Tests on priors for STACEY 1.1

Graham Jones

February 5, 2016

1 Introduction

This technical report describes some tests carried out on the BEAST2 package STACEY. Version 2.3.2 of BEAST and 1.1.1 of STACEY were used. The main purpose is to test that the MCMC operators correctly sample from the prior (that is, the posterior when there is no sequence data). The operators are StaceyNodeReheight, NodesNudge, FocusedScaler, and CoordinatedPruneRegraft, ThreeBranchAdjuster, abbreviated as H, N, F, R, A. The tests are very similar to those in the supplementary information of Jones (2015), or see `http://www.indriid.com/2015/2015-01-07-tests-on-priors.pdf`.

There are two sets of tests. One uses a fixed number (8) of species and samples from the prior on the species tree. The other has an unknown number of species (between 1 and 8). Although there is no sequence data, the assumptions about the number of species constitute some 'meta-data'. In both sets of tests, there is one gene tree with no data, that is with a sequence "?" at each tip.

2 Species tree prior with fixed delimitation

2.1 Scenario and settings

The XML files were generated using R scripts, namely `STACEY_XML` available at `https://github.com/Graham853` (commit a204e402fd09eb1fd3f0053c9e7c71359b71c8e1, 5 Feb 2016) and the script at the end of this note.

The population scaling factor σ is fixed to 0.01. A Yule model for the species tree prior was assumed. To achieve this, the relative death rate μ is set to 0 and the growth rate λ was 100. The combination $\lambda = 100$, $\sigma = 0.01$ means that a pair of sequences coalesce in a time equal to the mean branch length. This results in a moderate amount of incomplete lineage sorting.

for the fixed delimitation case, the collapse weight w is fixed at zero, and the collapse height is ignored.

The burnin was set to 10% of the run. There were about 9000 samples.

2.2 Results

Figure 1 shows how the operators sample from the possible topologies. Figure 2 shows how the operators sample the speciation heights. The theoretical values come from Gernhard (2008).

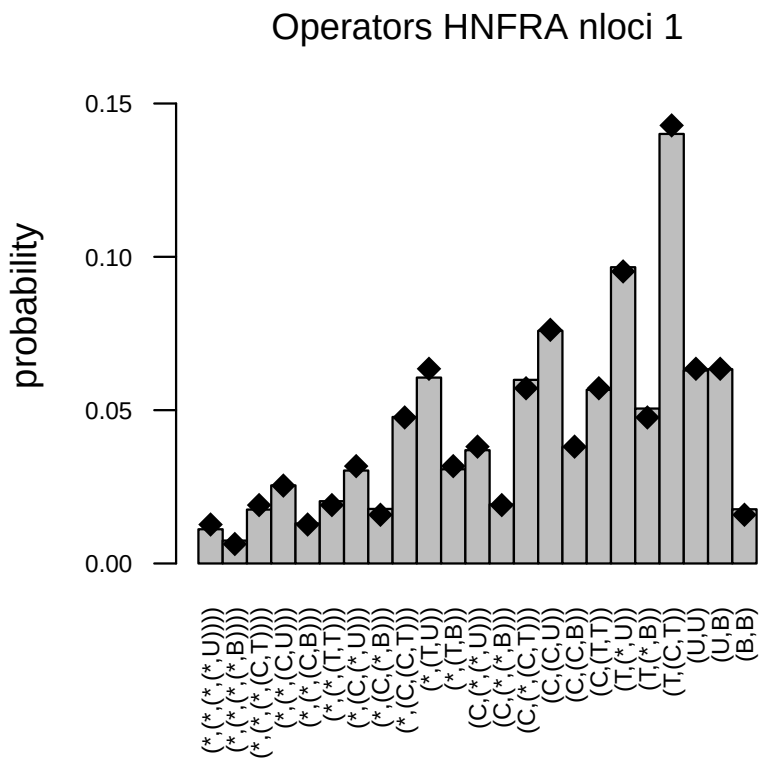

Figure 1: Estimated posterior probabilities for each of the 23 unlabeled rooted topologies with eight tips are shown in grey bars. The theoretical values are shown as diamonds. The x-axis annotations show the topologies in the following format. A tip is shown by *****. A cherry **(*,*)** is denoted as **C**, a 3-tip tree **(*,(*,*))** as **T**, an unbalanced 4-tip tree as **U** and a balanced 4-tip tree as **B**. Otherwise, the Newick format is used. For example **(T,(*,U))** is short for **((*,(*,*)),(*,(*,(*,(*,*)))))**.

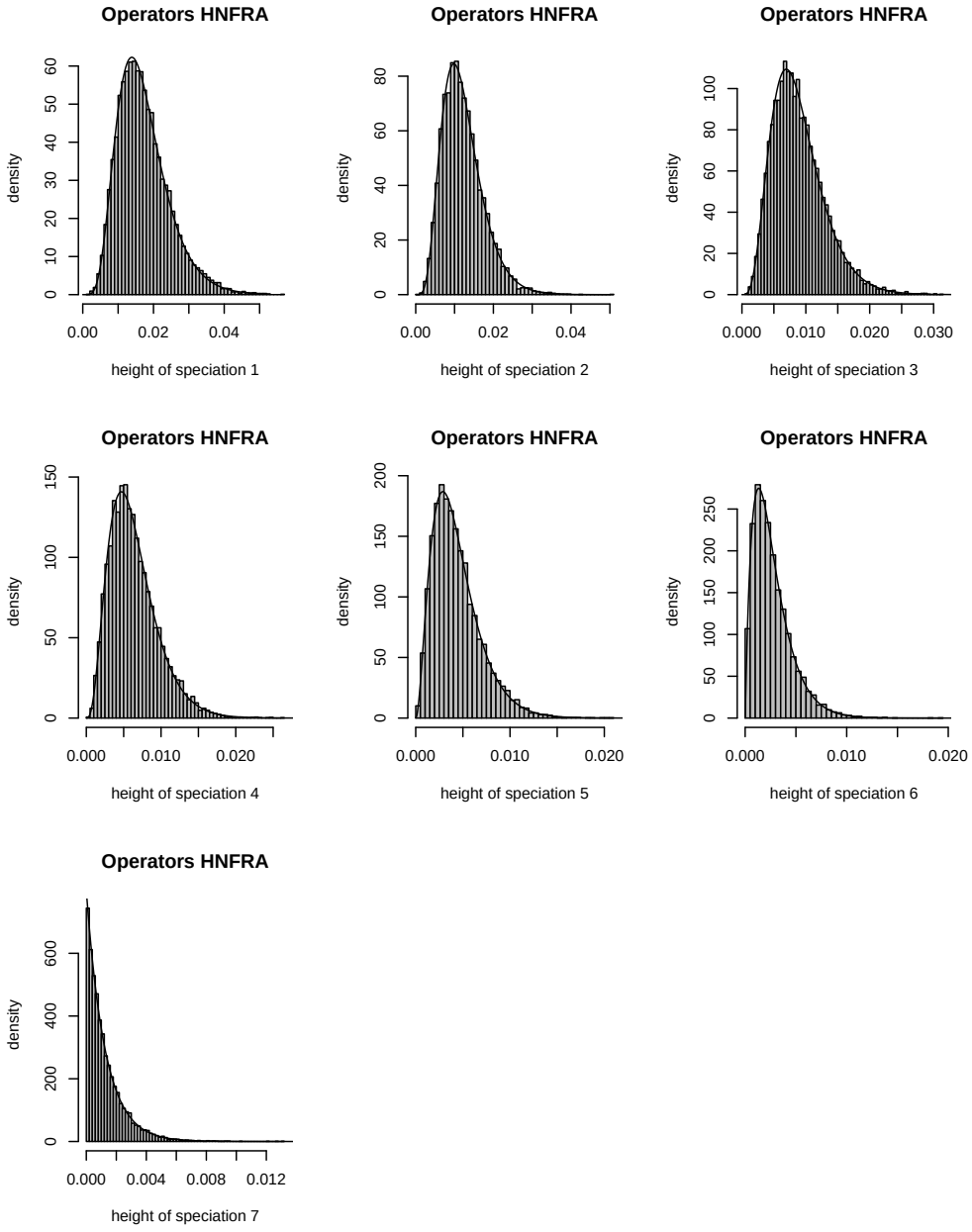

Figure 2: Histograms of the MCMC samples for the successive speciation heights. The theoretical densities are shown as curves.

3 Species delimitation

3.1 Scenario and settings

Here w is given a uniform prior on [0,1], resulting in a flat prior on the number of species between 1 and 8. The collapse height ϵ was set to 3e-5. This is about 1/300 of the expected branch lengths of 0.01. other settings were the same as for the fixed case.

3.2 Results

Figure 3 shows estimated and theoretical values for the 22 partitions of the number 8. Each partition of 8 represents one or more clusterings of 8 objects. There are a total of 4140 clusterings of 8 objects, and these can be grouped into 22 sets corresponding to the partitions of 8. For example suppose the 8 objects are a, b, c, d, e, f, g, h. One clustering is $\{\{a, b, c\}, \{d\}, \{e, f, g, h\}\}$, which corresponds to the partition 4+3+1 of 8. There are 280 clusterings with the shape 4+3+1, and whenever one of these are visited during the MCMC, it counts towards the posterior probability of this partition of 8.

Figures 4 show how the operators sample some parameters and other values. The likelihood is extremely close to zero as expected. The posteriors for w and the number of clusters are close to uniform, as they should be.

In the BirthDeathCollapseModel graph, which shows the contribution this makes to the posterior, distinct peaks can be seen which appear to represent the different numbers of clusters (although some peaks are merged).

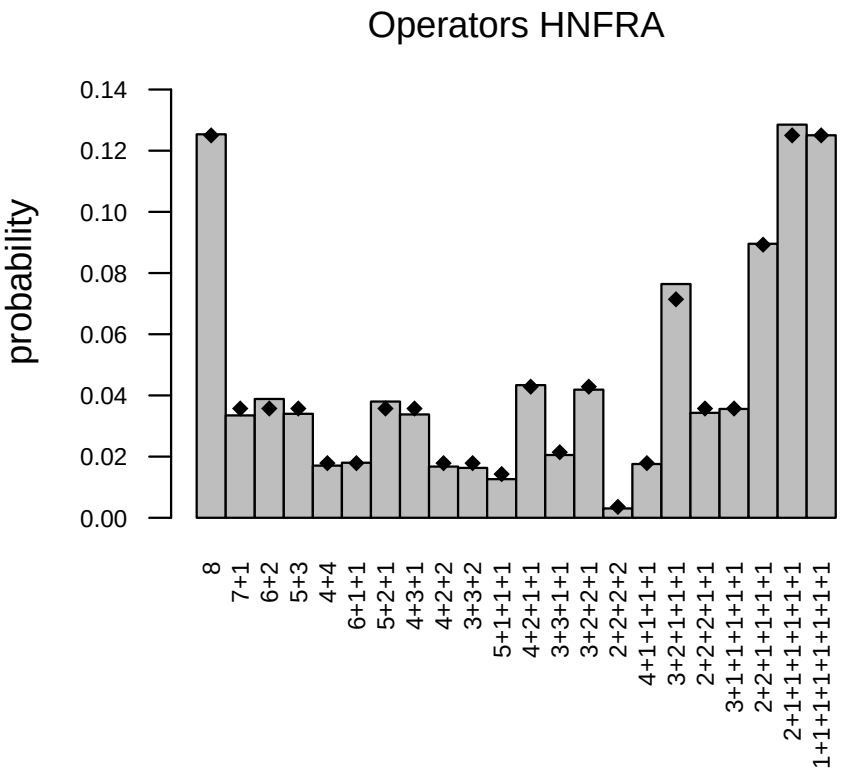

Figure 3: Estimated posterior probabilities for the 22 partitions of the integer 8 are shown in grey bars. The theoretical values are shown as diamonds. See text for further explanation.

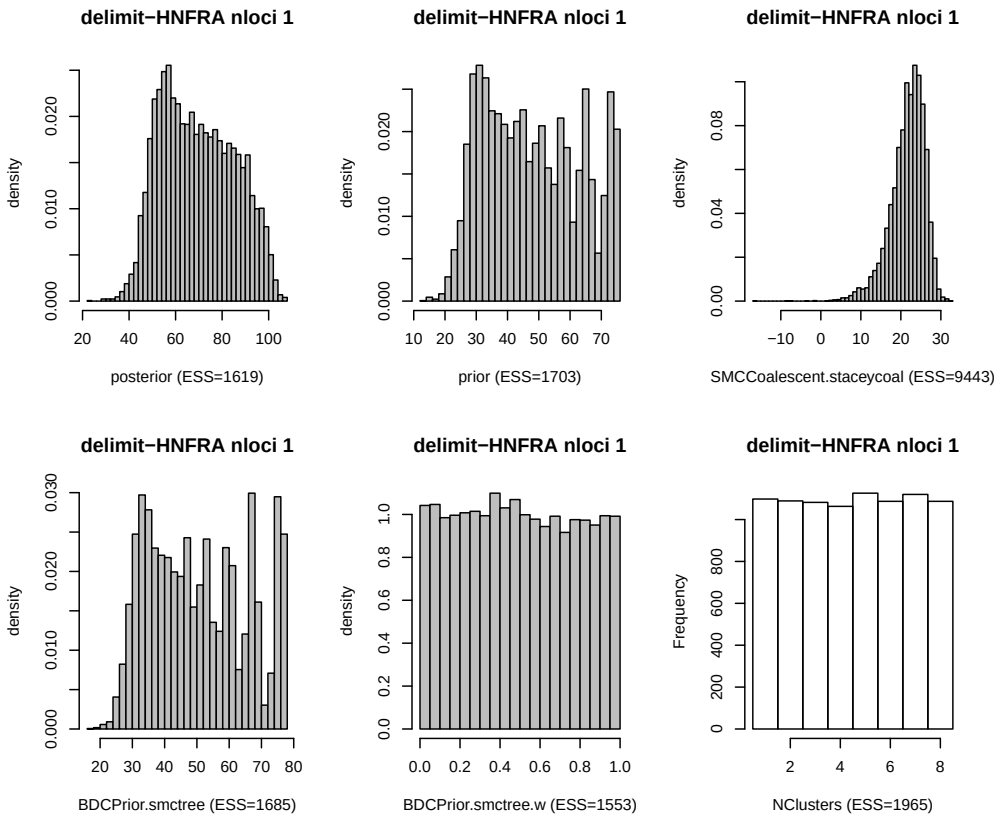

Figure 4: Histograms for various parameters and other quantities.

4 CODA summary for fixed test.

```
Operators  HNFRA  nloci  1

ESS values
                  Sample               posterior              likelihood
                1.500086             5458.099198                0.000000
                   prior SMCCoalescent.staceycoal     SMCCoalescent.popSF
             3306.626168             8403.021910                0.000000
        BDCPrior.smctree  BDCPrior.origin.height       GTreeLhood.seq01
             2871.358977             2062.480312                0.000000
      HKY.HKYkappa.seq01     HKY.HKYfreqs.seq011     HKY.HKYfreqs.seq012
             8752.000000             8405.197120             8752.000000
     HKY.HKYfreqs.seq013     HKY.HKYfreqs.seq014               NClusters
             8752.000000             8752.000000             8752.000000
                 PopSize
             8752.000000

Iterations = 1:8752
Thinning interval = 1
Number of chains = 1
Sample size per chain = 8752

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                             Mean        SD  Naive SE Time-series SE
Sample                   5.624e+08 2.527e+08 2.701e+06      2.063e+08
posterior                5.011e+01 5.091e+00 5.442e-02      6.891e-02
likelihood               5.963e-14 1.388e-14 1.484e-16      0.000e+00
prior                    2.825e+01 3.165e+00 3.383e-02      5.504e-02
SMCCoalescent.staceycoal 2.186e+01 3.988e+00 4.263e-02      4.351e-02
SMCCoalescent.popSF      1.000e-02 0.000e+00 0.000e+00      0.000e+00
BDCPrior.smctree         3.086e+01 2.825e+00 3.020e-02      5.272e-02
BDCPrior.origin.height   2.741e-02 1.219e-02 1.303e-04      2.684e-04
GTreeLhood.seq01         5.963e-14 1.388e-14 1.484e-16      0.000e+00
HKY.HKYkappa.seq01       5.761e+00 1.146e+01 1.225e-01      1.225e-01
HKY.HKYfreqs.seq011      2.515e-01 1.959e-01 2.094e-03      2.137e-03
HKY.HKYfreqs.seq012      2.501e-01 1.950e-01 2.085e-03      2.085e-03
HKY.HKYfreqs.seq013      2.477e-01 1.957e-01 2.092e-03      2.092e-03
HKY.HKYfreqs.seq014      2.507e-01 1.958e-01 2.093e-03      2.093e-03
NClusters                7.974e+00 1.590e-01 1.700e-03      1.700e-03
PopSize                  1.027e-02 1.145e-02 1.224e-04      1.224e-04
```

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
Sample	1.468e+08	3.437e+08	5.624e+08	7.812e+08	9.781e+08
posterior	3.893e+01	4.713e+01	5.056e+01	5.369e+01	5.862e+01
likelihood	3.308e-14	4.929e-14	5.818e-14	7.194e-14	8.438e-14
prior	2.110e+01	2.635e+01	2.863e+01	3.050e+01	3.340e+01
SMCCoalescent.staceycoal	1.229e+01	1.980e+01	2.249e+01	2.471e+01	2.763e+01
SMCCoalescent.popSF	1.000e-02	1.000e-02	1.000e-02	1.000e-02	1.000e-02
BDCPrior.smctree	2.460e+01	2.915e+01	3.115e+01	3.291e+01	3.548e+01
BDCPrior.origin.height	1.005e-02	1.863e-02	2.523e-02	3.376e-02	5.615e-02
GTreeLhood.seq01	3.308e-14	4.929e-14	5.818e-14	7.194e-14	8.438e-14
HKY.HKYkappa.seq01	2.335e-01	1.165e+00	2.660e+00	6.223e+00	2.906e+01
HKY.HKYfreqs.seq011	7.471e-03	9.034e-02	2.077e-01	3.757e-01	7.195e-01
HKY.HKYfreqs.seq012	8.528e-03	9.030e-02	2.064e-01	3.708e-01	7.094e-01
HKY.HKYfreqs.seq013	8.759e-03	8.825e-02	2.004e-01	3.674e-01	7.103e-01
HKY.HKYfreqs.seq014	7.396e-03	9.000e-02	2.054e-01	3.723e-01	7.086e-01
NClusters	7.000e+00	8.000e+00	8.000e+00	8.000e+00	8.000e+00
PopSize	2.770e-03	5.141e-03	7.553e-03	1.163e-02	3.329e-02

5 CODA summary for delimitation test.

```
Operators  HNFRA  nloci  1

ESS values
                  Sample               posterior              likelihood
                1.500086             1619.449787                0.000000
                   prior SMCCoalescent.staceycoal     SMCCoalescent.popSF
             1702.910430             9443.484279                0.000000
        BDCPrior.smctree      BDCPrior.smctree.w  BDCPrior.origin.height
             1685.138842             1553.052019             1226.731414
        GTreeLhood.seq01      HKY.HKYkappa.seq01     HKY.HKYfreqs.seq011
                0.000000             8752.000000             8752.000000
     HKY.HKYfreqs.seq012     HKY.HKYfreqs.seq013     HKY.HKYfreqs.seq014
             8752.000000             8752.000000             8752.000000
               NClusters                 PopSize
             1965.282064             8752.000000

Iterations = 1:8752
Thinning interval = 1
Number of chains = 1
Sample size per chain = 8752

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                             Mean        SD  Naive SE Time-series SE
Sample                   5.624e+08 2.527e+08 2.701e+06      2.063e+08
posterior                6.960e+01 1.572e+01 1.680e-01      3.906e-01
likelihood              -9.407e-14 2.327e-14 2.488e-16      0.000e+00
prior                    4.793e+01 1.522e+01 1.627e-01      3.688e-01
SMCCoalescent.staceycoal 2.168e+01 4.289e+00 4.585e-02      4.414e-02
SMCCoalescent.popSF      1.000e-02 0.000e+00 0.000e+00      0.000e+00
BDCPrior.smctree         5.056e+01 1.510e+01 1.615e-01      3.679e-01
BDCPrior.smctree.w       4.940e-01 2.881e-01 3.079e-03      7.310e-03
BDCPrior.origin.height   2.059e-02 1.298e-02 1.388e-04      3.707e-04
GTreeLhood.seq01        -9.407e-14 2.327e-14 2.488e-16      0.000e+00
HKY.HKYkappa.seq01       5.988e+00 1.201e+01 1.284e-01      1.284e-01
HKY.HKYfreqs.seq011      2.535e-01 1.955e-01 2.090e-03      2.090e-03
HKY.HKYfreqs.seq012      2.484e-01 1.921e-01 2.053e-03      2.053e-03
HKY.HKYfreqs.seq013      2.501e-01 1.948e-01 2.082e-03      2.082e-03
HKY.HKYfreqs.seq014      2.480e-01 1.928e-01 2.060e-03      2.060e-03
NClusters                4.509e+00 2.293e+00 2.451e-02      5.172e-02
PopSize                  9.882e-03 9.425e-03 1.007e-04      1.007e-04
```

2. Quantiles for each variable:

	2.5%	25%	50%	75%	97.5%
Sample	1.468e+08	3.437e+08	5.624e+08	7.812e+08	9.781e+08
posterior	4.446e+01	5.646e+01	6.831e+01	8.235e+01	9.892e+01
likelihood	-1.343e-13	-1.107e-13	-9.881e-14	-7.416e-14	-5.174e-14
prior	2.453e+01	3.449e+01	4.621e+01	6.040e+01	7.448e+01
SMCCoalescent.staceycoal	1.130e+01	1.939e+01	2.233e+01	2.474e+01	2.811e+01
SMCCoalescent.popSF	1.000e-02	1.000e-02	1.000e-02	1.000e-02	1.000e-02
BDCPrior.smctree	2.780e+01	3.710e+01	4.888e+01	6.323e+01	7.654e+01
BDCPrior.smctree.w	2.675e-02	2.463e-01	4.866e-01	7.448e-01	9.752e-01
BDCPrior.origin.height	1.661e-03	1.114e-02	1.859e-02	2.738e-02	5.188e-02
GTreeLhood.seq01	-1.343e-13	-1.107e-13	-9.881e-14	-7.416e-14	-5.174e-14
HKY.HKYkappa.seq01	2.434e-01	1.147e+00	2.680e+00	6.306e+00	3.388e+01
HKY.HKYfreqs.seq011	8.125e-03	9.212e-02	2.100e-01	3.796e-01	7.109e-01
HKY.HKYfreqs.seq012	8.067e-03	9.057e-02	2.046e-01	3.685e-01	6.986e-01
HKY.HKYfreqs.seq013	8.957e-03	9.103e-02	2.059e-01	3.651e-01	7.173e-01
HKY.HKYfreqs.seq014	8.295e-03	8.974e-02	2.051e-01	3.664e-01	6.989e-01
NClusters	1.000e+00	3.000e+00	5.000e+00	7.000e+00	8.000e+00
PopSize	2.775e-03	5.033e-03	7.400e-03	1.145e-02	3.081e-02

6 R script used with STACEY_XML

```
library(ape)
library(hash)

RcodeSourceDirectory <- "C:/Users/Work/AAA/Programming/biology/STACEY_XML"

source(paste0(RcodeSourceDirectory, "/", "analysis-structure-lib.r"))
source(paste0(RcodeSourceDirectory, "/", "STACEYxml-lib.r"))



##############################################################################################
#########################  Making BEAST XML files ############################################
##############################################################################################

run.options <- function(ops.set) {
  ro <- c(
    sampledgtrees.fpathbase = "gtrees",
    sampledsmctrees.fpath = "smctrees.txt",
    sampledparams.fpath = "params.txt",
    chainlength =    "1000000000",
    store.every =      "10000000",
    params.logevery =    "100000",
    smctree.logevery =   "100000",
    gtrees.logevery =   "1000000",
    screen.logevery =    "100000",
    op.wt.Reheight =     operator.weights.table[ops.set, "H"],
    op.wt.Nudge =        operator.weights.table[ops.set, "N"],
    op.wt.Focused =      operator.weights.table[ops.set, "F"],
    op.wt.Regraft =      operator.weights.table[ops.set, "R"],
    op.wt.BAdjust =      operator.weights.table[ops.set, "A"]
  )
  ro
}



make.xml <- function(scenario, ops.set, nloci) {

  dname <- paste0(scenario, "-", ops.set, "-g", nloci)
  beastxml.dpath <-  paste0(base.dpath, "/", dname)
  if (!dir.exists(beastxml.dpath)) {
    dir.create(beastxml.dpath)
  }
  beastxml.fpath <-  paste0(beastxml.dpath, "/beast.xml")
  data.fnames <- paste0("a-", sprintf("%02d", 1:nloci), ".nex")
  names(data.fnames) <- paste0("seq", sprintf("%02d", 1:nloci))
```

```
taxa.table <- matrix("", nrow=8, ncol=2)
for (ind in 1:8) {
  txname <- sprintf("a%02d", ind)
  taxa.table[ind,] <- c(paste0(txname,"A"), txname)
}
colnames(taxa.table) <- c("taxon", "mincluster")

TheAnalysis(id=paste0("Test-", dname),
            data.dpath=data.dpath,
            data.fnames=data.fnames,
            taxa.table=taxa.table,
            run.options=run.options(ops.set))
cat("made TheAnalysis for scenario ", scenario, " operator set ", ops.set, " and ", nloci, " loci\n")

gtree.priors <- get.GTreePriors()
for (i in 1:length(gtree.priors)) {
  GTreePrior(gtree.priors[[i]]$id, SMCtree("smctree"))
}

SMCtree("smctree", BDCPrior("smctree"), SMCCoalescent("staceycoal"))
if (scenario == "fixed") {
  BDCPrior("smctree", growthrate=Fixed("smctree.g"),
           reldeath=Fixed("smctree.rd"),
           w=Fixed("smctree.w"),
           oh="origin.height",
           eps=epsilon)
  Fixed("smctree.w", 0)
} else if (scenario == "delimit") {
  BDCPrior("smctree", growthrate=Fixed("smctree.g"),
           reldeath=Fixed("smctree.rd"),
           w=Beta("smctree.w"),
           oh="origin.height",
           eps=epsilon)
  Beta("smctree.w", 1, 1)
} else {
  stop("Unknown scenario in make.xml()")
}

Fixed("smctree.g", 100)
Fixed("smctree.rd", 0)
SMCCoalescent("staceycoal", invgammamix=InvGammaMix("popBV"), popSF=Fixed("popSF"))
Fixed("popSF", 0.01)
InvGammaMix("popBV", weights=c(1), alphas=c(3), betas=c(2))

cat("made SMCtree, BDCPrior, SMCCoalescent", date(), "\n")

gtree.brms <- get.gtree.BranchRMs()
for (i in 1:length(gtree.brms)) {
```

```
    id <- gtree.brms[[i]]$id
    BranchRM(id, StrictClock(id))
    StrictClock(id, 1.0)
  }

  gtree.ploidys <- get.gtree.Ploidys()
  for (i in 1:length(gtree.ploidys)) {
    Ploidy(gtree.ploidys[[i]]$id, 2)
  }

  gtree.prms <- get.gtree.PartitionRateMs()
  for (i in 1:length(gtree.prms)) {
    id <- gtree.prms[[i]]$id
    if (i==1) {
      PartitionRateM(id, 1.0)
    } else {
      PartitionRateM(id, LogNorm(id))
      LogNorm(id, 0.0, 1.0)
    }
  }

  gtree.substs <- get.gtree.SubstMs()
  for (i in 1:length(gtree.substs)) {
    id <- gtree.substs[[i]]$id
    SubstM(id, HKY(id))
    freqsID <- paste0("HKYfreqs.", id)
    kappaID <- paste0("HKYkappa.", id)
    #   HKY(id, UniformUnitSimplex(freqsID), LogNorm(kappaID))
    #   UniformUnitSimplex(freqsID, 4)
    #   LogNorm(kappaID, 1.0, 1.25)
    HKY(id, UniformUnitSimplex(freqsID), LogNorm(kappaID))
    UniformUnitSimplex(freqsID, 4)
    LogNorm(kappaID, 1.0, 1.25)
  }

  gtree.sitehets <- get.gtree.SiteHets()
  for (i in 1:length(gtree.sitehets)) {
    SiteHet(gtree.sitehets[[i]]$id, "None")
  }

  cat(date(),"Made the analysis structure\n")
  #cat.structure("C:/Users/Work/Desktop/TheAnalysisStructure.txt")

  xmlFile.from.analysis.structure(beastxml.fpath)
  reset.the.analysis()
  reset.xml.lib()
}
```

```
make.set.of.xmls <- function() {
  for (scenario in c("delimit", "fixed")) {
    for (ops.set in rownames(operator.weights.table)) {
      for (nloci in c(1,1)) {
        make.xml(scenario, ops.set, nloci)
      }
    }
  }
}



##############################################################################################
#################### Running BEAST and SDA ###################################################
##############################################################################################


run.beast <- function(ops.sets) {
  SDA.jar.fpath <- "C:/Users/Work/AAA/Programming/biology/Simulations/SpeciesDelimLookAtResults/speciesl
  BEAST.jar.fpath <- "C:/Users/Work/AAA/biology/SoftWare/BEAST.v2.3.2/lib/beast.jar"

  for (scenario in c("delimit", "fixed")) {
    for (ops in ops.sets) {
      for (nloci in c(1)) {
        dname <- paste0(scenario, "-", ops, "-g", nloci)
        beastxml.fpath <- paste0(base.dpath, "/", dname, "/beast.xml")
        command <- paste0("java -jar ", BEAST.jar.fpath, " -java -overwrite -working -seed 46 ", beastx
        cat(command, "\n")
        system(command)
      }
    }
  }
}



##############################################################################
##############################################################################
##############################################################################

base.dpath <- "C:/Users/Work/AAA/Programming/ProgramOutput/STACEY-DevVersion/Jan2016TESTS"
data.dpath <- paste0(base.dpath, "/data")

epsilon <- 0.00003

operator.weights.table <-
  matrix(c(
    7/3,   0,   0,   0,   0,
```

```
    7/6, 7/2,   0,   0,   0,
    7/6,   0, 7/2,   0,   0,
    7/6,   0,   0, 7/2,   0,
    7/6,   0,   0,   0, 7/2,
    1,   1,   1,   1,   1
  ), nrow = 6, ncol = 5, byrow=TRUE)
rownames(operator.weights.table) <- c("H", "HN", "HF", "HR", "HA", "HNFRA")
colnames(operator.weights.table) <- c("H", "N", "F", "R", "A")

make.set.of.xmls()
#run.beast(rownames(operator.weights.table))
run.beast("HNFRA")
```

References

T Gernhard. The conditioned reconstructed process. *J. Theo. Biol.*, 253:769–778, 2008.

G Jones. Species delimitation and phylogeny estimation under the multispecies coalescent. *bioRχiv*, 2015. doi: 10.1101/010199. URL http://biorxiv.org/content/early/2015/03/22/010199.