Short manual for programs to be used in workshop "Phylodynamic analysis with BEAST"

BEAUti

Partitions

- Import alignment through File → Import Alignment or use + button in left-bottom corner
- You can import multiple alignments for instance for different genes of the samples.
- Double click alignment to view data
- Use Split to take different codon positions into account (this accommodates the possibility that 3rd codon position can have a higher mutation rate). Use only if you know your codon positions!

Tip Dates

- Use tip dates, when you know the sample times of the alignment
- Use Auto-configure function to extract sampling time from the sample name; several ways available
- The height denotes the time before the most recent sample (that has a height of 0.0)

Site Model

- No site heterogeneity: when Gamma Category Count is set to 0 or 1
- Site heterogeneity: when Gamma Category Count is set to 2 or higher; this is the number of categories in which the gamma distribution for site heterogeneity is discretized (usually 4 categories gives the best trade-off between accuracy and computational effort)
- When using gamma distribution for site heterogeneity, make sure to estimate shape parameter
- Another way to capture site heterogeneity: proportion invariant, toggle to estimate, but set starting value between 0 and 1
- Combination of gamma distributed heterogeneity and invariant proportion often used, but concerns are raised because results are highly interdependent
- Standard available substitution models: JC69, HKY, TN93, GTR; make sure the parameters of these models are estimated
- Frequencies of A, C, G, T nucleotides: "all equal" means each frequency is fixed at 0.25; "empirical" uses the frequencies in the data; "estimated" estimates the frequencies according to the model using the observed frequencies in the data (latter is best option)

Clock model

- Strict clock: all branches evolve at same rate
- Relaxed clock exponential: branches can evolve at different rates governed by exponential distribution
- Relaxed clock lognormal: branches can evolve at different rates governed by a lognormal distribution (more general and usually preferred over exponential)
- Random local clock: evolutionary rate can change at some point(s) in tree. Use only when there is a specific reason to suspect a sudden change in evolutionary rate, such as a change in host species.

Priors

- Only coalescent tree priors used in this workshop:
  - Constant population
  - Exponential population
  - Bayesian Skyline: estimates population size in predefined number of groups/periods
  - Extended Bayesian Skyline: estimates number of groups/periods and the population size in each
- Clock rate: default is a uniform prior between -∞ and ∞, which is highly unlikely. Use a strictly positive prior instead, e.g. lognormal (default parameters suffice for now)
- Population size: default is Jeffrey's prior (1/X) which is shown to be a conservative choice
- Other prior specifications needed depend on your model choices; review each critically

MCMC

- Specify chain length: default is 10 million (which is generally short)
- Tracelog: stores parameter values during analysis; specify filename (which will be saved in same directory as the Beauti xml file) and logging frequency
- Screenlog: writes some parameters to screen to follow the analysis; when specifying a filename it will write to file
- Treelog: stores trees during analysis; specify filename (which will be saved in same directory as the Beauti xml file) and logging frequency
- When using Bayesian Skyline Coalescent: make sure the tracelog and treelog have same logging frequency, as this is needed in the Bayesian Skyline Analyisis (alternative is to subsample logfile post analysis in LogCombiner that accompanies BEAST software).

Initialization panel

- Show by toggling in View → Show Initialization panel
- Usually not needed to inspect, as most initial values can also be set in priors panel
- When using Bayesian Skyline Coalescent: possible to set the dimension of bGroupSizes (number of samples in group) and bPopSizes (population size per group); default is 5 for both (keep them equal for ease of interpretation)

Operators panel

- Show by toggling in View → Show Operators panel
- Shows different types of operators for different parameters and their relative weights (operations with higher weights are performed more frequently)
- Operator parameters are (often) optimized during analysis, and recommendations are given at end of analysis
- In general: default weights are good starting point, and recommendations are given at end of analysis

Save BEAUti specification as .xml file:

- Xml file is the input file for BEAST
- Xml file can be loaded into BEAUti to change your specifications

BEAST2

- Load xml file
- Default: Only write new log files: will give an error if your output files already exist
- Overwrite: will overwrite log files if your output files already exist
- Resume: will append log files to existing output files provided that a .state file is available (that specifies the starting position to resume from)
- Use BEAGLE only when installed (library to speed up calculations, especially when using GPU instead of CPU)
- Run!
- Screenlog shows a.o. sample number, posterior and speed of calculation (important to time your coffee break)
- Close window when finished (you can save the screenlog, but it is not needed for further analysis)

Tracer

- Import trace file (= .log file containing all parameter values in MCMC chain) through File → Import Trace File or use + button below Trace Files on the left
- You can also import trace file while BEAST analysis is still running
- States: length of MCMC chain, but number of samples is chain length divided by log frequency
- Burn-in: default is 10% of number states, but check if this is sufficient in trace plots
- In left panel: all estimated parameters, with mean and ESS
- ESS: effective sample size gives the number of independent samples (more autocorrelation leads to lower ESS)
  - ESS > 200 (black values) recommended
  - 100 < ESS < 200 (orange values) acceptable if it's not a key parameter
  - ESS < 100 (red values) not acceptable
  - Increase ESS by running longer MCMC chain or increasing updating frequency (weights) in operators panel.
- Estimates: summary statistics of a parameter with histogram; when you select multiple parameters (by holding down Shift or Ctrl key) the HPD intervals will be displayed
- Marginal Prob Distribution: probablity density plots of one or more parameters
- Joint-Marginal: two parameters plotted against each other
- Trace: parameter values in MCMC chain; burnin samples are in grey (these are not used for summary statistics), useful to check whether MCMC chain is in equilibrium (i.e. samples represent posterior distribution)
- In Joint-Marginal and Trace: when "sample only" is toggled, the ESS samples are shown, otherwise all samples

- Bayesian Skyline Analysis
  - In Analysis → Bayesian Skyline Analysis
  - Choose Trees Log File from BEAST analysis
  - Choose Stepwise (Constant) variant (linear and exponential change are not supported in standard BEAUti set-up)
  - Check if Population Size, Group Size and Root Height have correct names (corresponding to the parameter names)
  - Maximum time decides how early x-axis starts
  - Give age of youngest tip to have real time on x-axis

TreeAnnotator

- Program to summarize trees logfile in one summary tree
- Burnin percentage: specify based on your findings in the trace plots
- Posterior probability limit: 0.0 means all nodes will be used
- Target tree type: Maximum clade credibility (recommended, tree that maximizes product of posterior probabilities), Maximum sum of clade credibilities (less often used, tree that maximizes sum of posterior probabilities), or User target tree (needs to be specified)
- Node heights: Common ancestor heights (mean of tMRCA of all pairs in clade), Median heights, Mean heights, Keep target heights (uses heights of MCC tree or user target tree)
- Using mean and median heights can lead to negative branch lengths in tree (when uncertainty is large); use common ancestor heights instead to avoid this
- Input Tree File: trees log file produced by BEAST
- Output File: provide an output filename, e.g. with extension .tree
- Run!
- Close window (not needed for further analysis)

FigTree

- Open summary tree file (you can also open trees log file, but it can take some time to load all trees)
- There are a lot of options what and how to display; try them to see what you like. Find some tips&tricks below
- Trees → Order nodes: ladderizes tree from lower left corner
- Color taxa by Filtering (right top corner) on part of name and select color
- Get proper time axis:
  - Time Scale → Scale by factor → Offset by (date of most recent sample)
  - Scale Axis → Reverse axis