BP&P

VERSION 4.7

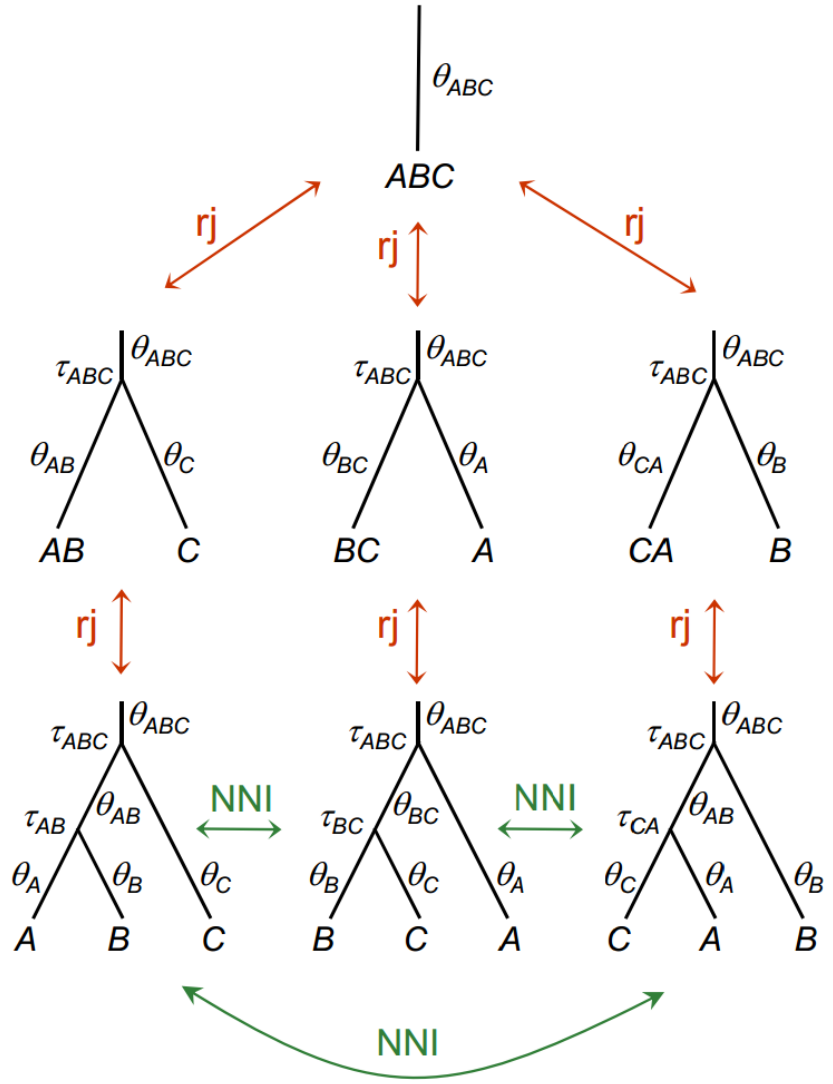
- Bayesian (MCMC) program for analyzing DNA sequence alignments from multiple loci under the multispecies coalescent (MSC)
- https://github.com/bpp/bpp

Unguided Species Delimitation Using DNA Sequence Data from Multiple Loci

Ziheng Yang[1,2] and Bruce Rannala*[,1,3]

Bayesian Phylogenetics and Phylogeography (BPP)

- Bayesian modeling approach generating speciation posterior probabilities
- Sequence data as input
- Multiple loci
- Uncertainty in gene phylogeny
- With or without guide species tree as input
- Prior for θ: effective population size parameter
- Prior for τ: species divergence time parameter
- Available at: https://github.com/bpp/bpp

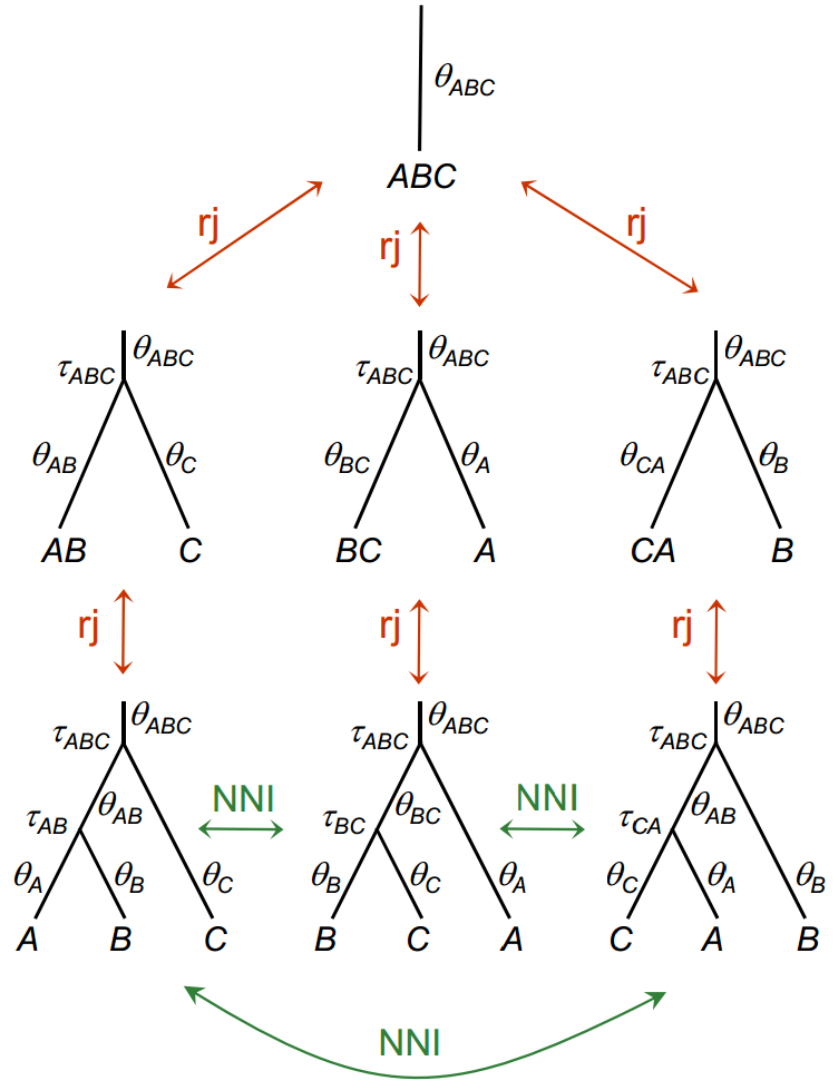


Rannala & Yang. 2003. Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. – *Genetics* 164: 1645-1656.
Yang & Rannala. 2010. Bayesian species delimitation using multilocus sequence data. – *PNAS* 107: 9264–9269.

Bayesian Phylogenetics and Phylogeography (BPP)

- Nearest-neighbor interchange (NNI) used to move between species phylogenies with the species delimitation fixed

- Reversible-jump Markov Chain Monte Carlo (rjMCMC) to move between species delimitations when the underlying guide tree is fixed



Rannala & Yang. 2003. Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. – *Genetics* 164: 1645-1656.
Yang & Rannala. 2010. Bayesian species delimitation using multilocus sequence data. – *PNAS* 107: 9264–9269.

Assumptions of the Coalescent Model

1) No recombination within a locus

2) Free recombination between loci

3) No migration (gene flow) between species

4) Neutral evolution

5) clock-like evolution.

Recent implementations of BPP can accommodate for:

1) Gene-flow

PNAS RESEARCH ARTICLE | EVOLUTION OPEN ACCESS

Efficient Bayesian inference under the multispecies coalescent with migration

Tomáš Flouri[a], Xiyun Jiao[b], Jun Huang[c], Bruce Rannala[d,1], and Ziheng Yang[a,1]

Edited by Rasmus Nielsen, University of California, Berkeley, CA; received June 25, 2023; accepted August 15, 2023

Recent implementations of BPP can accommodate for:

1) Gene-flow

2) Non-clockwise evolution

Bayesian Phylogenetic Inference using Relaxed-clocks and the Multispecies Coalescent

Tomáš Flouri [1], Jun Huang [1,2], Xiyun Jiao, [1,3] Paschalia Kapli [1], Bruce Rannala [4], and Ziheng Yang [*,1]

Recent implementations of BPP can accommodate for:

1) Gene-flow
2) Non-clockwise evolution
3) Scalable to phylogenomic data

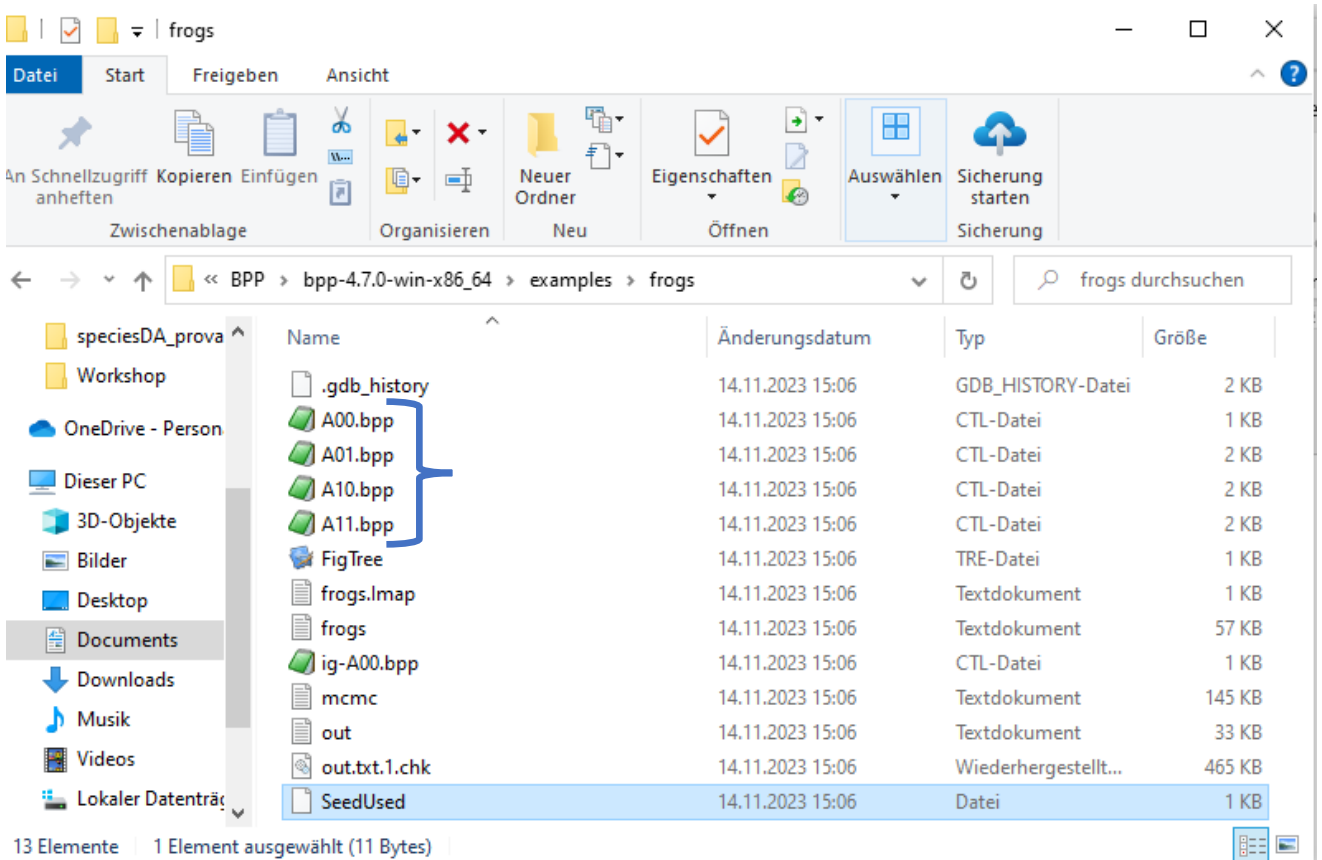
- **A00** (speciesdelimitation = 0, speciestree = 0): estimation of the parameters τ (species divergence times) and θ (effective population sizes) under the MSC when the species tree model is given (Rannala and Yang, 2003; Flouri et al., 2020a);
- **A01** (speciesdelimitation = 0, speciestree = 1): inference of the species tree when the assignments are given by the user (Rannala and Yang, 2017);
- **A10** (speciesdelimitation = 1, speciestree = 0): species delimitation using a user-specified guide tree (Yang and Rannala, 2010; Rannala and Yang, 2013);
- **A11** (speciesdelimitation = 1, speciestree = 1): joint species delimitation and species tree inference of unguided species delimitation (Yang and Rannala, 2014).

- Command line software.
- No graphical interface (unlike BEAST)

  take a look to the software folder and examples…

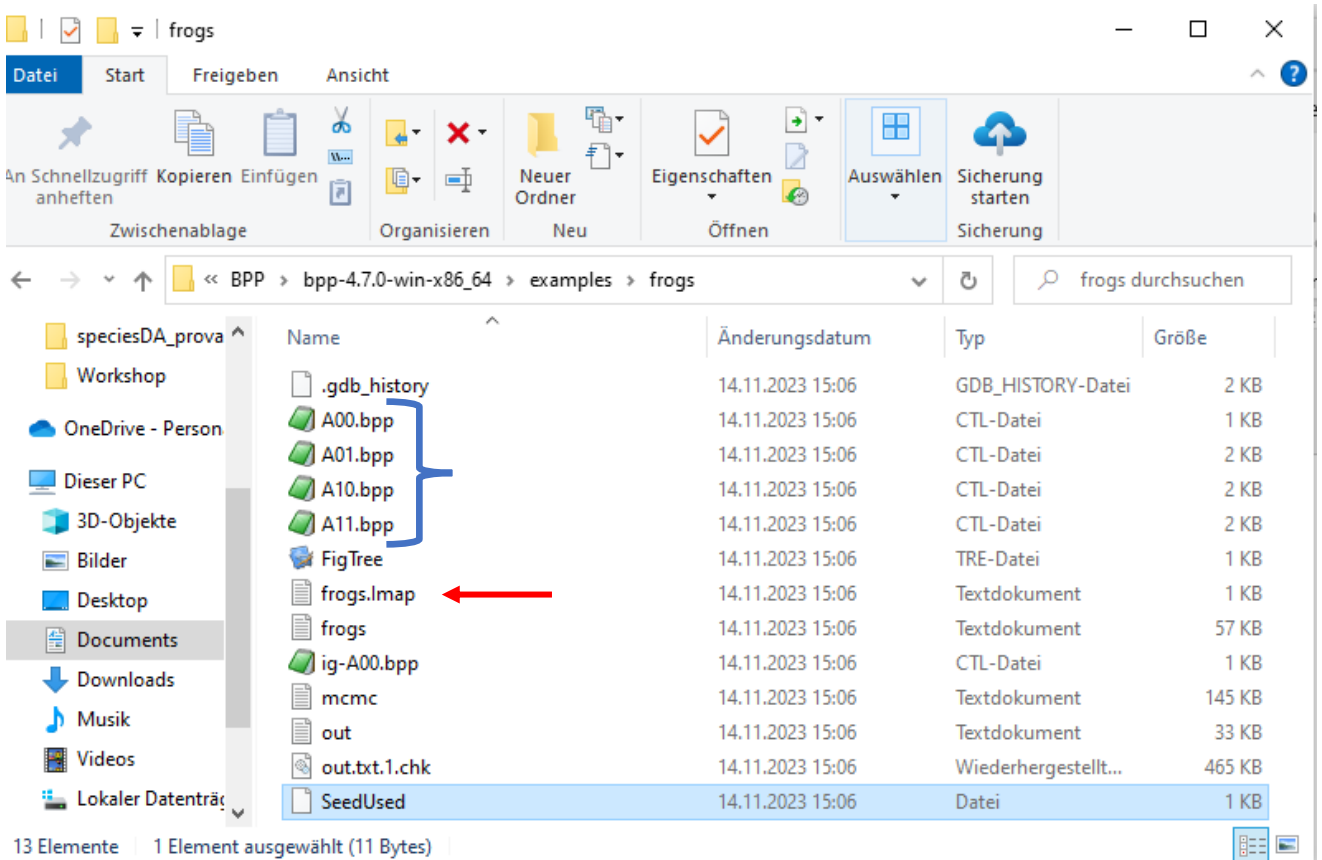
- Command line software.
- No graphical interface (unlike BEAST)

  - **Settings file**



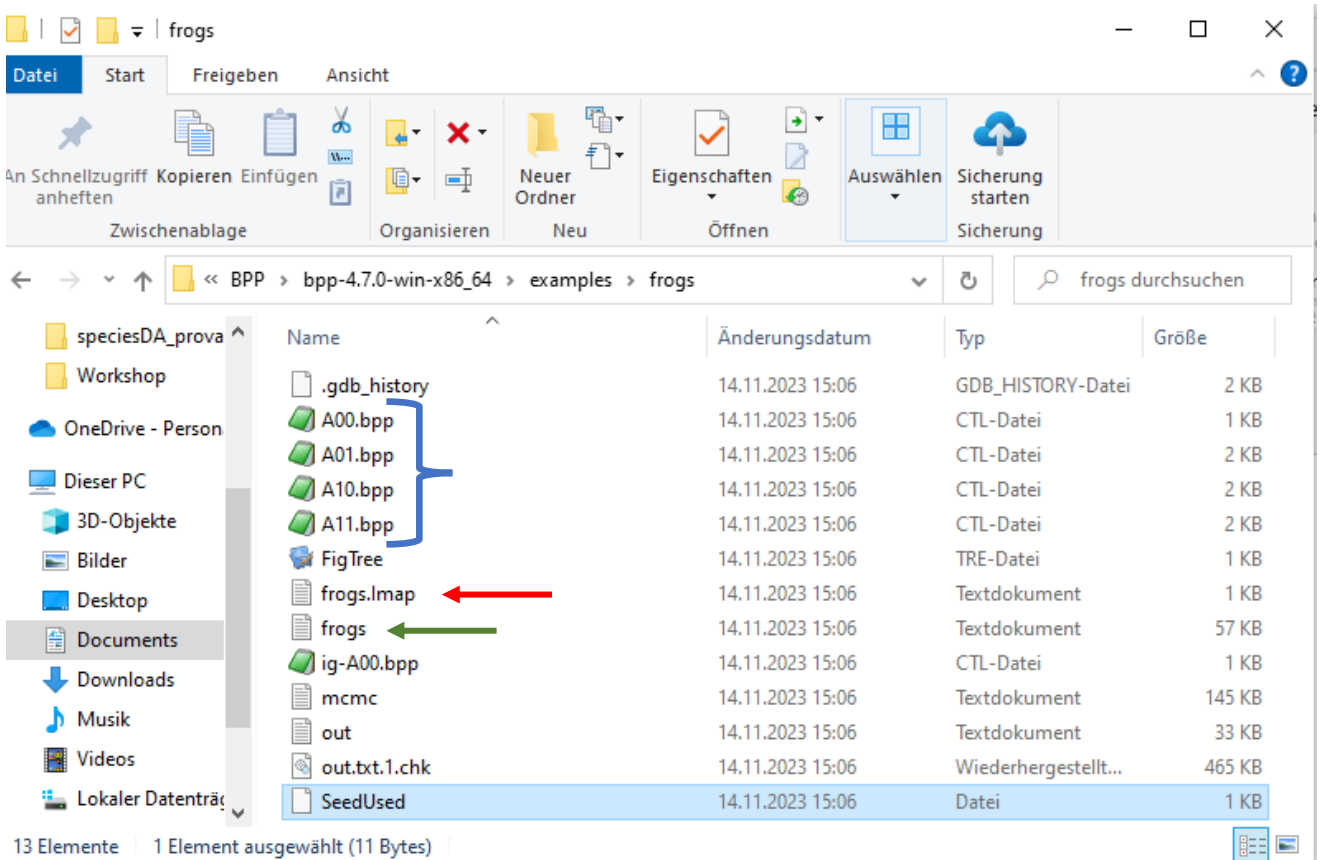
- Command line software.
- No graphical interface (unlike BEAST)

  - **Settings file**
  - **Imap (sequence-taxon association)**



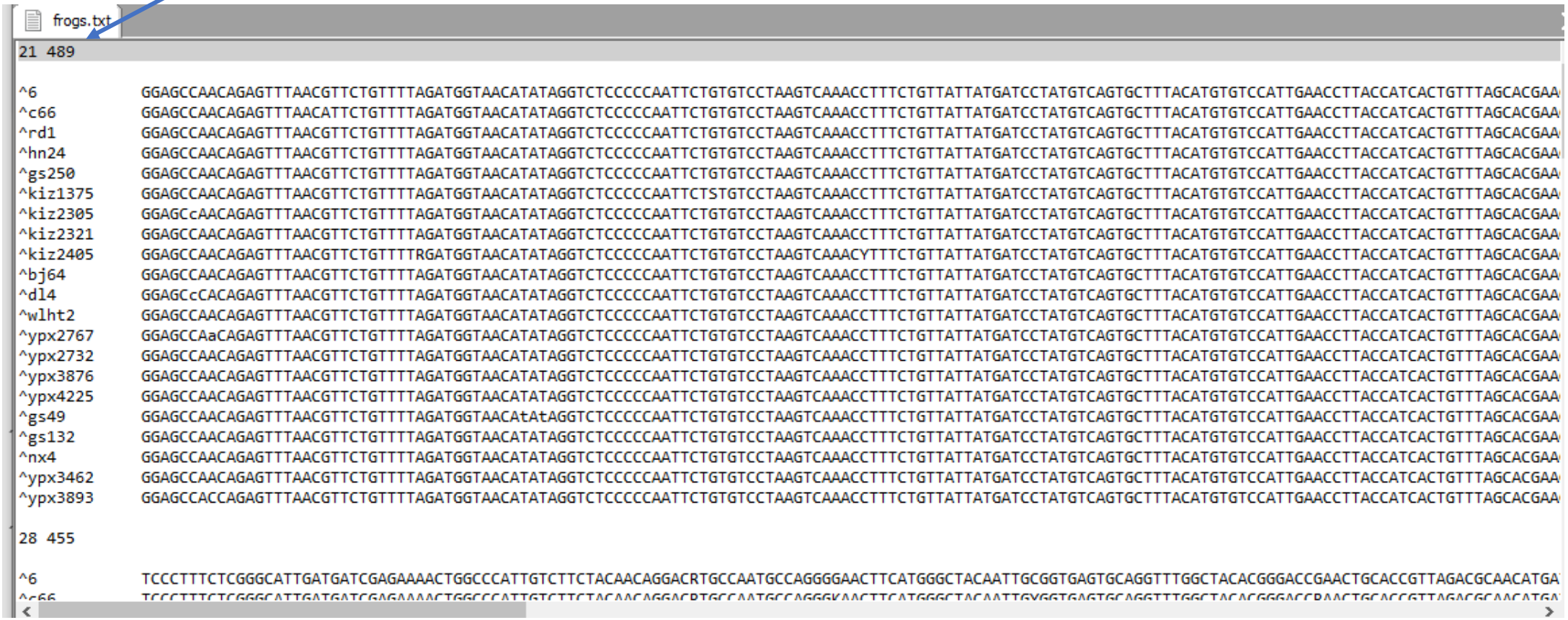
- Command line software.
- No graphical interface (unlike BEAST)

  - **Settings file**
  - **Imap (sequence-taxon association)**
  - **Alignments file**

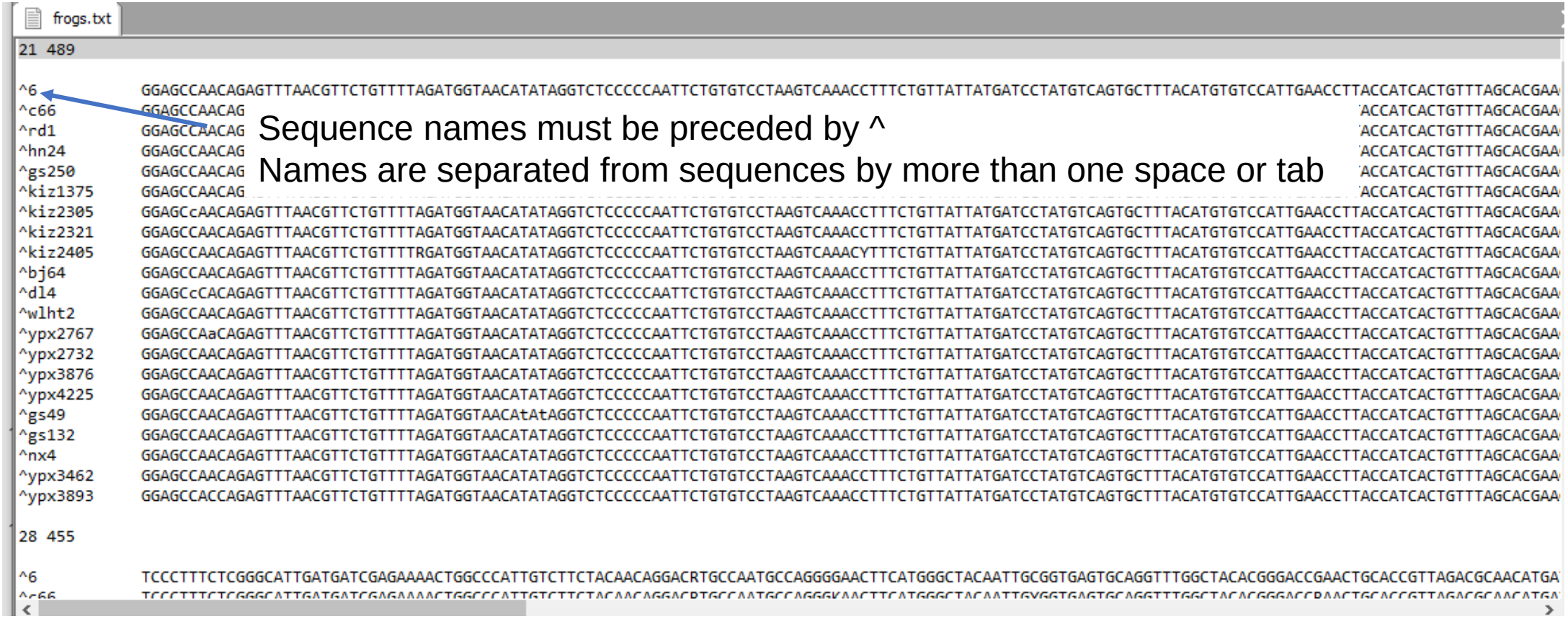


Sequence alignments must be in phylip format

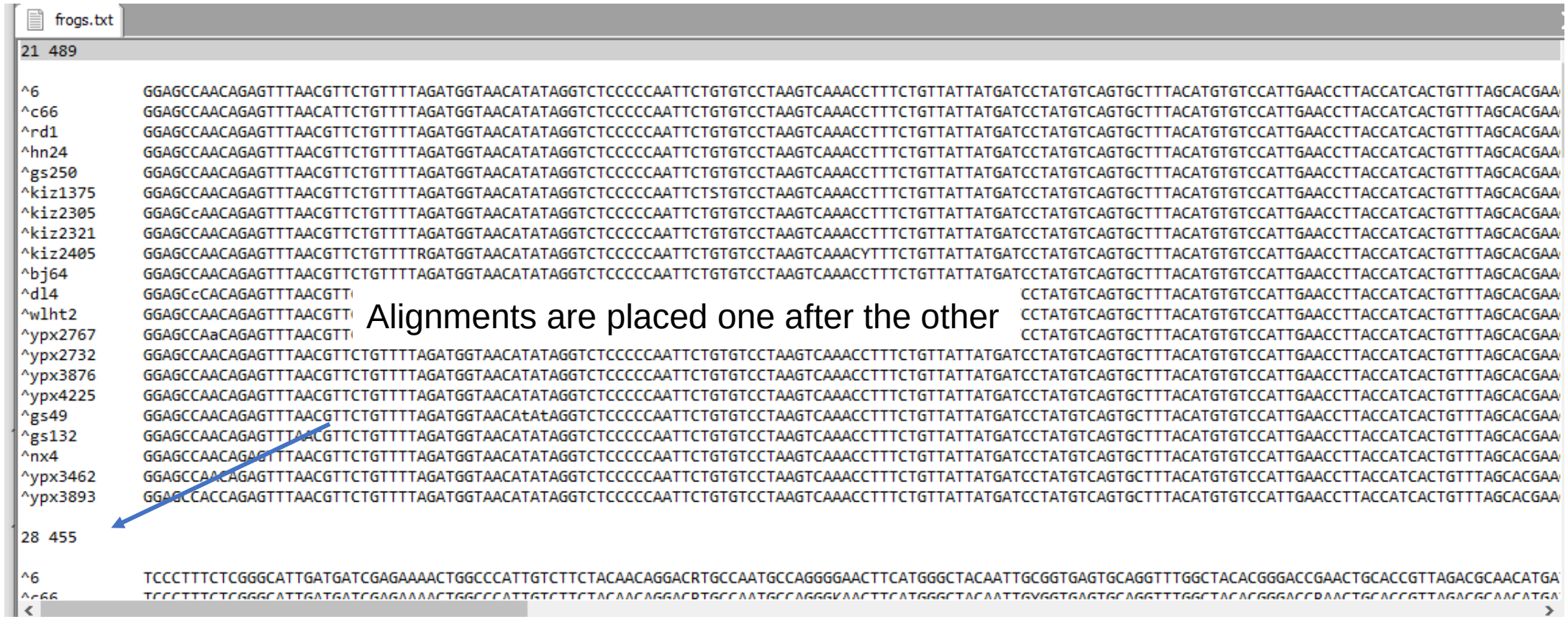
First line contains information on sequence number and length of the alignment



Sequence alignments must be in phylip format

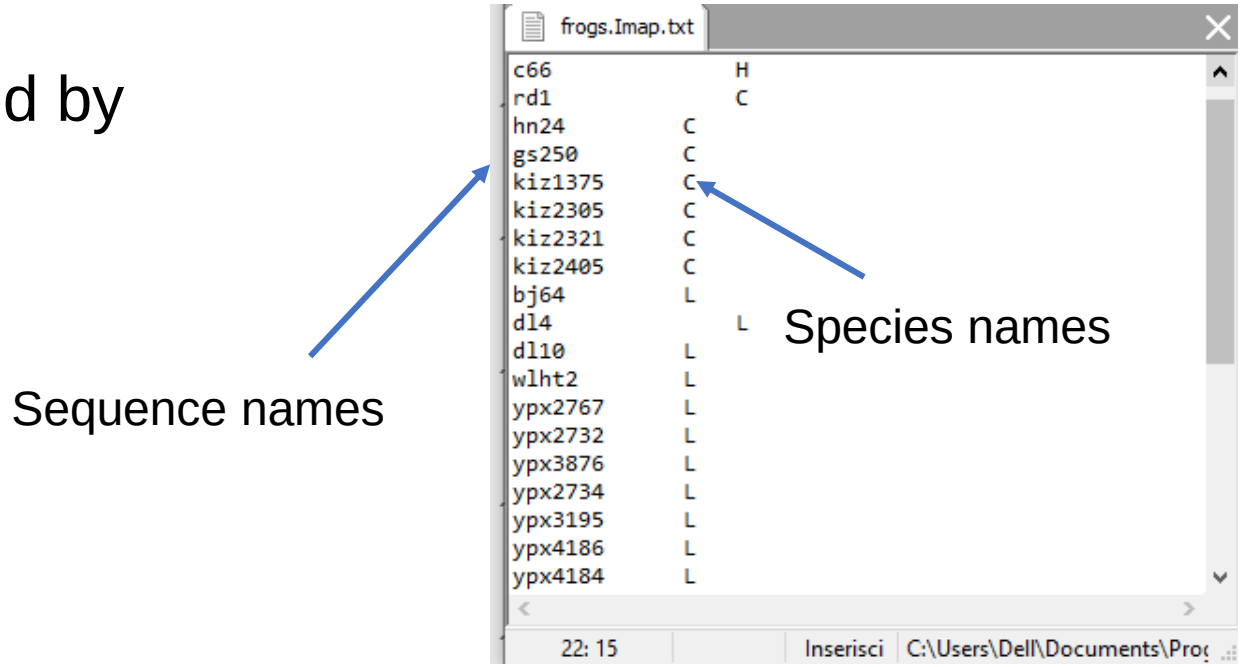


Sequence alignments must be in phylip format



The Imap files contain information of sequence-taxa assignment:

- sequences from a sample can have different names in different analyses
- One could assign sequence from two or more samples to the same taxon (e.g., two individuals of the same population). They will not be divided during the analyses
- Two columns separated by spaces or tabs



The setting file

```
         seed = -1
      seqfile = ../frogs.txt
     Imapfile = ../frogs.Imap.txt
      outfile = out.txt
     mcmcfile = mcmc.txt

 speciesdelimitation = 0       # fixed species delimitation
         speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1       # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H        # number of species and list of species labels
                     9  7 14  2        # max number of sequences from each species at a loci
                  ((K, C), (L, H));   # initial species phylogeny
        diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0            * 0: nothing;  1 : save;  2: read
       usedata = 1             # 0: no data (prior); 1: seq like
         nloci = 5             # number of data sets to read in seqfile
     cleandata = 0             # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e    # invgamma(a,b) for theta parameters
       tauprior = 3 0.004      # invgamma(a,b) for root tau

*      heredity = 1 4 4    # (0: no variation, 1: estimate, 2: from file)
*     locusrate = 0 2.0    # (0: no variation, 1: estimate, 2: from file)

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

          print = 1 0 0 0      # print MCMC samples, locusrate, heredity scalars, gene trees
         burnin = 8000         # burn-In
       sampfreq = 2            # frequency of sampling (sample every second MCMC iteration)
        nsample = 100000       # total number of samples to log
```

File Names

The setting file

```
         seed = -1
      seqfile = ../frogs.txt
     Imapfile = ../frogs.Imap.txt
      outfile = out.txt
     mcmcfile = mcmc.txt

 speciesdelimitation = 0        # fixed species delimitation
        speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1        # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H        # number of species and list of species labels
                     9  7 14  2        # max number of sequences from each species at a loci
                  ((K, C), (L, H));   # initial species phylogeny
        diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0             * 0: nothing;  1 : save;  2: read
       usedata = 1              # 0: no data (prior); 1: seq like
         nloci = 5              # number of data sets to read in seqfile
     cleandata = 0              # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e     # invgamma(a,b) for theta parameters
       tauprior = 3 0.004       # invgamma(a,b) for root tau

*      heredity = 1 4 4    # (0: no variation, 1: estimate, 2: from file)
*     locusrate = 0 2.0    # (0: no variation, 1: estimate, 2: from file)

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

          print = 1 0 0 0       # print MCMC samples, locusrate, heredity scalars, gene trees
         burnin = 8000          # burn-In
       sampfreq = 2             # frequency of sampling (sample every second MCMC iteration)
        nsample = 100000        # total number of samples to log
```

If or not you want to perform species delimitation (1=yes)

The setting file

```
          seed = -1
       seqfile = ../frogs.txt
      Imapfile = ../frogs.Imap.txt
       outfile = out.txt
      mcmcfile = mcmc.txt

 speciesdelimitation = 0          # fixed species delimitation
         speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1       # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H        # number of species and list of species labels
                     9  7 14  2        # max number of sequences from each species at a loci
                  ((K, C), (L, H));   # initial species phylogeny
        diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0            * 0: nothing;  1 : save;  2: read
       usedata = 1             # 0: no data (prior); 1: seq like
         nloci = 5             # number of data sets to read in seqfile
     cleandata = 0             # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e    # invgamma(a,b) for theta parameters
       tauprior = 3 0.004      # invgamma(a,b) for root tau

*      heredity = 1 4 4    # (0: no variation, 1: estimate, 2: from file)
*     locusrate = 0 2.0    # (0: no variation, 1: estimate, 2: from file)

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

          print = 1 0 0 0      # print MCMC samples, locusrate, heredity scalars, gene trees
         burnin = 8000         # burn-In
       sampfreq = 2            # frequency of sampling (sample every second MCMC iteration)
        nsample = 100000       # total number of samples to log
```

If or not you want to infer the species tree (1=yes)

The setting file

```
         seed = -1
      seqfile = ../frogs.txt
     Imapfile = ../frogs.Imap.txt
      outfile = out.txt
     mcmcfile = mcmc.txt

 speciesdelimitation = 0       # fixed species delimitation
        speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1       # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H
                  9  7 14  2          # max number of sequences from each species at a loci
                ((K, C), (L, H));    # initial species phylogeny
       diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0                  * 0: nothing;  1 : save;  2: read
       usedata = 1                   # 0: no data (prior); 1: seq like
         nloci = 5                   # number of data sets to read in seqfile
     cleandata = 0                   # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e    # invgamma(a,b) for theta parameters
       tauprior = 3 0.004      # invgamma(a,b) for root tau

*      heredity = 1 4 4    # (0: no variation, 1: estimate, 2: from file)
*     locusrate = 0 2.0    # (0: no variation, 1: estimate, 2: from file)

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

          print = 1 0 0 0      # print MCMC samples, locusrate, heredity scalars, gene trees
         burnin = 8000         # burn-In
       sampfreq = 2            # frequency of sampling (sample every second MCMC iteration)
        nsample = 100000       # total number of samples to log
```

Number and names of the samples/taxa/pre-defined species

The setting file

```
         seed = -1
      seqfile = ../frogs.txt
     Imapfile = ../frogs.Imap.txt
      outfile = out.txt
     mcmcfile = mcmc.txt

 speciesdelimitation = 0       # fixed species delimitation
        speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1       # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H
                     9  7 14  2
                  ((K, C), (L, H));   # initial species phylogeny
        diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0            * 0: nothing;  1 : save;  2: read
       usedata = 1             # 0: no data (prior); 1: seq like
         nloci = 5             # number of data sets to read in seqfile
     cleandata = 0             # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e    # invgamma(a,b) for theta parameters
       tauprior = 3 0.004      # invgamma(a,b) for root tau

*      heredity = 1 4 4    # (0: no variation, 1: estimate, 2: from file)
*     locusrate = 0 2.0    # (0: no variation, 1: estimate, 2: from file)

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

          print = 1 0 0 0      # print MCMC samples, locusrate, heredity scalars, gene trees
         burnin = 8000         # burn-In
       sampfreq = 2            # frequency of sampling (sample every second MCMC iteration)
        nsample = 100000       # total number of samples to log
```

Max. number of sequences per samples/taxa/pre-defined species

The setting file

```
         seed = -1
      seqfile = ../frogs.txt
     Imapfile = ../frogs.Imap.txt
      outfile = out.txt
     mcmcfile = mcmc.txt

 speciesdelimitation = 0       # fixed species delimitation
        speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1       # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H        # number of species and list of species labels
                     9  7 14  2
                  ((K, C), (L, H));
        diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0            * 0: nothing;  1 : save;  2: read
       usedata = 1             # 0: no data (prior); 1: seq like
         nloci = 5             # number of data sets to read in seqfile
     cleandata = 0             # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e    # invgamma(a,b) for theta parameters
       tauprior = 3 0.004      # invgamma(a,b) for root tau

*      heredity = 1 4 4    # (0: no variation, 1: estimate, 2: from file)
*     locusrate = 0 2.0    # (0: no variation, 1: estimate, 2: from file)

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

          print = 1 0 0 0      # print MCMC samples, locusrate, heredity scalars, gene trees
         burnin = 8000         # burn-In
       sampfreq = 2            # frequency of sampling (sample every second MCMC iteration)
        nsample = 100000       # total number of samples to log
```

Guide species tree (if not inferred) or starting tree

The setting file

```
          seed = -1
       seqfile = ../frogs.txt
      Imapfile = ../frogs.Imap.txt
       outfile = out.txt
      mcmcfile = mcmc.txt

 speciesdelimitation = 0       # fixed species delimitation
         speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1       # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H        # number of species and list of species labels
                     9  7 14  2        # max number of sequences from each species at a loci
                  ((K, C), (L, H));   # initial species phylogeny
        diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0            * 0: nothing;  1 : save;  2: read
       usedata = 1             # 0: no data (prior); 1: seq like
         nloci = 5             # number of data sets to read in seqfile
      cleandata = 0            # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e    :heta parameters
      tauprior = 3 0.004       :oot tau

*      heredity = 1 4 4    # (0: no variation, 1: estimate, 2: from file)
*     locusrate = 0 2.0    # (0: no variation, 1: estimate, 2: from file)

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

          print = 1 0 0 0      # print MCMC samples, locusrate, heredity scalars, gene trees
         burnin = 8000         # burn-In
       sampfreq = 2            # frequency of sampling (sample every second MCMC iteration)
        nsample = 100000       # total number of samples to log
```

Priors for τ and θ

Prior on θ

- Since BPP3.4, both the θ and τ parameters have inverse gamma priors
- The mean of an inverse gamma distribution G(α, β) is:

$$\beta / (\alpha - 1)$$

Prior on θ

- The effective population size (N_e) is an estimate of the population size base sequence data
- It differs among species and even among different genomic regions (e.g., nuclear vs. organellar)

$$N_e = \theta/\mu * 4$$

Depending on the ploidy of the genomic regions… nuclear=4, mitochondrion (human)= 1, in our case??

Effective Population Size

Mutation rate

Genome Wide Analyses Reveal Little Evidence for Adaptive Evolution in Many Plant Species

Toni I. Gossmann,[1] Bao-Hua Song,[2] Aaron J. Windsor,[2] Thomas Mitchell-Olds,[2] Christopher J. Dixon,[3] Maxim V. Kapralov,[3] Dmitry A. Filatov,[3] and Adam Eyre-Walker*,[1]

Prior on θ

- Since BPP3.4, both the θ and τ parameters have inverse gamma priors
- The mean of an inverse gamma distribution IG(α, β) is:

$$\beta / (\alpha - 1)$$

- Say you have an effective population size of 100.000, and a mutation rate of $1*10^{-8}$
- What would you expect for θ?
- giving $\alpha = 3$ (for a diffuse prior) what would you give to α in order to center the distribution to the expected value?

Prior on τ

- Since BPP3.4, both the θ and τ parameters have inverse gamma priors
- The mean of an inverse gamma distribution IG(α, β) is:

$$\beta / (\alpha - 1)$$

Generations (from root to leaves) $= \tau / \mu$

- And for τ? Let's say the samples used diverged 3 millions year ago, the mutation rate of $1*10^{-8}$ (mutations per site per year)
- *Xanthium* is annual
- What would you expect for τ?
- giving $\alpha = 3$ (for a diffuse prior) what would you give to α in order to center the distribution to the expected value?

The setting file

```
         seed = -1
      seqfile = ../frogs.txt
     Imapfile = ../frogs.Imap.txt
      outfile = out.txt
     mcmcfile = mcmc.txt

 speciesdelimitation = 0       # fixed species delimitation
         speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1       # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H        # number of species and list of species labels
                     9  7 14  2        # max number of sequences from each species at a loci
                  ((K, C), (L, H));   # initial species phylogeny
        diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0            * 0: nothing;  1 : save;  2: read
       usedata = 1             # 0: no data (prior); 1: seq like
         nloci = 5             # number of data sets to read in seqfile
      cleandata = 0            # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e    # invgamma(a,b) for theta parameters
       tauprior = 3 0.004      # invgamma(a,b) for root tau

*      heredity = 1 4 4    # (0:
*     locusrate = 0 2.0    # (0:

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

          print = 1 0 0 0      # print MCMC samples, locusrate, heredity scalars, gene trees
         burnin = 8000         # burn-In
       sampfreq = 2            # frequency of sampling (sample every second MCMC iteration)
        nsample = 100000       # total number of samples to log
```

if you want to scale θ according to the genomic region (e.g., nuclear organellar, X or Y chromosomes…)

heredity

- scale the θ parameter in different genomic regions

  heredity = 0 (no variation)
  = 1 (estimated from gamma distribution (provide α, β)
  = 2 (from file)

Genome	Heredity scalar
Nuclear autosome	1
X chromosome	0.75
Y chromosome	0.25
Mitochondrial	0.25

What in our case?

The setting file

```
         seed = -1
      seqfile = ../frogs.txt
     Imapfile = ../frogs.Imap.txt
      outfile = out.txt
     mcmcfile = mcmc.txt

 speciesdelimitation = 0       # fixed species delimitation
        speciestree = 1 0 0 0        # estimate species tree
   speciesmodelprior = 1       # 0:uniform LH; 1:uniform rooted trees; 2:uniformSLH
3:uniformSRooted

   species&tree = 4  K  C  L  H        # number of species and list of species labels
                     9  7 14  2        # max number of sequences from each species at a loci
                  ((K, C), (L, H));   # initial species phylogeny
        diploid =    1  1  1  1        # 0: phased sequences; 1: diploid unphased sequences.

*    checkpoint = 0            * 0: nothing;  1 : save;  2: read
       usedata = 1             # 0: no data (prior); 1: seq like
         nloci = 5             # number of data sets to read in seqfile
     cleandata = 0             # remove sites with ambiguity data (1: yes, 0: no)

     thetaprior = 3 0.002 e    # invgamma(a,b) for theta parameters
       tauprior = 3 0.004      # invgamma(a,b) for root tau

*      heredity = 1 4 4   # (0: no variation, 1: estimate, 2: from file)
*     locusrate = 0 2.0   # (0: no variation, 1: estimate, 2: from file)

       finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MCMC step lengths

         print = 1 0 0 0      # print MCMC samples, locusrate, heredity scalars, gene trees
        burnin = 8000
      sampfreq = 2
       nsample = 100000       # total number of samples to log
```

mcmc settings (chain length, sampling every, etc…)

The setting file

Run the program from a command box (rather than double-clicking the executable) so that you will see the error messages. In the bpp/ folder, run the program by typing the following command:

On WINDOWS	On LINUX/UNIX/MAC OSX
`bin\bpp --cfile bpp.5s.ctl`	`bin/bpp --cfile bpp.5s.ctl`

- Copy the sequence, Imap and setting files in the BPP folder
- Navigate to the bpp folder and run:

```
>bpp.exe --cfile SettingFile.ctl
```

Screen output

```
Initial parameters, np = 3.
Genetrees generated from the MSC density.

lnpG0 =  711.7015  lnL0 = -3350.8540
 -3% 0.70 0.13 0.00 0.22 0.35   4  3 0.0039 0.1455 P(4)=0.9965  -1.0000 0.0007  775.17 -3126.4790
 -2% 0.70 0.31 0.00 0.21 0.28   4  3 0.0100 0.1260 P(4)=0.9925  -1.0000 0.0008  745.60 -3125.8524
 -1% 0.70 0.30 0.00 0.22 0.35   4  3 0.0039 0.1000 P(4)=0.9980  -1.0000 0.0007  750.41 -3126.2339
  0% 0.70 0.30 0.00 0.23 0.25   4  3 0.0077 0.0920 P(4)=0.9845  -1.0000 0.0009  775.04 -3126.2217  0:13
  5% 0.70 0.30 0.00 0.24 0.34   4  3 0.0231 0.0967 P(4)=0.9368  -1.0000 0.0007  750.67 -3125.9098  0:30
 10% 0.70 0.30 0.00 0.24 0.34   4  3 0.0189 0.0954 P(4)=0.9474  -1.0000 0.0007  732.19 -3125.9363  0:46
     ^^ Pjump for MCMC moves ^^     S  p   Prj   PSPR  P(S=4)       theta   tau0   lnpG    E(lnL)
```

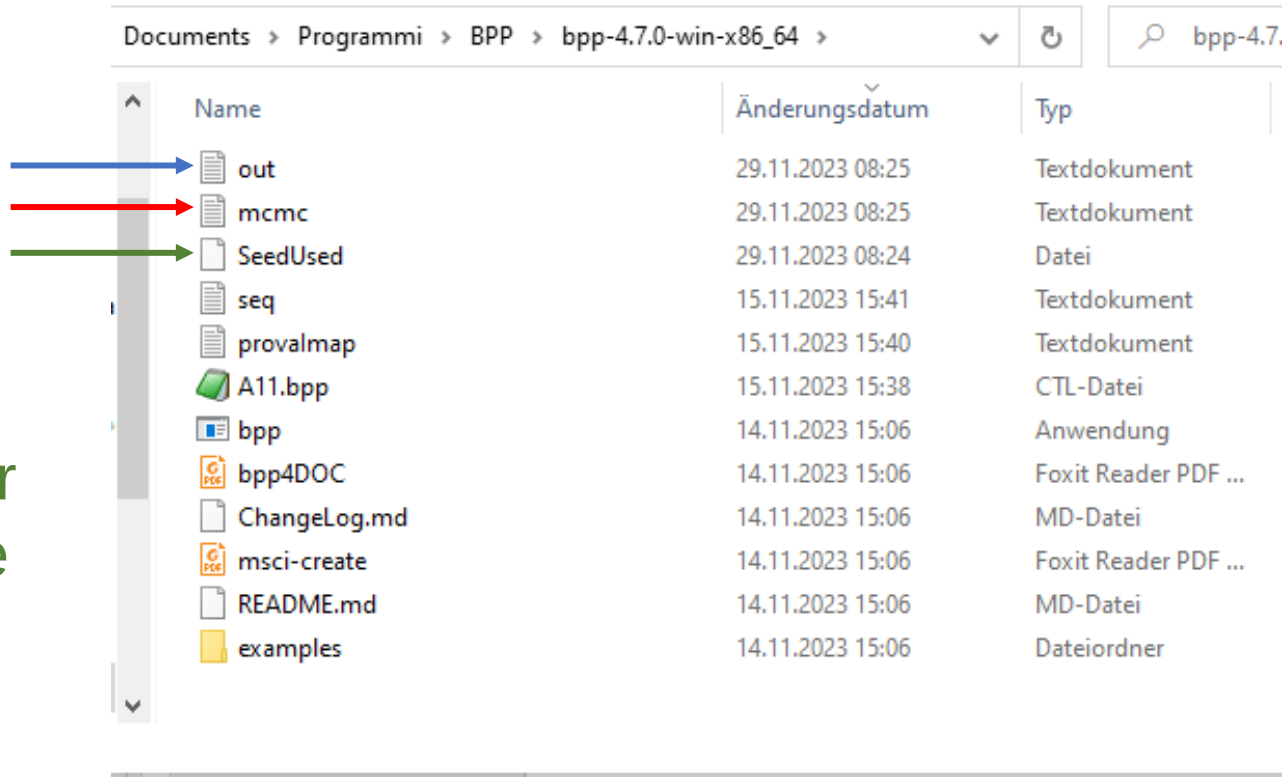
Acceptance proportions of the step lengths for MCMC moves:

Should be values around 0.3, between 0.1 and 0.8. If some of the values is too low, the respective step lengths should be increased in the settings file and vice versa

```
      tauprior = 3 0.004      # invgamma(a,b) for root tau

*      heredity = 1 4 4    # (0: no variation, 1: estimate, 2: from file)
*    locusrate = 0 2.0    # (0: no variation, 1: estimate, 2: from file)

     finetune = 1: .01 .01 .01 .01 .01 .01 .01 .01  # auto (0 or 1): MC

        print = 1 0 0 0      # print MCMC samples, locusrate, heredity
       burnin = 8000         # burn-In
     sampfreq = 2            # frequency of sampling (sample every secc
     nsample = 100000        # total number of samples to log
```

Output files

- output with species delimitation and species tree results
- mcmc file
- seed used (useful if one wants to perform repeat the analyses with another seed (to see if results are consistent)



Output files

- List of models (species delimitation and species tree with the posterior probabilities)

Senza nome | out.txt

```
(A) List of best models (count postP #species SpeciesTree)
16898 0.168980 0.168980 3  (X12X20 X13X29X33 X19X8)  ((X12X20, X13X29X33), X19X8);
14648 0.146480 0.315460 4  (X12 X13X29X33 X19X8 X20)  (((X12, X20), X13X29X33), X19X8);
12391 0.123910 0.439370 2  (X12X13X20X29X33 X19X8)  (X12X13X20X29X33, X19X8);
 7884 0.078840 0.518210 4  (X12X20 X13X29X33 X19 X8)  ((X12X20, X13X29X33), (X19, X8));
 6949 0.069490 0.587700 5  (X12 X13X29X33 X19 X20 X8)  (((X12, X20), X13X29X33), (X19, X8));
 6358 0.063580 0.651280 3  (X12X13X20X29X33 X19 X8)  (X12X13X20X29X33, (X19, X8));
 2351 0.023510 0.674790 4  (X12 X13X29X33 X19X8 X20)  (((X12, X13X29X33), X20), X19X8);
 2224 0.022240 0.697030 4  (X12 X13X29X33 X19X8 X20)  ((X12, (X13X29X33, X20)), X19X8);
 1464 0.014640 0.711670 4  (X12X20 X13X33 X19X8 X29)  ((X12X20, (X13X33, X29)), X19X8);
 1382 0.013820 0.725490 4  (X12X20 X13 X19X8 X29X33)  ((X12X20, (X13, X29X33)), X19X8);
 1337 0.013370 0.738860 4  (X12X20 X13X29 X19X8 X33)  ((X12X20, (X13X29, X33)), X19X8);
 1116 0.011160 0.750020 5  (X12 X13X29 X19X8 X20 X33)  (((X12, X20), (X13X29, X33)), X19X8);
 1065 0.010650 0.760670 5  (X12 X13X29X33 X19 X20 X8)  (((X12, X13X29X33), X20), (X19, X8));
 1052 0.010520 0.771190 5  (X12 X13X29X33 X19 X20 X8)  ((X12, (X13X29X33, X20)), (X19, X8));
 1032 0.010320 0.781510 5  (X12 X13 X19X8 X20 X29X33)  (((X12, X20), (X13, X29X33)), X19X8);
 1019 0.010190 0.791700 5  (X12 X13X33 X19X8 X20 X29)  (((X12, X20), (X13X33, X29)), X19X8);
  891 0.008910 0.800610 3  (X12 X13X20X29X33 X19X8)  ((X12, X13X20X29X33), X19X8);
  809 0.008090 0.808700 3  (X12X13X29X33 X19X8 X20)  ((X12X13X29X33, X20), X19X8);
  701 0.007010 0.815710 5  (X12X20 X13 X19 X29X33 X8)  ((X12X20, (X13, X29X33)), (X19, X8));
  638 0.006380 0.822090 5  (X12X20 X13X29 X19 X33 X8)  ((X12X20, (X13X29, X33)), (X19, X8));
  620 0.006200 0.828290 5  (X12X20 X13X33 X19 X29 X8)  ((X12X20, (X13X33, X29)), (X19, X8));
  524 0.005240 0.833530 4  (X12 X13X20X29X33 X19 X8)  ((X12, X13X20X29X33), (X19, X8));
  466 0.004660 0.838190 3  (X12X20X29X33 X13 X19X8)  ((X12X20X29X33, X13), X19X8);
  461 0.004610 0.842800 6  (X12 X13 X19 X20 X29X33 X8)  (((X12, X20), (X13, X29X33)), (X19, X8));
  447 0.004470 0.847270 6  (X12 X13X33 X19 X20 X29 X8)  (((X12, X20), (X13X33, X29)), (X19, X8));
  428 0.004280 0.851550 6  (X12 X13X29 X19 X20 X33 X8)  (((X12, X20), (X13X29, X33)), (X19, X8));
  426 0.004260 0.855810 4  (X12 X13X29X33 X19X8 X20)  (((X12, X20), X19X8), X13X29X33);
  386 0.003860 0.859670 4  (X12X13X29X33 X19 X20 X8)  ((X12X13X29X33, X20), (X19, X8));
  358 0.003580 0.863250 5  (X12X20 X13 X19X8 X29 X33)  ((X12X20, ((X13, X33), X29)), X19X8);
```

Trova & sostituisci

Output files

- List the best species delimitation models (just species delimitation!)

Posterior probability of the model

out.txt

```
    1 0.000010 1.000000 2  (X12X13X19X20X29X33 X8)  (X12X13X19X20X29X33, X8);

(B) 93 species delimitations & their posterior probabilities
20944 0.209440   4 (X12 X13X29X33 X19X8 X20)
18400 0.184000   3 (X12X20 X13X29X33 X19X8)
11908 0.119080   2 (X12X13X20X29X33 X19X8)
 9804 0.098040   5 (X12 X13X29X33 X19 X20 X8)
 8811 0.088110   4 (X12X20 X13X29X33 X19 X8)
 5915 0.059150   3 (X12X13X20X29X33 X19 X8)
 1725 0.017250   5 (X12 X13X29 X19X8 X20 X33)
 1653 0.016530   5 (X12 X13 X19X8 X20 X29X33)
 1608 0.016080   4 (X12X20 X13 X19X8 X29X33)
 1595 0.015950   4 (X12X20 X13X29 X19X8 X33)
 1494 0.014940   5 (X12 X13X33 X19X8 X20 X29)
 1466 0.014660   6 (X12 X13 X19X8 X20 X29 X33)
 1433 0.014330   4 (X12X20 X13X33 X19X8 X29)
 1190 0.011900   3 (X12 X13X20X29X33 X19X8)
 1110 0.011100   5 (X12X20 X13 X19X8 X29 X33)
  949 0.009490   6 (X12 X13X29 X19 X20 X33 X8)
  838 0.008380   6 (X12 X13X33 X19 X20 X29 X8)
  837 0.008370   6 (X12 X13 X19 X20 X29X33 X8)
  795 0.007950   5 (X12X20 X13X29 X19 X33 X8)
  761 0.007610   7 (X12 X13 X19 X20 X29 X33 X8)
  741 0.007410   5 (X12X20 X13X33 X19 X29 X8)
  701 0.007010   5 (X12X20 X13 X19 X29X33 X8)
  612 0.006120   4 (X12 X13X20X29X33 X19 X8)
  607 0.006070   6 (X12X20 X13 X19 X29 X33 X8)
  416 0.004160   3 (X12X13X29X33 X19X8 X20)
  390 0.003900   3 (X12X20X29X33 X13 X19X8)
  306 0.003060   3 (X12X13X20X33 X19X8 X29)
  207 0.002070   4 (X12X20X29X33 X13 X19 X8)
  206 0.002060   4 (X12X13X29X33 X19 X20 X8)
  183 0.001830   3 (X12X20X33 X13X29 X19X8)
```

model

Output files

- delimited species and their posterior probabilities

Posterior probabilities

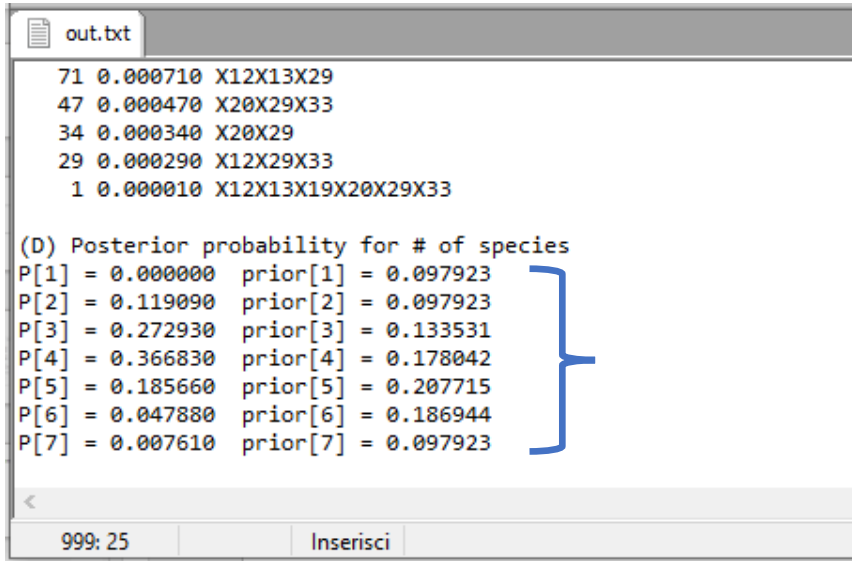
out.txt

```
    2 0.000020   5 (X12X33 X13X20 X19 X29 X8)
    1 0.000010   5 (X12X29 X13X20 X19 X33 X8)
    1 0.000010   2 (X12X13X19X20X29X33 X8)

(C) 34 delimited species & their posterior probabilities
67198 0.671980 X19X8
 7959 0.579590 X13X29X33
 2913 0.429130 X12
 1457 0.414570 X20
35801 0.358010 X12X20
32802 0.328020 X8
32801 0.328010 X19
17823 0.178230 X12X13X20X29X33
 9825 0.098250 X13
 9627 0.096270 X33
 9475 0.094750 X29
 5547 0.055470 X13X29
 5196 0.051960 X29X33
 4757 0.047570 X13X33
 1802 0.018020 X13X20X29X33
  622 0.006220 X12X13X29X33
  597 0.005970 X12X20X29X33
```

species

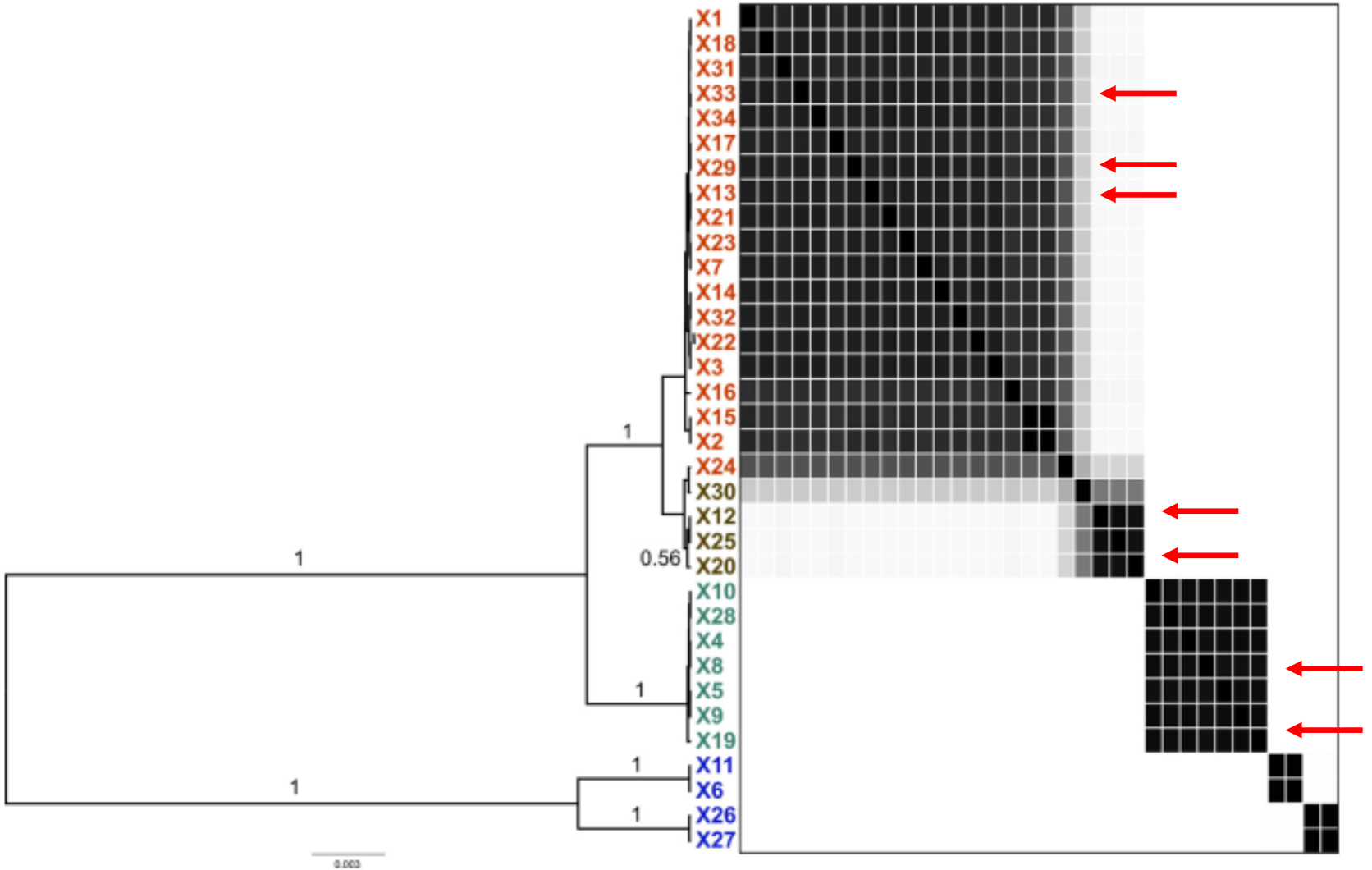
Output files

- posterior probabilities (and their respective priors) of species delimitation models of *n* species



- What come next?
  - Are different analyses, with different starting tree producing the same results?
  - Do different values of influence heavily the results?

- Are results consistent with those in Tomasello (2018)?



Coffe?

- Extension on BPP capable to co-analyse molecular and morphometric data
- https://github.com/cecileane/iBPP/

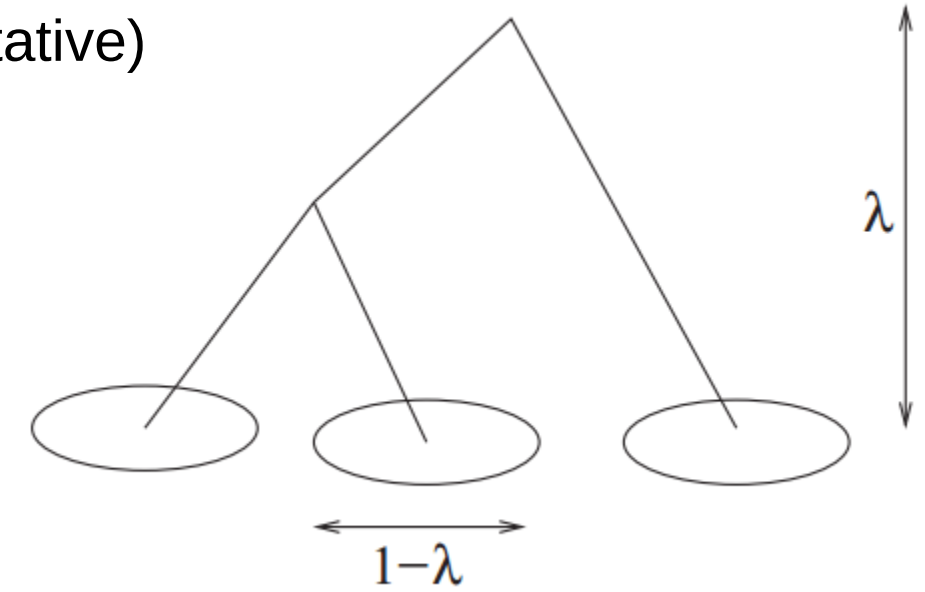
doi:10.1111/evo.12582

Bayesian species delimitation combining multiple genes and traits in a unified framework

Claudia Solís-Lemus,[1] L. Lacey Knowles,[2] and Cécile Ané[1,3,4]

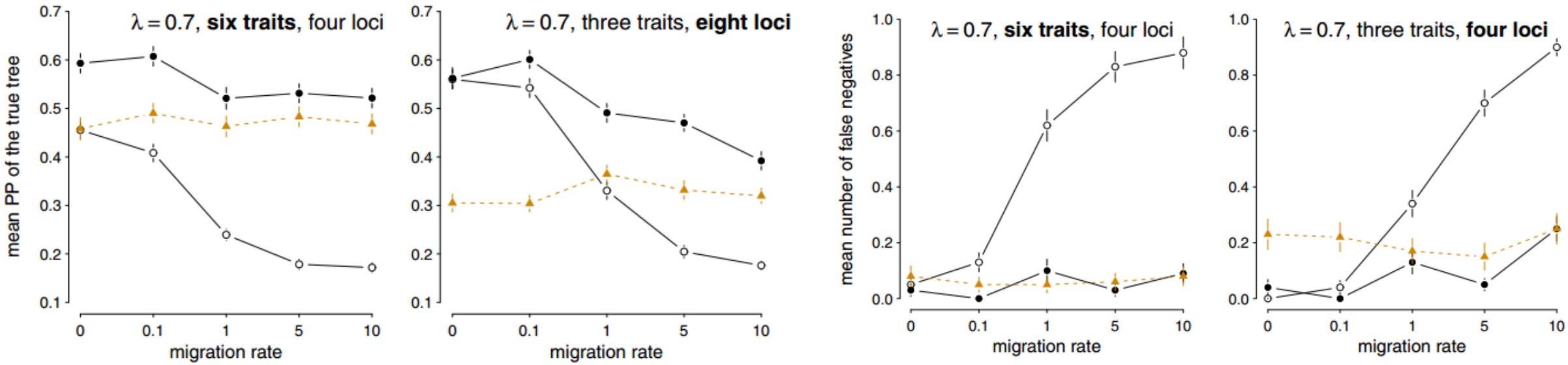
Integrative Bayesian Phylogenetics and Phylo-geography (iBPP)

- Extension of BPP
- Sequence and morphological (continuous quantitative) data as input
- Multiple loci
- Each trait is assumed to have a normal distribution
- A parameter λ models the between-to-within species variance ratio
- Each trait is allowed to have its own λ parameter



Solís-Lemus et al. 2015. *Evolution* 69: 492-507. http://doi.org/10.1111/evo.12582

Integrative Bayesian Phylogenetics and Phylo-geography (iBPP)



Integrating morphological traits increase accuracy!!

Solís-Lemus et al. 2015. *Evolution* 69: 492-507. http://doi.org/10.1111/evo.12582

It uses BPP, basically it works in the same way

but

- It is based on an older version of BPP, layout of inputs and outputs might be a bit different
- Not all model implemented in BPP v4 are supported in iBPP (e.g., MSC with gene flow, etc…)
- The setting files need a few adjustments for the analyses of morphometric data

The following settings need to be added to the setting file:

`traitfile = 5s.morph.txt`	replace with name of file with the trait data
`useseqdata = 1`	0 if sequence data should not be used in the analysis, and 1 if it should
`usetraitdata = 1`	0 if trait data should not be used in the analysis, and 1 if it should
`ntraits = 10`	number of traits in the trait file
`nu0 = 0`	prior parameter for σ^2
`kappa0 = 0`	prior parameter for μ given σ^2

Change also:

- “speciesmodelprior” in the control file of BPP 4.7 with “uniformrootedtrees” in the control of the iBPP
- “useseqdata” instead of “usedata”

Ind	Population	length	width	beaks_length	spines_length	spines	Tomento
X13_	X13	1.9342	0.7334	0.4805	0.32504	6.6	2
X29_	X29	2.166	0.8384	0.5847	0.49068	12.4	3
X33_	X33	1.86875	0.81375	0.493875	0.36105	12.5	5
X8_	X8	0.9154	0.4918	0.2038	0.16696	14.2	1
X12_	X12	1.5122	0.6432	0.3623	0.32412	9.8	0
X19_	X19	1.024	0.6048	0.2703	0.24212	10.6	1
X20_	X20	1.0568	0.5278	0.3459	0.25172	15.2	0

- Bur length

- Bur length
- Bur width

- Bur length
- Bur width
- Beaks’ length

- Bur length
- Bur width
- Beaks’ length
- Spines’ length

- Bur length
- Bur width
- Beaks’ length
- Spines’ length
- Spines / 0.5 cm^2
- Tomentosity

Prior on θ

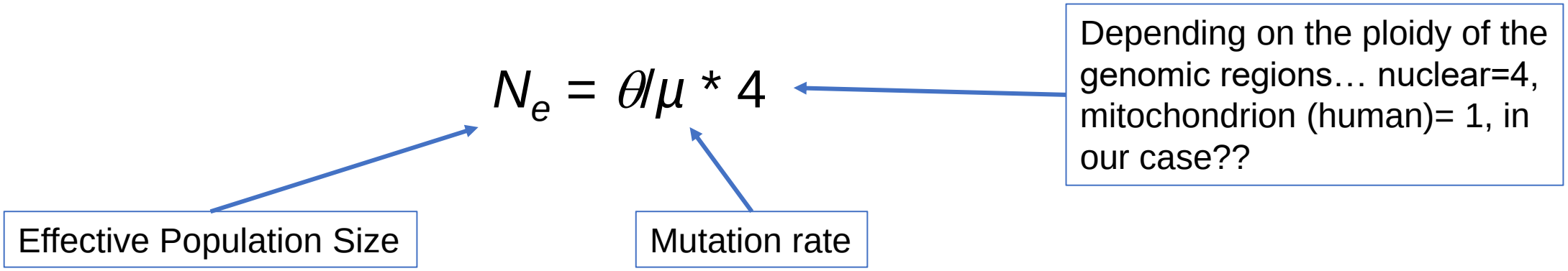
- In older version of BPP a gamma (not inverse gamma!!!) distribution was given to both the θ and τ parameters
- The mean of an gamma distribution G(α, β) is:

$$\alpha / \beta$$

- Say you have an effective population size of 100.000, and a mutation rate of $1*10^{-8}$
- What would you expect for θ?
- giving $\beta = 2$ (for a diffuse prior) what would you give to α in order to center the distribution to the expected value?

Prior on θ

- Perform different analyses with different values of θ, for both big and small effective population size



possible prior values:

- G(2, 50) Large
- G(2, 1000) Small

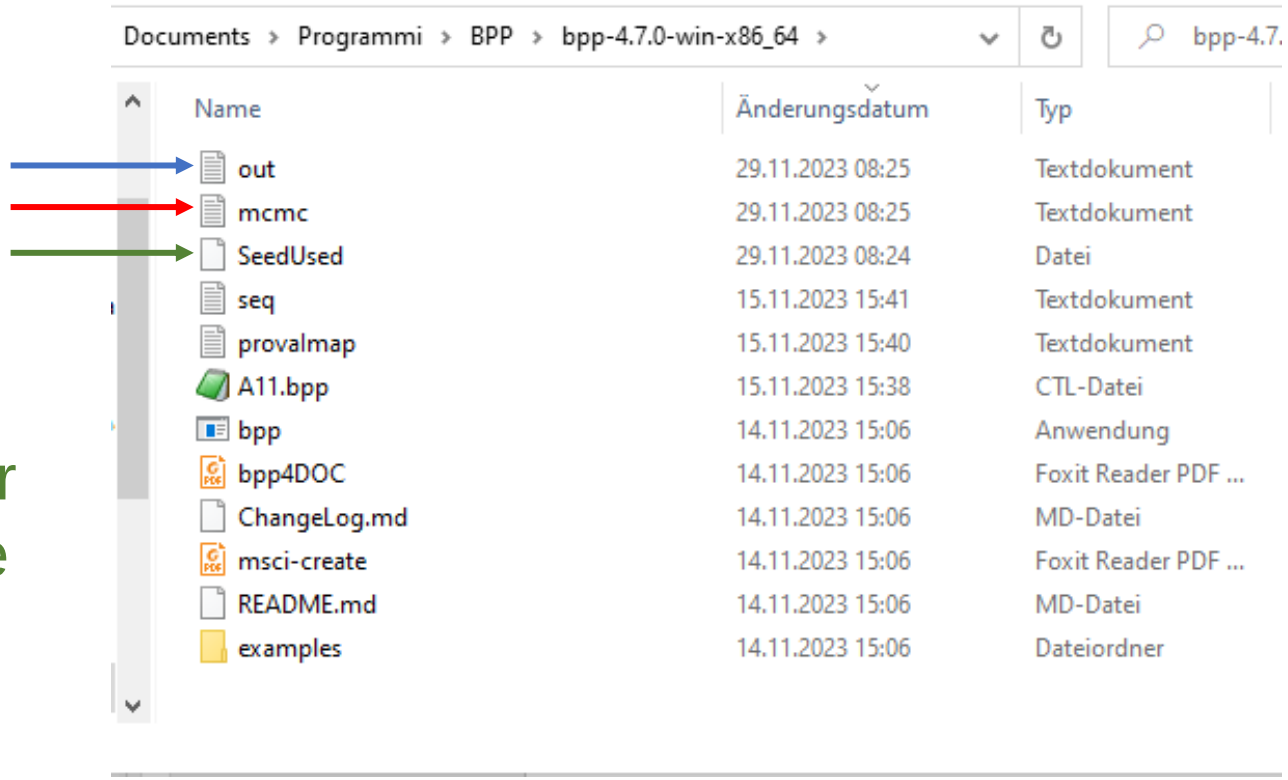
The setting file

- Copy the sequence, Imap and setting files in the iBPP folder
- Navigate to the iBPP folder and run:

```
>ibpp.exe SettingFile.ctl
```

Output files

- output with species delimitation and species tree results
- mcmc file
- seed used (useful if one wants to perform repeat the analyses with another seed (to see if results are consistent)



Output files

weather or not internal nodes are collapsed (0=collapsed, 1=not)

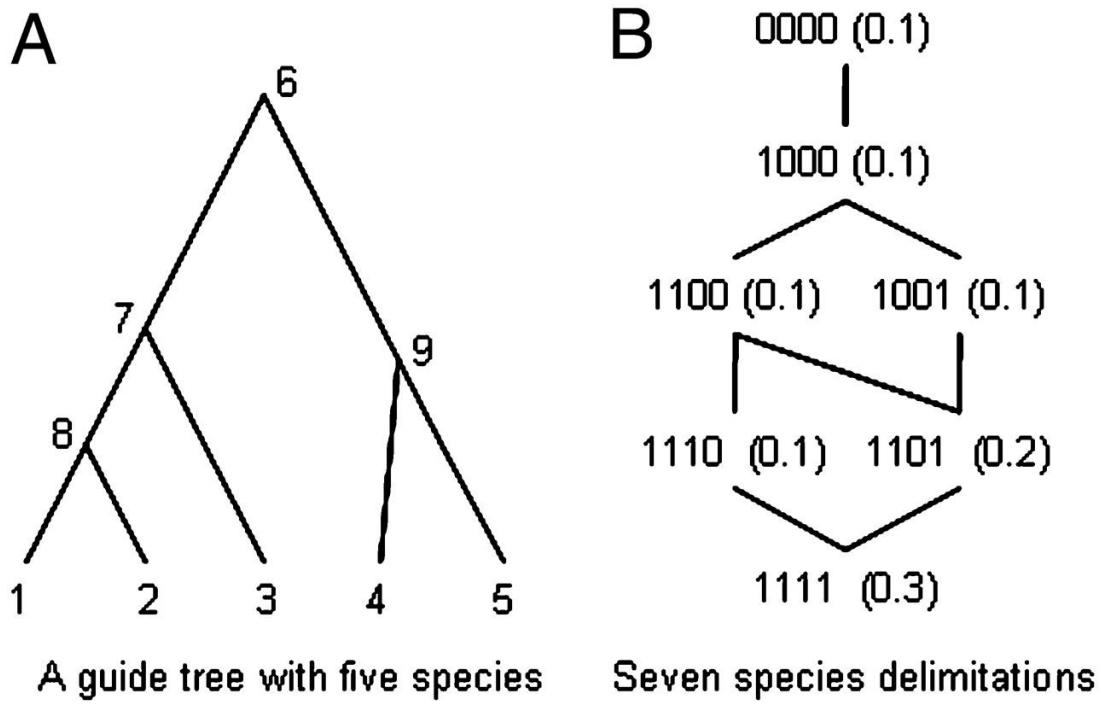
Order in which internal node are listed

troubleShooting.txt | A11.bpp.ctl | morpho.txt | A11.bpp.ctl | 5s.morph.txt | out.txt | provaImap.txt

```
X19   -0.93533 -0.44307 -0.90188 -0.64509 -0.34993 -0.39694
X20   -0.87044  -1.0123  -0.3399 -0.55223   1.2371 -0.95266

Tree frequencies (Ancestral nodes in order:  8 X13X29X33X12X20X8X19  9 X13X29X33X12X20 10 X13X29X33 11 X29X33 12 X12X20 13 X8X19)
   48  110000   58621  0.293105
   49  110001   16832  0.084160
   50  110010   70134  0.350670
   51  110011   19584  0.097920
   56  111000    9973  0.049865
   57  111001    2960  0.014800
   58  111010    9275  0.046375
   59  111011    3329  0.016645
   60  111100    4063  0.020315
   61  111101    1117  0.005585
   62  111110    2972  0.014860
   63  111111    1140  0.005700

Guide tree with posterior probability for presence of nodes
(((X13, (X29, X33)'#0.04646')'#0.17415', (X12, X20)'#0.53217')'#1.00000', (X8, X19)'#0.22481')'#1.00000';

Summarizing the posterior of parameters under the MAP tree 110010
```

Posterior probability of the reconstructed species delimitations

Output files



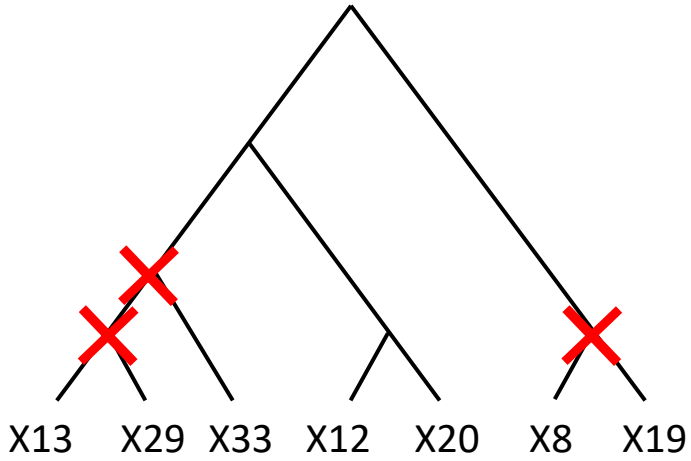
A guide tree with five species

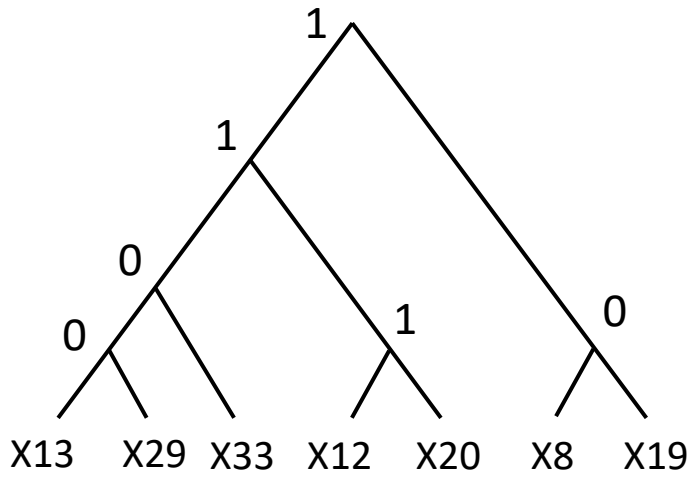
Seven species delimitations

- Given the guide species tree (fixed to the one provided in the setting files
- Each species delimitation is represented by a set of flags indicating whether each of the four ancestral nodes (6, 7, 8, 9) is collapsed (0) or resolved (1)

Yang & Rannala (2010) PNAS 107: 9264-9269

Output files

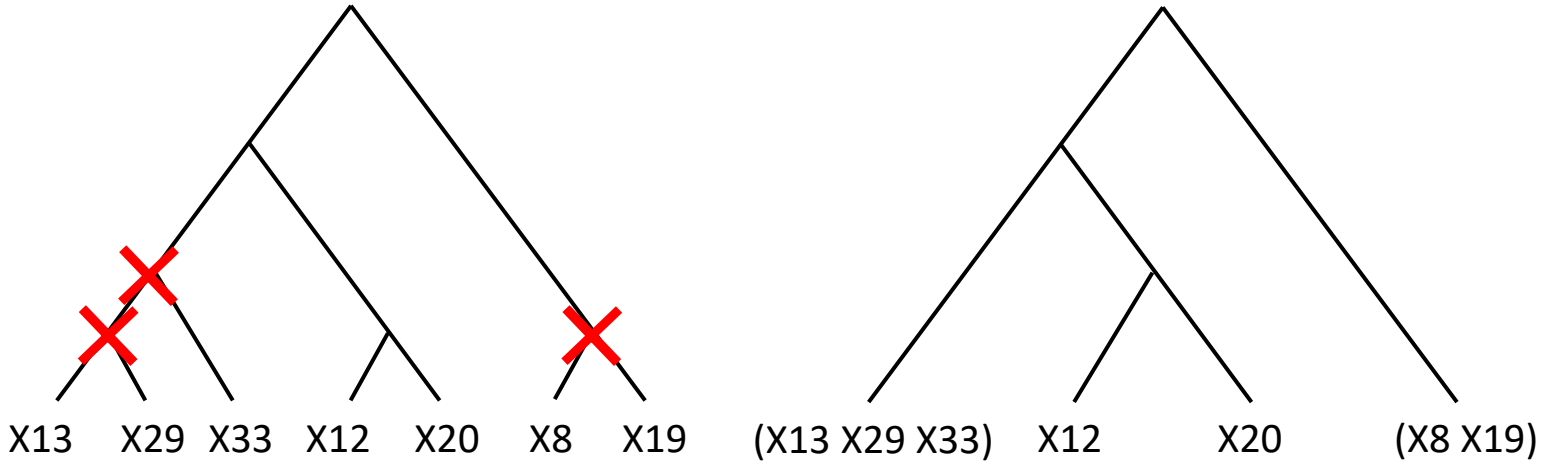




troubleShooting.txt | A11.bpp.ctl | morpho.txt | A11.bpp.ctl | 5s.morph.txt | out.txt | provaImap.txt

```
X19   -0.93533 -0.44307 -0.90188 -0.64509 -0.34993 -0.39694
X20   -0.87044  -1.0123  -0.3399 -0.55223   1.2371 -0.95266

Tree frequencies (Ancestral nodes in order:   8 X13X29X33X12X20X8X19  9 X13X29X33X12X20 10 X13X29X33 11 X29X33 12 X12X20 13 X8X19)
   48  110000    58621  0.293105
   49  110001    16832  0.084160
   50  110010    70134  0.350670
   51  110011    19584  0.097920
   56  111000     9973  0.049865
   57  111001     2960  0.014800
   58  111010     9275  0.046375
   59  111011     3329  0.016645
   60  111100     4063  0.020315
   61  111101     1117  0.005585
   62  111110     2972  0.014860
   63  111111     1140  0.005700

Guide tree with posterior probability for presence of nodes
(((X13, (X29, X33)'#0.04646')'#0.17415', (X12, X20)'#0.53217')'#1.00000', (X8, X19)'#0.22481')'#1.00000';

Summarizing the posterior of parameters under the MAP tree 110010
```

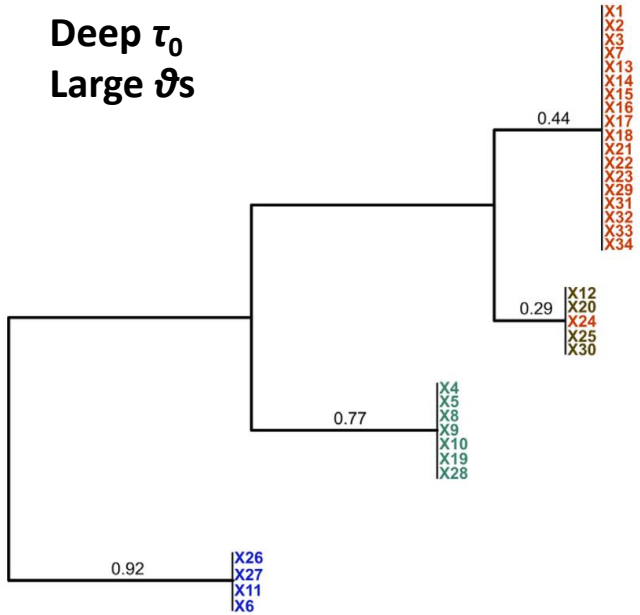
Output files



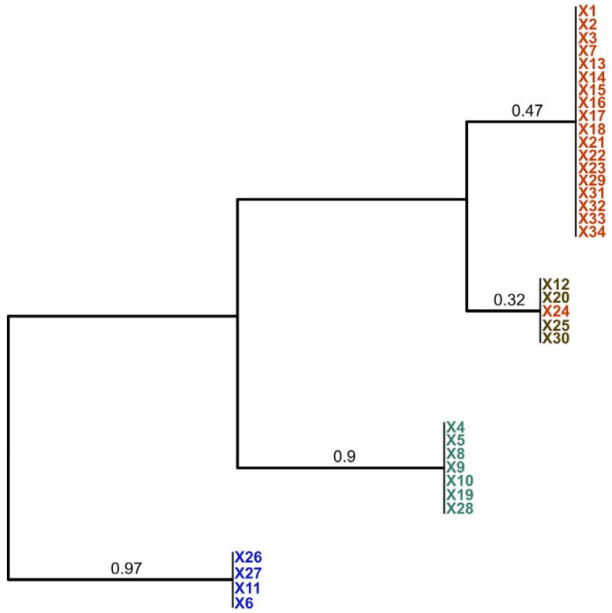
troubleShooting.txt | A11.bpp.ctl | morpho.txt | A11.bpp.ctl | 5s.morph.txt | out.txt | provaImap.txt

```
X19   -0.93533 -0.44307 -0.90188 -0.64509 -0.34993 -0.39694
X20   -0.87044  -1.0123  -0.3399 -0.55223   1.2371 -0.95266

Tree frequencies (Ancestral nodes in order:   8 X13X29X33X12X20X8X19  9 X13X29X33X12X20 10 X13X29X33 11 X29X33 12 X12X20 13 X8X19)
   48  110000    58621  0.293105
   49  110001    16832  0.084160
   50  110010    70134  0.350670
   51  110011    19584  0.097920
   56  111000     9973  0.049865
   57  111001     2960  0.014800
   58  111010     9275  0.046375
   59  111011     3329  0.016645
   60  111100     4063  0.020315
   61  111101     1117  0.005585
   62  111110     2972  0.014860
   63  111111     1140  0.005700

Guide tree with posterior probability for presence of nodes
(((X13, (X29, X33)'#0.04646')'#0.17415', (X12, X20)'#0.53217')'#1.00000', (X8, X19)'#0.22481')'#1.00000';

Summarizing the posterior of parameters under the MAP tree 110010
```

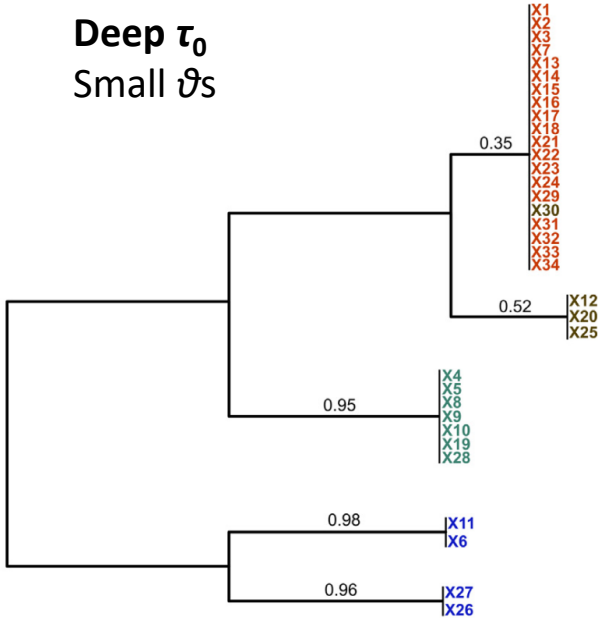
Deep τ_0
Large ϑs



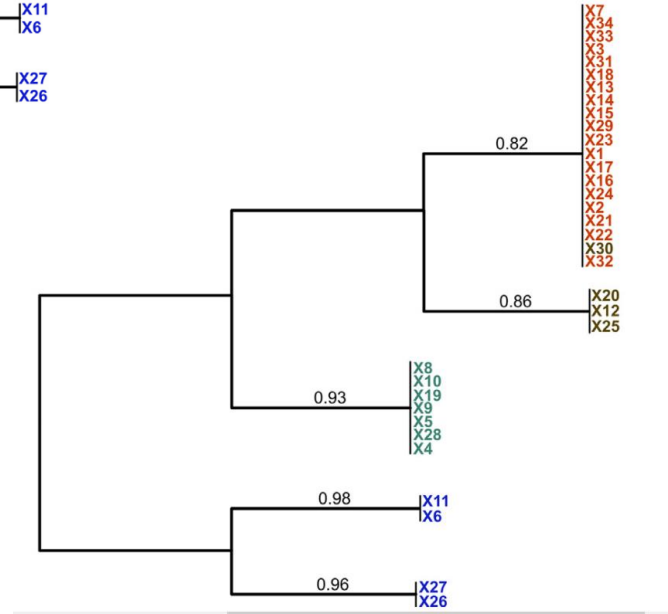
Shallow τ_0
Large ϑs



Deep τ_0
Small ϑs



Shallow τ_0
Small ϑs



Tomasello 2018. *Molecular Phylogenetics and Evolution* 127: 135-145. https://doi.org/10.1016/j.ympev.2018.05.024

contacts:

https://www.uni-goettingen.de/de/staff/185774.html

salvatore.tomasello@uni-goettingen.de