

¹ *Working Notes:* Single-celled bottlenecks, germlines and the
² evolution of complex multi-cellularity

³

⁴ **Contents**

⁵ 1 Data Collection	¹
⁶ 1.1 Search strategy	¹
⁷ 1.2 Production of phylogenetic tree	²
⁸ 2 Statistical Analyses	²
⁹ 2.1 MCMCglmm parameters	²
¹⁰ 2.2 Without Phylogeny	³
¹¹ 2.2.1 Model 1: Fission vs Cell Number	³
¹² 2.2.2 Model 2: Fission vs Cell Types	⁵
¹³ 2.2.3 Model 3: Germline vs Cell Numbers	⁵
¹⁴ 2.2.4 Model 4: Germline vs Cell Types	⁵
¹⁵ 2.3 Phylogenetically Informed Models	⁵
¹⁶ 2.3.1 Model 5: Fission vs Cell Number	⁵
¹⁷ 2.3.2 Model 6: Fission vs Cell Types	⁵
¹⁸ 2.3.3 Model 7: Germline vs Cell Number	⁵
¹⁹ 2.3.4 Model 8: Germline vs Cell Types	⁵
²⁰ 2.4 Open Questions	⁵
²¹ 3 References	13

²² **1 Data Collection**

²³ **1.1 Search strategy**

²⁴ Searches were conducted broadly for literature focussed on reproductive mode and germline development
²⁵ across the tree of life. This included chapters reviews and chapters within textbooks.

²⁶ As Fisher paper was used for estimates of individual complexity (which in turn used Bell paper as a foundation),
²⁷ narrower searches were conducted for each species/genus from Bell paper on Web of Knowledge and
²⁸ on Google Scholar.

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

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

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

44 Caveats:

- 45 • Sexual organisms contain more cells because they have gonads that asexual organisms lack...
- 46
- 47 • hard to fit algae with gametophyte/sporophyte stages: they have life cycles where some stages can
48 fragment, where some stages can reproduce by spores, parthenogenesis or by sex. If an algae can stay
49 in one loop and reproduce exclusively through parthenogenesis/fragmentation, then counts– similar to
50 organisms that can fission, but don't always.

51 1.2 Production of phylogenetic tree

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

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

62 2 Statistical Analyses

63 2.1 MCMCglmm parameters

64 All analyses were conducted in R [@R-base] using the package MCMCglmm [@MCMCglmm], while docu-
65 ments were produced using [RMarkdown]. All data and code are accessible at [github](#).

66 Model parameters were optimised using the first model described in our results, for which we ran a total of
67 38 MCMCglmm chains of varying lengths (500000 - 10000000 iterations), with varying warm-ups (100000
68 - 1000000, and with thinning of either 100 or 1000 fold, see Figure S1. All subsequent models were then
69 fit using the combination of these parameters where the autocorrelation of successive sampled mean and
70 variance were minimal: 8×10^6 iterations, a warm-up of 10^6 iterations and thinning by a factor of 100. In all
71 fitted models, the autocorrelation was well below the suggested tolerable maximum of 0.1 [@hadfield?]. For

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

75 The four models described in section @ref(Without Phylogeny) are phylogenetically naive, and treat each
76 species as independent data points. The default priors used for fixed effects, and residual variance prior of
77 $V = 1$ and $\nu = 0.002$.

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

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

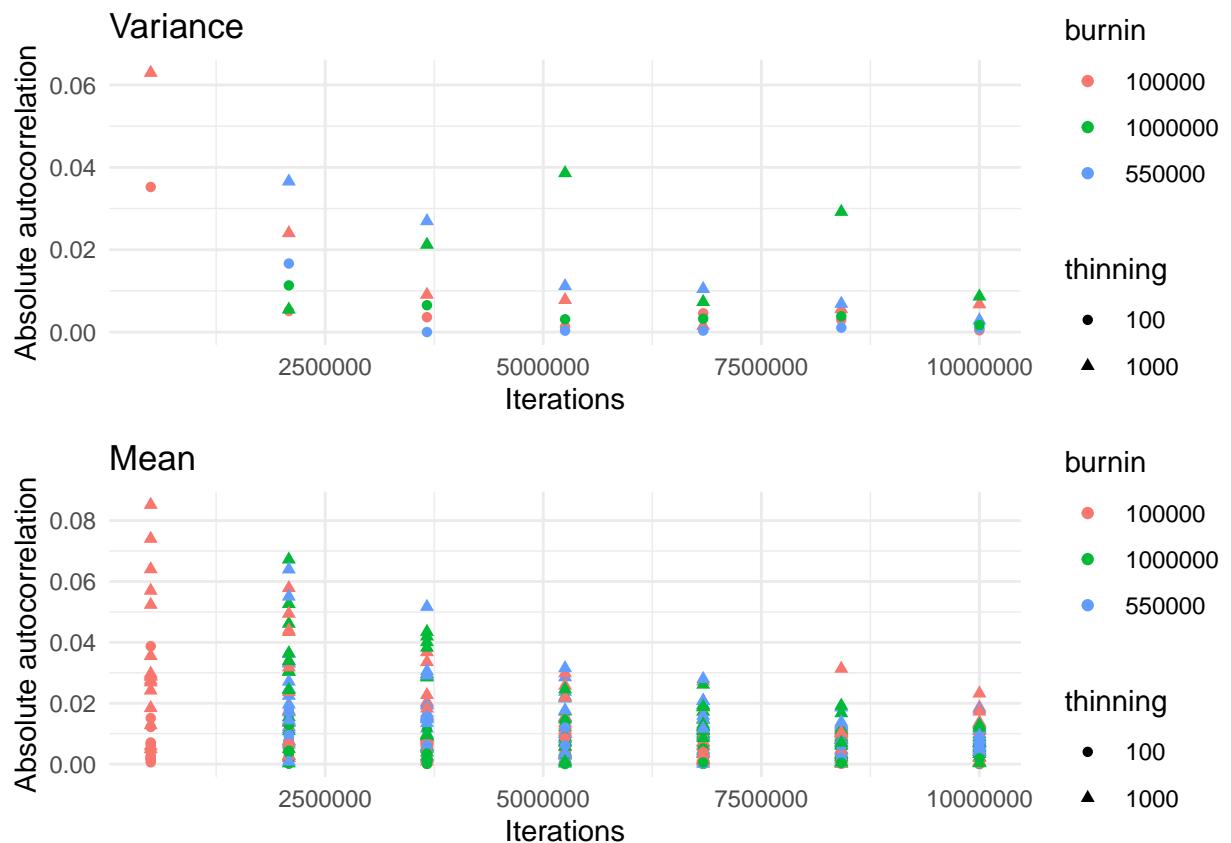


Figure 1: Autocorrelation of successively sampled mean and variance values from posterior distribution

84 2.2 Without Phylogeny

85 2.2.1 Model 1: Fission vs Cell Number

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

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

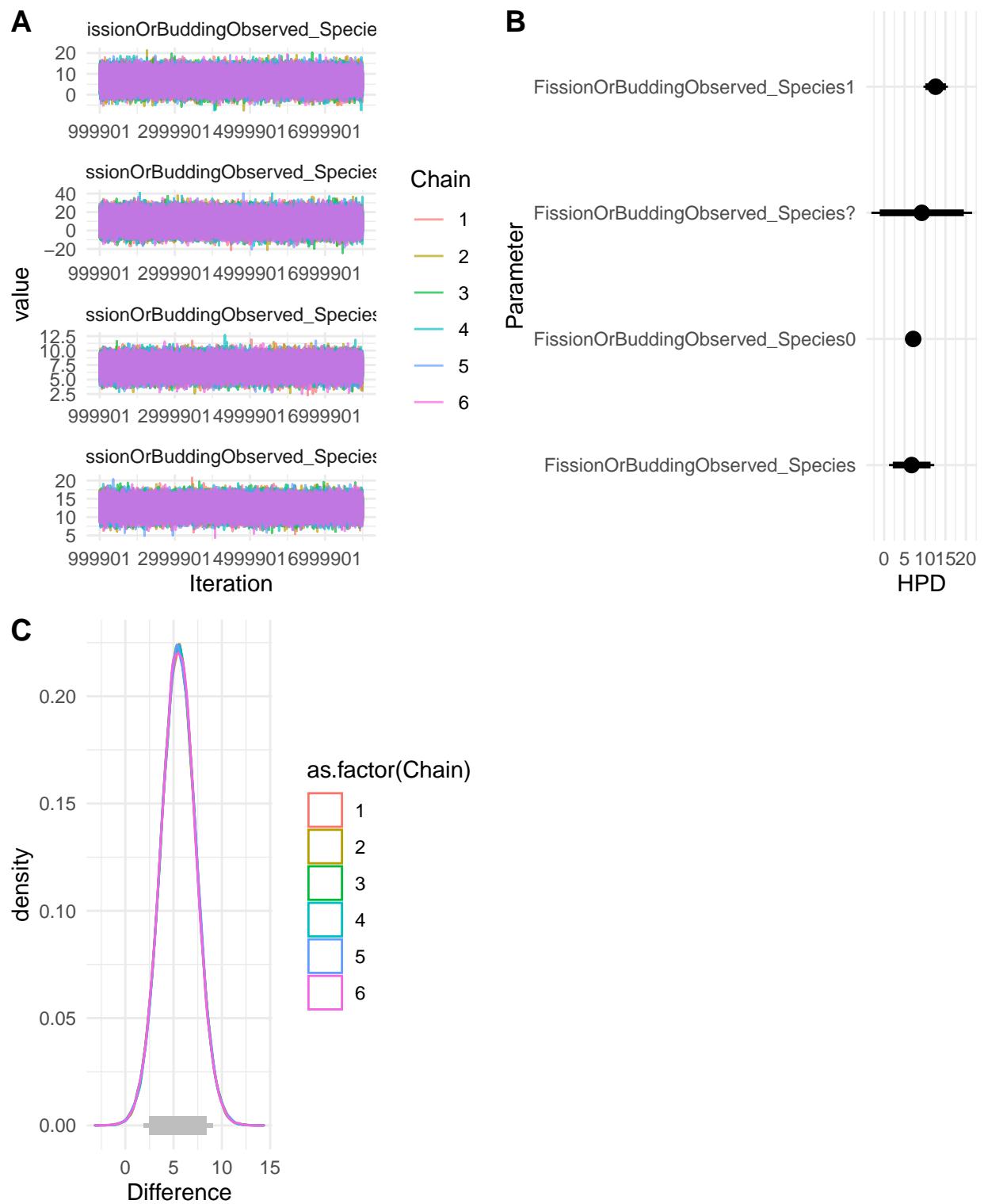


Figure 2: **Model 1: Cell Numbers vs Fission** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively.

90 **2.2.2 Model 2: Fission vs Cell Types**

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

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

94 **2.2.3 Model 3: Germline vs Cell Numbers**

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

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

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

99 **2.2.4 Model 4: Germline vs Cell Types**

100 **Do organisms that segregate germline early have more cell types?** Again, just about overlaps with
101 0, so not clear.

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

103 **2.3 Phylogenetically Informed Models**

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

106 **2.3.1 Model 5: Fission vs Cell Number**

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

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

109 **2.3.2 Model 6: Fission vs Cell Types**

110 Overlaps with 0, no difference

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

112 **2.3.3 Model 7: Germline vs Cell Number**

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

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

115 **2.3.4 Model 8: Germline vs Cell Types**

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

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

118 **2.4 Open Questions**

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

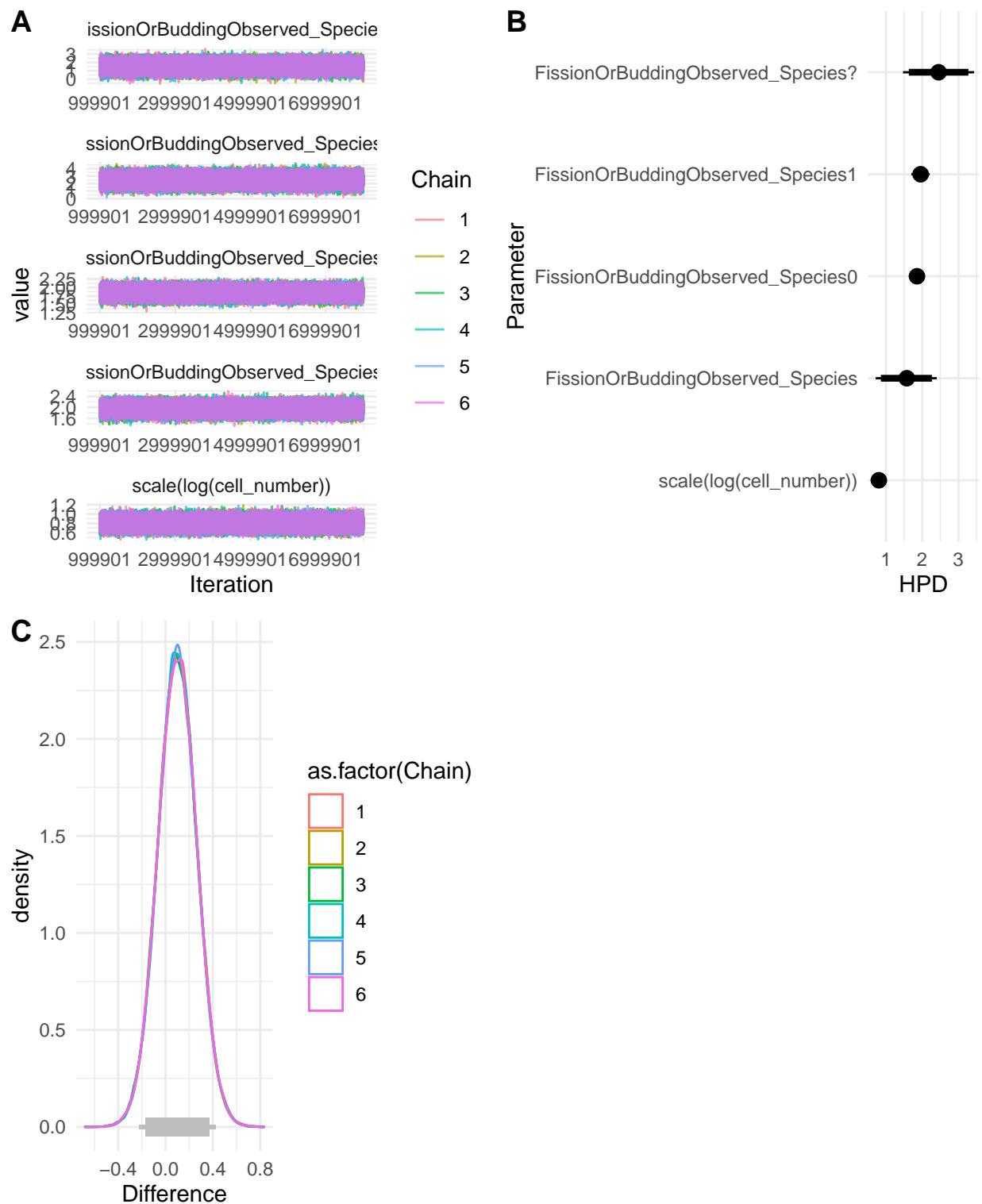


Figure 3: **Model 2: Cell Types vs Fission** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively.

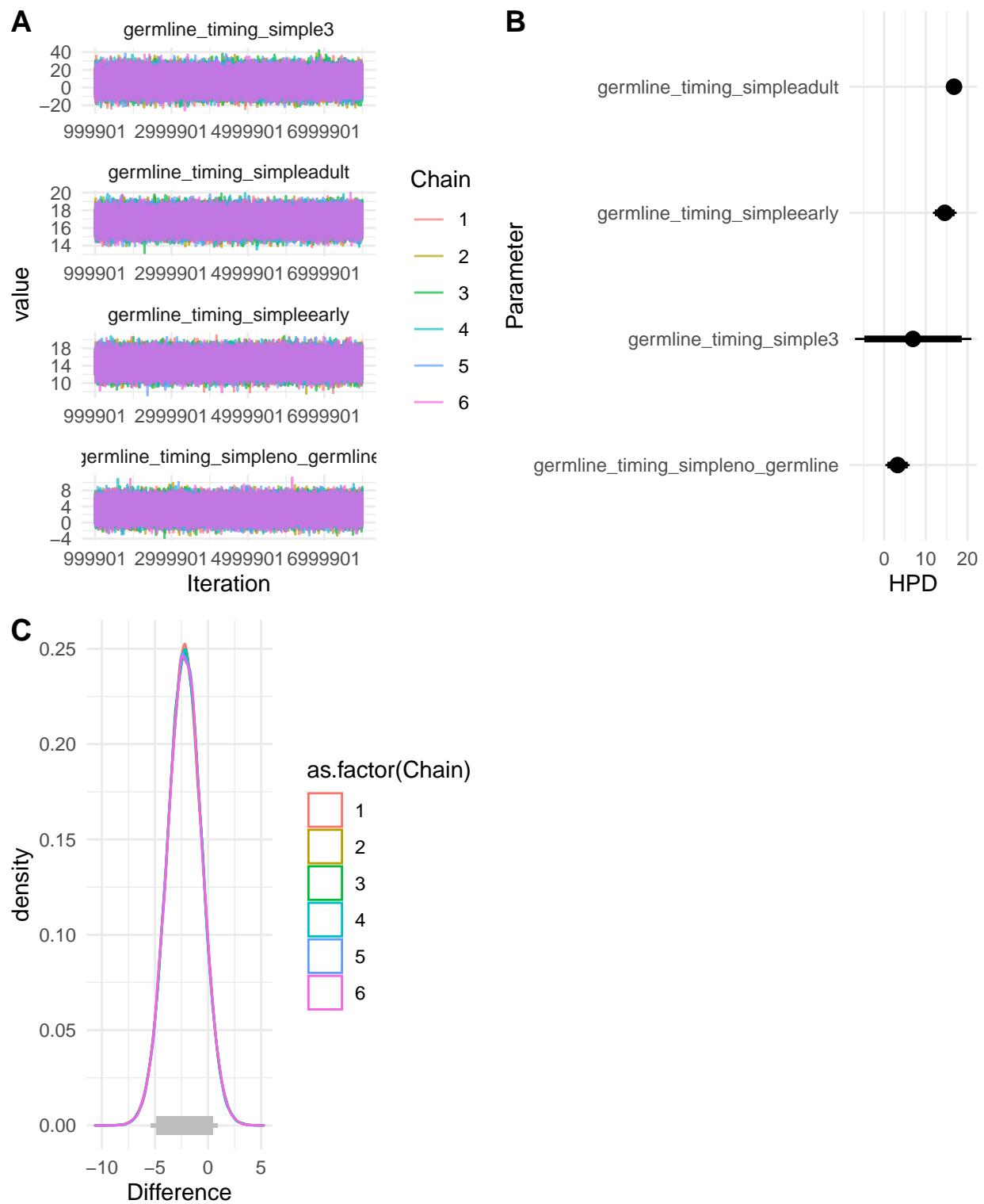


Figure 4: **Model 3: Cell number vs Germline** A Traceplots for the estimated means, B Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively.

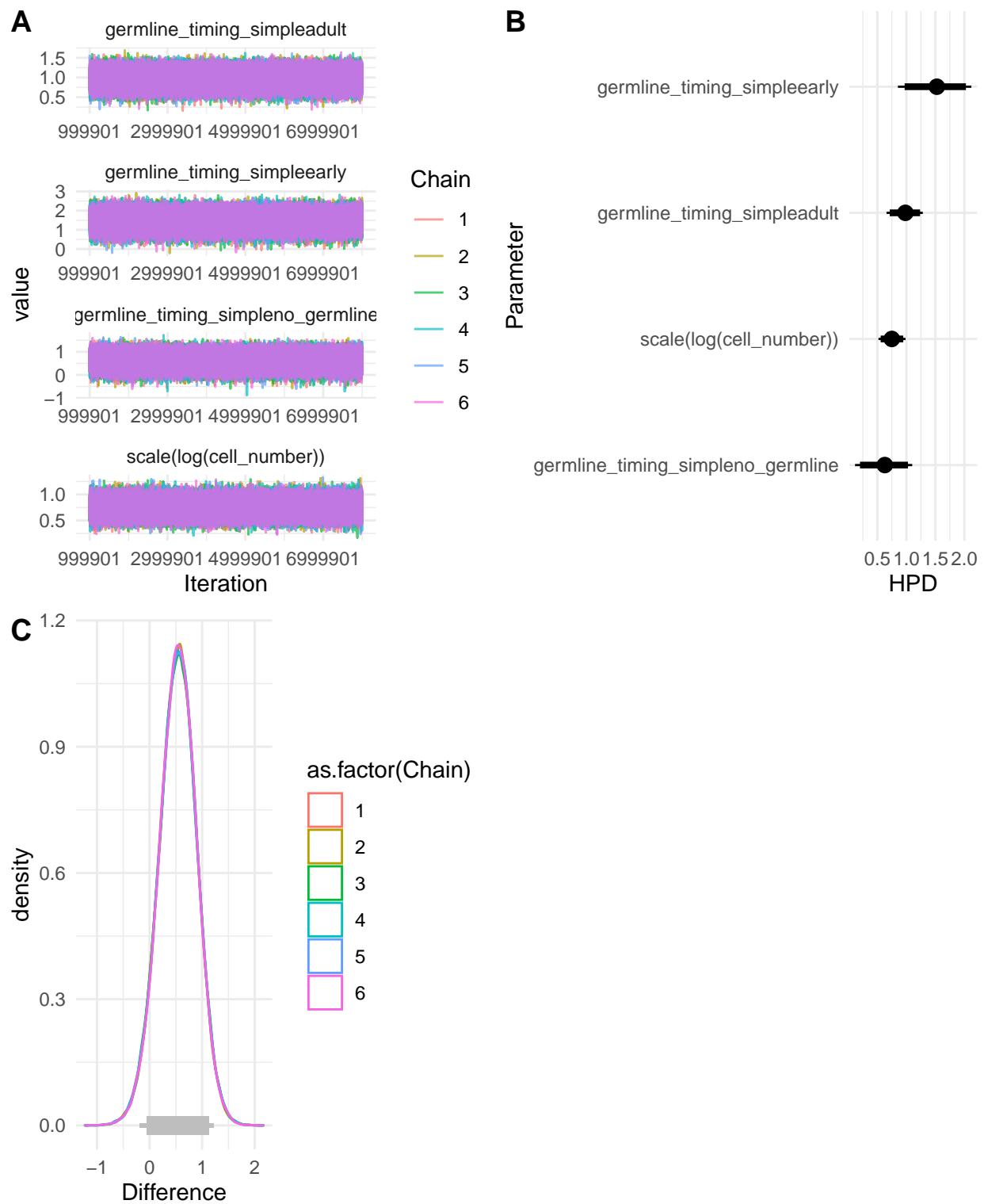


Figure 5: **Model 4: Cell Types vs Germline A** Traceplots for the estimated means, **B** Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively.

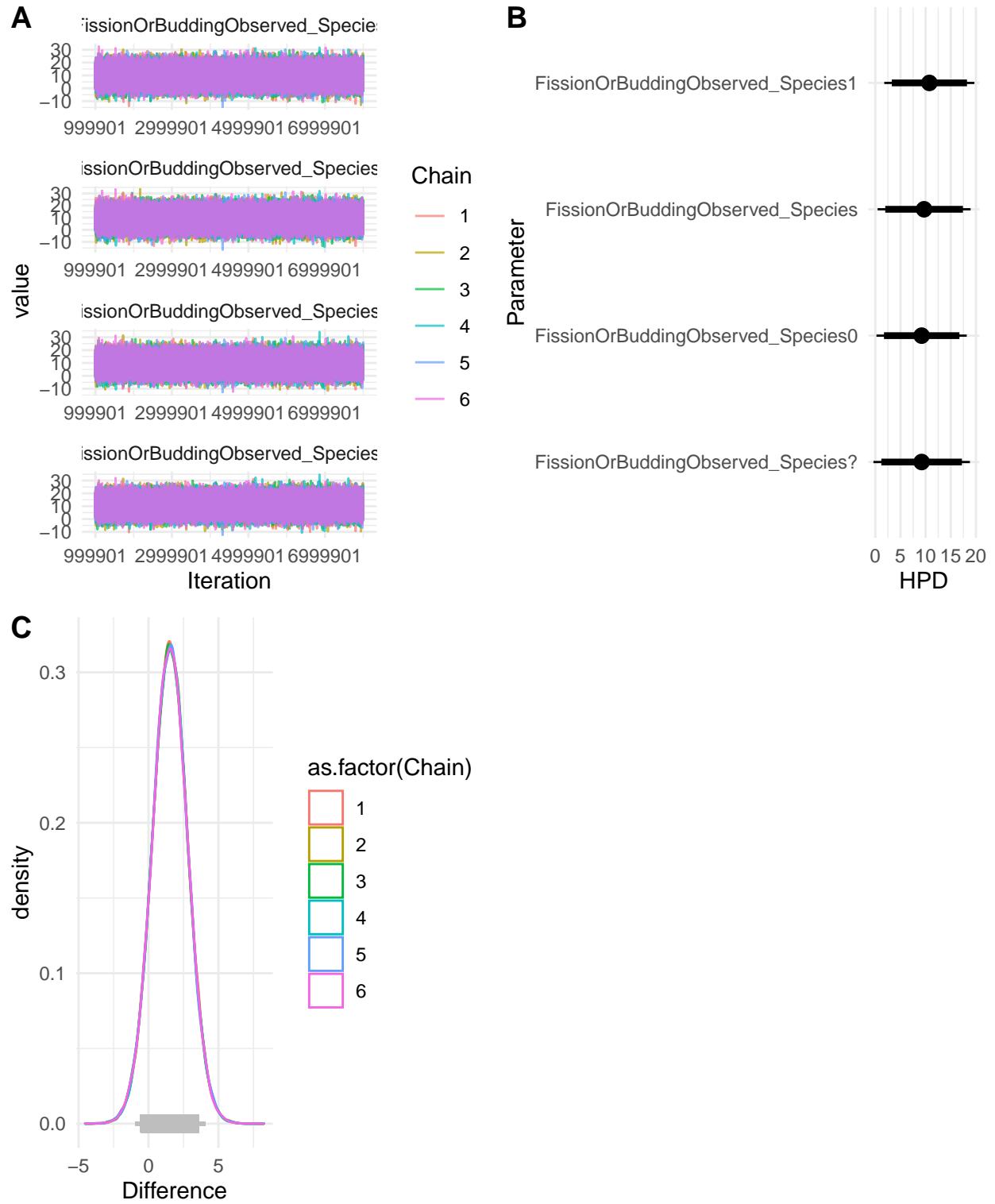


Figure 6: **Model 5: Cell Number vs Fission with germline** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively.

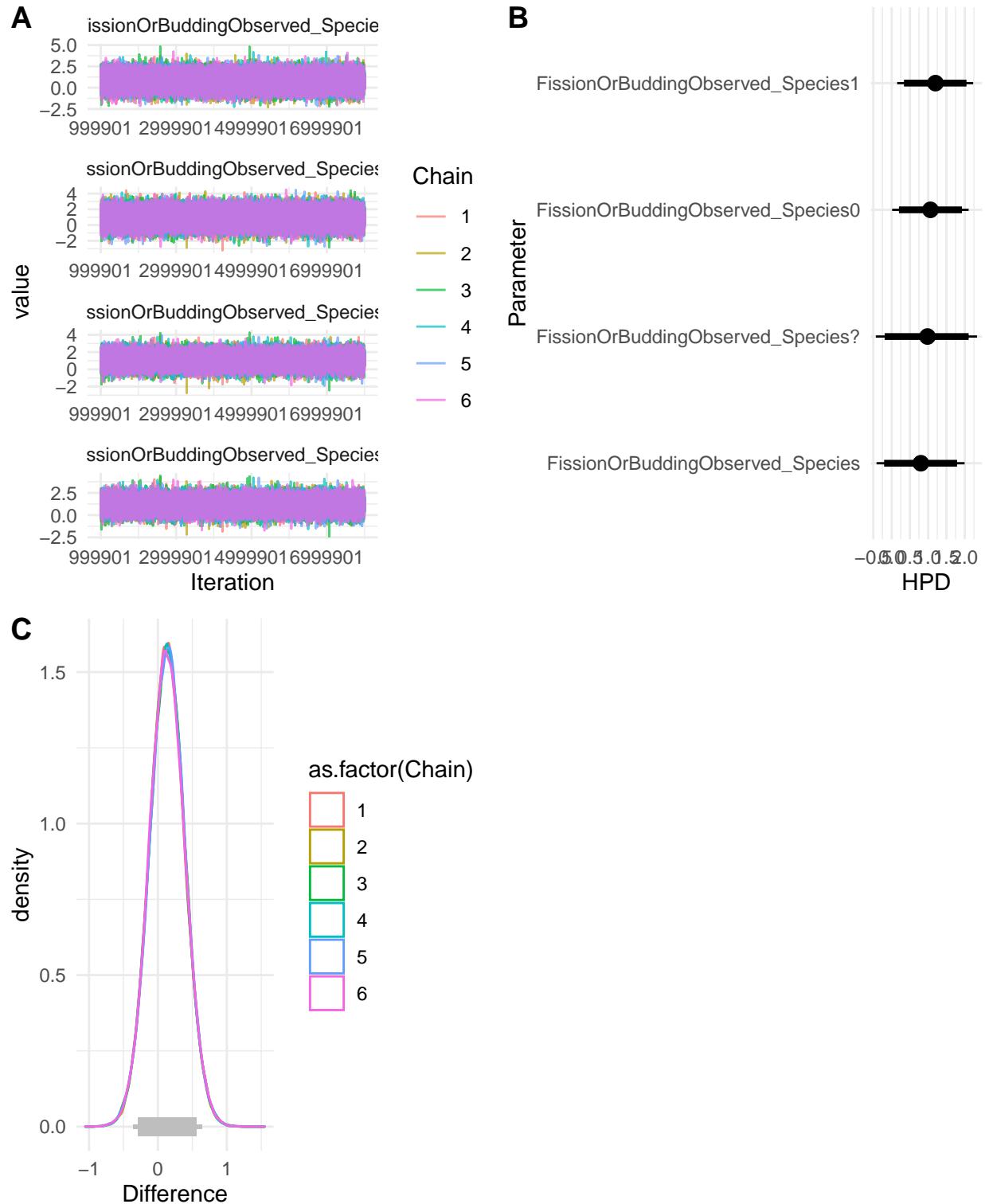


Figure 7: **Model 6: Cell Number vs Fission with germline** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively.

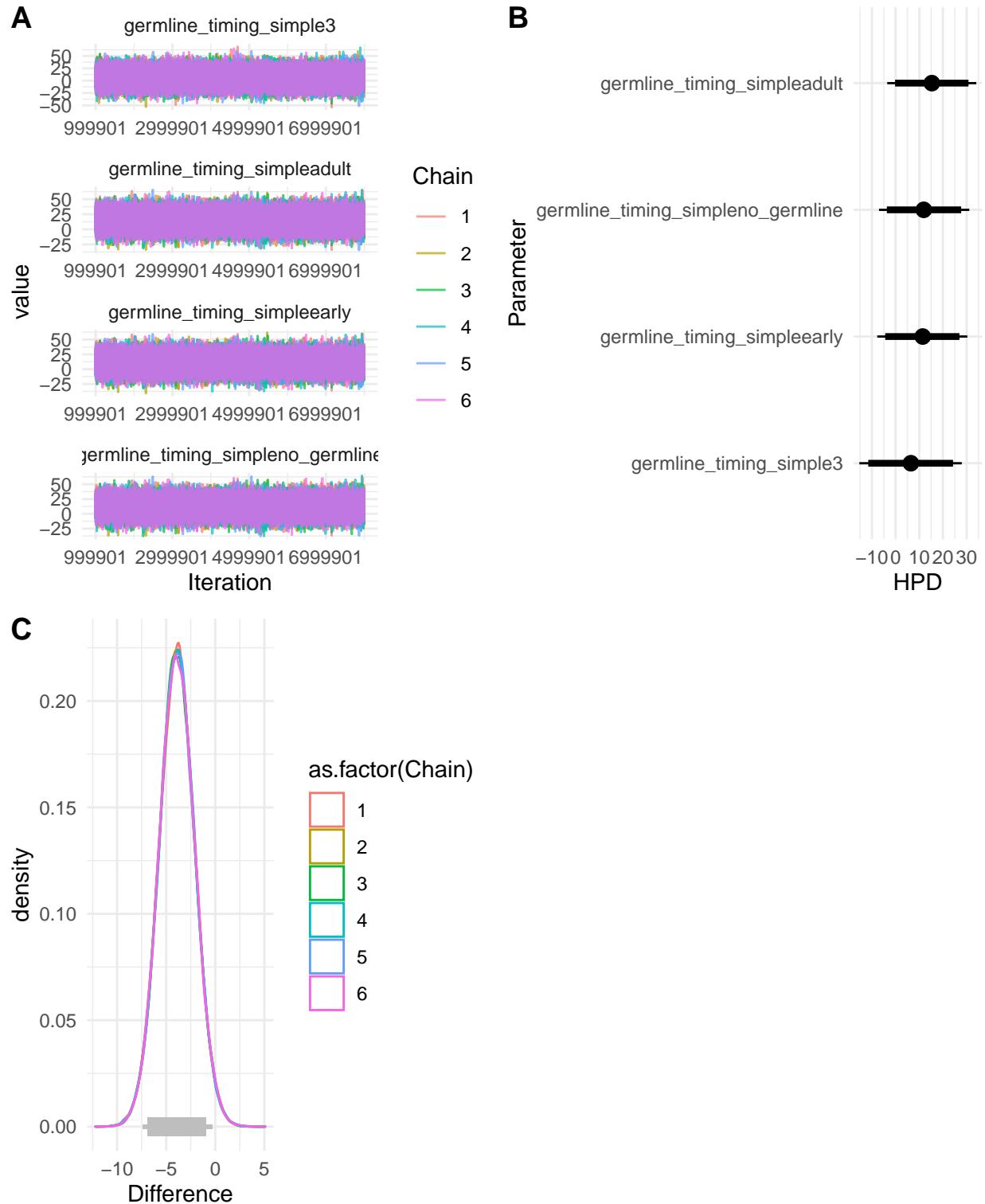


Figure 8: **Model 7: Cell Number vs Fission with germline** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively.

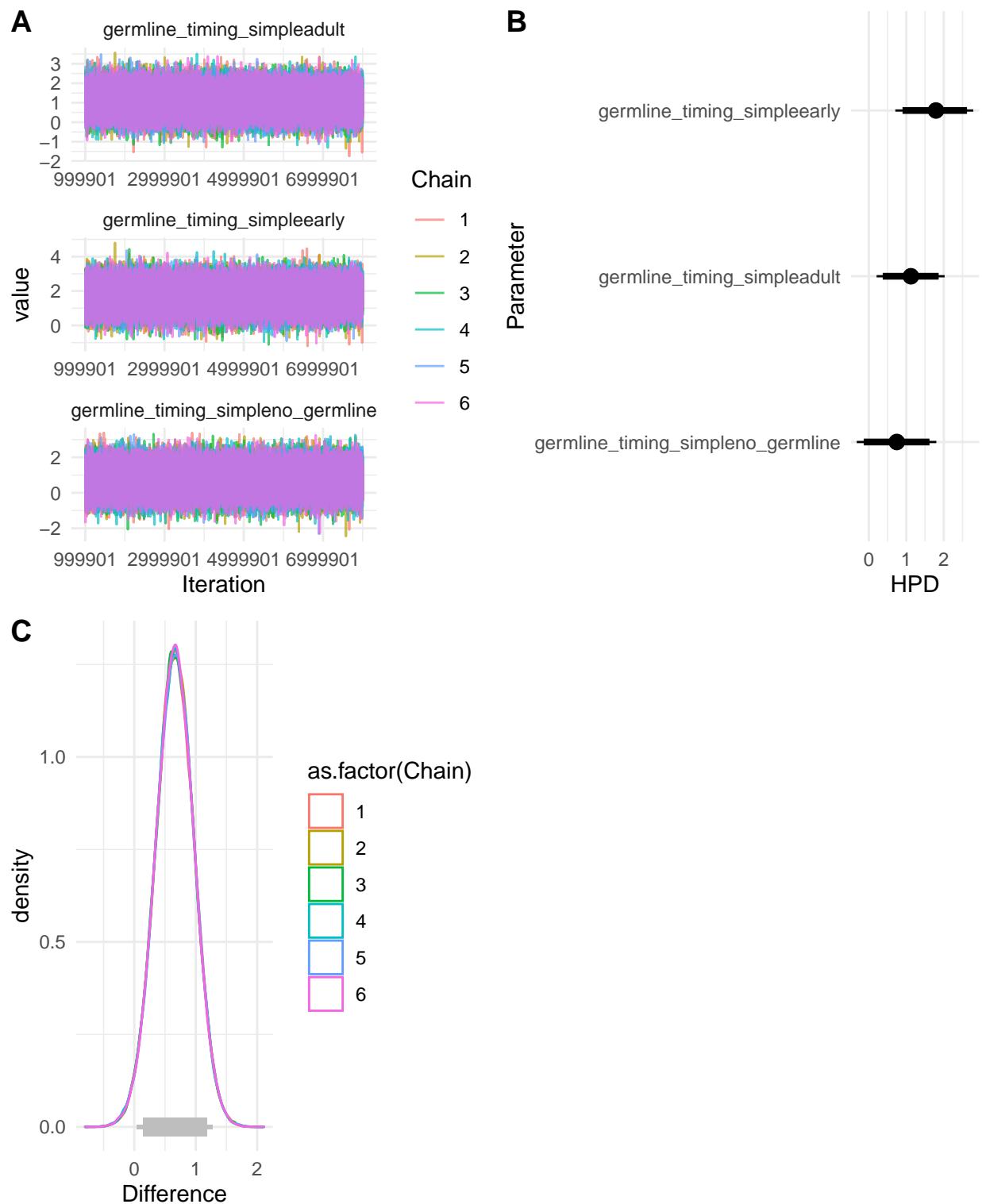


Figure 9: **Model 8: Cell Number vs Fission with germline** *A* Traceplots for the estimated means, *B* Estimates for means from posterior distribution, dots represent median, thick and thin lines indicate 90% and 95% of highest posterior density regions, respectively.

₁₂₀ **3 References**