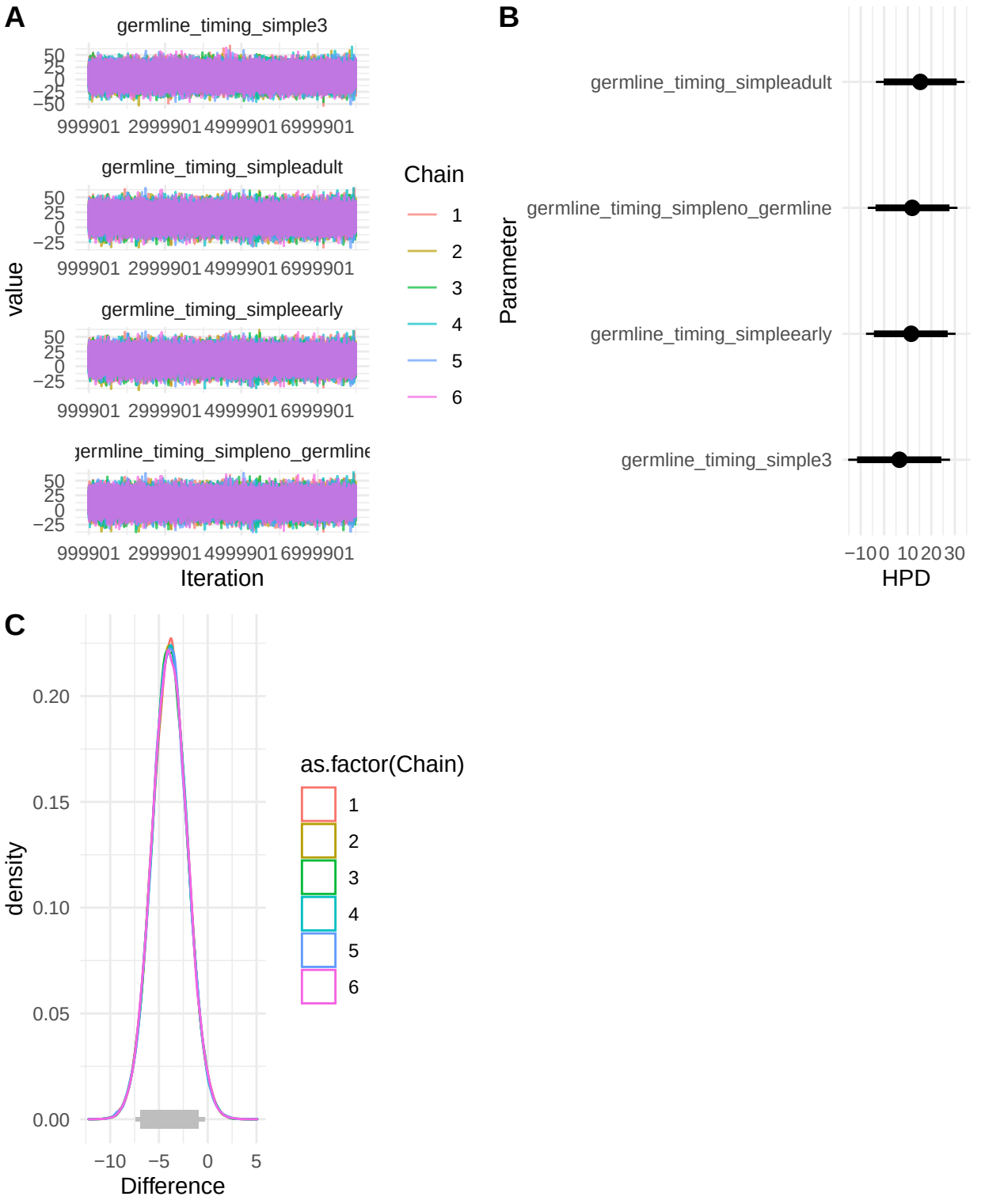

Figure 8: **Model 7: Cell Number vs Fission with phylogeny** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively. *C* Density plot of estimated differences, bar represents 90% and 95% credible intervals.

2.3.4 Model 8: Germline vs Cell Types

```
##
##  Iterations = 1000001:7999901
##  Thinning interval  = 100
##  Sample size  = 70000
##
##  DIC: 226.7579
##
##  G-structure:  ~species
##
##         post.mean l-95% CI u-95% CI eff.samp
## species    0.7402   0.2043    1.418    34743
##
##  R-structure:  ~units
##
##       post.mean  l-95% CI u-95% CI eff.samp
## units    0.0123 0.0001529  0.04578    50977
##
##  Location effects: cell_types ~ germline_timing_simple - 1 + scale(log(cell_number))
##
##                                  post.mean l-95% CI u-95% CI eff.samp    pMCMC
## germline_timing_simpleadult          1.1211   0.2479   2.0623    55252  0.02006
## germline_timing_simpleearly          1.7797   0.7297   2.8027    50297  0.00271
## germline_timing_simpleno_germline    0.7392  -0.3472   1.7870    30346  0.16497
## scale(log(cell_number))              0.5993   0.3087   0.8953    35619 8.57e-05
##
## germline_timing_simpleadult       *
## germline_timing_simpleearly       **
## germline_timing_simpleno_germline
## scale(log(cell_number))           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Seems like they may be smaller, but that they have more cell types per cell– HCI is just above 0 (0.0379847).

p2=list(R = list(V = 1, nu=0.002), G = list(G1=list(V=1, nu=0.002)))

Is pMCMC just the number of simulated cases where difference is <0? In which case:

2.4 Open Questions

Phylogenetic correlation between germline and fission– multivariate model? How to test this?

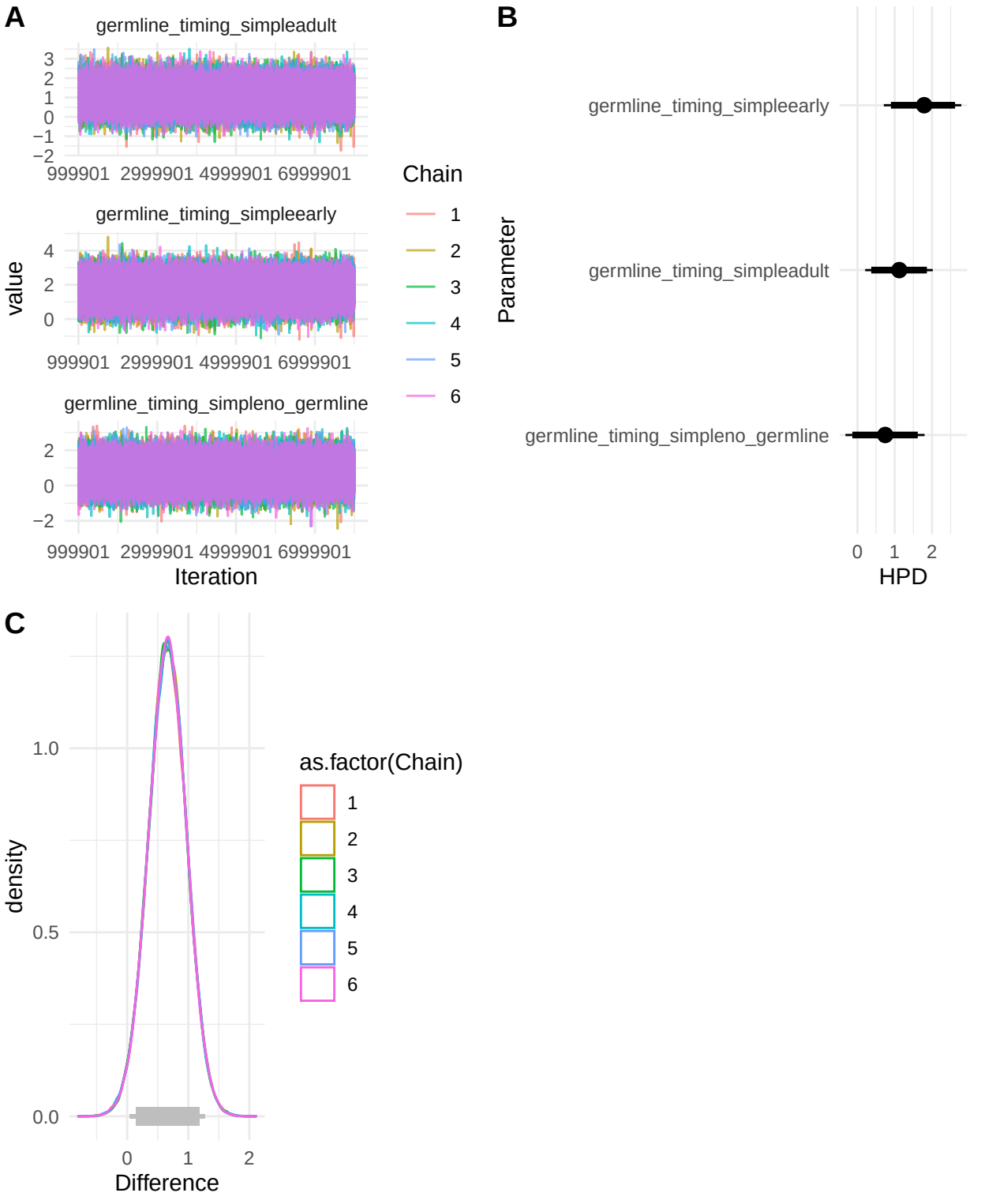

Figure 9: **Model 8: Cell Number vs Fission with phylogeny** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively. *C* Density plot of estimated differences, bar represents 90% and 95% credible intervals.

3 References