StarBEAST2 v1.0.0 tutorial Estimating species trees using StarBEAST2

Huw A. Ogilvie (huw. a. ogilvie@rice. edu) Department of Computer Science, Rice University, Houston, Texas

1 Introduction

StarBEAST2 is a method for multilocus, multispecies coalescent phylogenetic inference (Ogilvie et al. 2017). Multilocus methods like StarBEAST2 take as input one or more multiple sequence alignments derived from genomic loci. There is assumed to be no recombination within each locus and free recombination between loci. Therefore the sequence alignments should be relatively short, and relatively distantly spaced along the genome. Genomic data sets which are typically a good fit to the StarBEAST2 model include RAD-seq loci (Ogilvie et al. 2016) or exons (Scornavacca and Galtier 2017).

StarBEAST2 can be used to estimate species tree topologies and divergence times, the topologies and coalescence times of gene trees, the substitution model rates and base frequencies for those gene trees, per-species population sizes and per-species molecular clock rates.

Newer versions of StarBEAST2 supports the integration of molecular sequences and morphological characters for so-called "total evidence" analyses, and the use of fossil data for tip-dating (**Ogilvie2021**). Fossil data can include ancient DNA, morphological characters, and the estimated ages of fossils. Tip-dating calibrates the trees in absolute time, typically millions of years.

This tutorial will cover using StarBEAST2 to estimate a species tree including per-species population sizes of the genus *Canis* from exon sequences (Section 3), to estimate per-species molecular clock rates (Section 4), and to conduct a total evidence tip-dating study (Section 5).

2 Preliminaries

Before doing anything, we need to install StarBEAST2, which is available as a package for the phylogenetic software platform BEAST2. This tutorial is for version 2.7 of the platform, which you can download from http://beast2.org. Make sure you have the latest update to BEAST2 version 2.7 before proceeding.

After downloading and installing BEAST2 on your computer, open BEAUti — the graphical user interface for BEAST2. To install StarBEAST2 (or any BEAST2 package), open the File menu and select Manage Packages (Figure 1).



Figure 1: Opening the package manager

Select StarBEAST2 from the list of available packages and click the install button, which will take care of the installation for you (Figure 2). After the installation of StarBEAST2 is finished, just like for any BEAST2 package, you **must** quit and relaunch BEAUti.

	BEAST 2 Package Manager 📃 🗆 🗙										
List of available package	s for BEAST v2.7.*										
Name	Install L	atest	Dependen	cies	Link	Implementation of the MSRD multi-type birth-death tree					
MultiTypeTree	8	0.0	BEAST base BEAST	ann	c-9	Structured coalescent inference					
NS	1	2.1	BEAST ann BEAST h	ase MOD		Nested sampling for model selection and posterior infere					
	1	11	BEAST app, bModel	Test RFA		OBAMA for Bayesian Amino-acid Model Averaging					
ORC	112 1	1.2	BEAST app, BRIGGEI	abs BEAS		Optimised Relayed Clock model					
PhyDyn	1.1.2	4.0	BEAST app, BEAST h	ase	c-3	PhyDyn: Epidemiological modelling with BEAST					
phylodynamics	1	4.0	BDSKY BEAST base	Juse	6-3	BDSIR and Stochastic Coalescent					
phylonco	1	.0.1	BEASTLabs, BEAST	app. BEAS	6-3	A package for single cell cancer evolution					
PIOMEE	1	0.0	BDSKY BEAST base	BEAST app	6-3	Birth-death skyline-based method efficiently dealing with					
PoMo	1	.1.1	BEAST.base. BEAST.	арр	c-9	PoMo, a substitution model that separates mutation and					
Recombination	1.	.0.1	BEAST.app. BEAST.b	ase	c-3	Inference of Recombination networks					
SA	2.1.1 2	.1.1	BEASTLabs, BEAST.	pase, BEA	c-3	Sampled ancestor trees					
SCOTTI	3.	.0.0	BEAST.base, BEAST.	арр	6-3	Structured COalescent Transmission Tree Inference					
SNAPP	1	.6.1	BEAST.base, BEASTI	Labs, BEA	6-3	SNP and AFLP Phylogenies					
snapper	1.	.1.3	BEAST.base, BEAST.	app, SNAPP	6-3	Diffusion based SNP and AFLP Phylogenies					
SpeciesNetwork	1.	.0.0	BEAST.app, BEAST.b	base	6-3	Multispecies network coalescent (MSNC) inference of intr					
speedemon	1.	.1.0	BICEPS, BEAST.app,	BEASTLa	6-3	Fast species delimitation under the multispecies coalesce					
SSM	1.	.2.0	BEAST.app, BEAST.b	base	6-3	Standard Nucleotide Substitution Models					
STACEY	1.	.3.1	BEAST.app, BEAST.b	oase	6-3	Species delimitation and species tree estimation					
StarBEAST2	1.	.0.0	BEAST.app, MM, BE	AST.base,	c-3	Multispecies coalescent inference using multi-locus and f					
starbeast3	1.1.7 1	.1.7	BEAST.base, SA, BEA	AST.app, B	c-3	StarBeast3 multispecies coalescent using advanced MCM					
timtam	0.	.4.0	BEAST.base, BEAST.	арр	6-3	Distribution for birth-death models with occurrence data					
TreeStat2	0	.1.1	BEAST.base		6-9	Utility for calculating tree statistics from tree log file					
< [>					
✓ Latest	Install/Upgrade		Uninstall	Package rej	oositor	ories ? Close					

Figure 2: Installing StarBEAST2

After relaunching BEAUti, you will notice four new templates (Figure 3). The first is "StarBeast2", which is for a strict molecular clock or relaxed gene tree clocks uncorrelated with the species tree lineages. The others are "SpeciesTreeUCLN", "SpeciesTreeRLC" and "SpeciesTreeUCED" which enable per-species clock rates, as described by Ogilvie et al. 2016. The relaxed clock models used by those templates are uncorrelated log-normal (UCLN), relaxed local clock (RLC), and uncorrelated exponential distribution (UCED) respectively.

To visualize posterior distributions of trees, we will be using DensiTree (Bouckaert 2010) which is included with BEAST2. We will also use the web application IcyTree to view summary trees, which is available at http://icytree.org/. To analyze continuous parameters we will use Tracer, which can be downloaded from https://beast.community/tracer.

3 Reconstructing the species tree of *Canis*

The genus *Canis* includes species such as *Canis lupus* (wolves), *Canis latrans* (coyotes), and *Canis dirus* (dire wolves, extinct). Because the taxonomy of this genus is quite messy, it also includes *Cuon alpinus* and *Lycaon pictus*. In this section we will reconstruct the species tree of *Canis* using StarBEAST2.

				BE	AUti 2: St	andard		- • ×
File Mode View Help)	_						
D New	Ctrl+N	el C	lock Model	Priors	ИСМС			
🗃 Load	Ctrl+O	Link	Clock Models	Unlink Clo	k Models	Link Trees	Unlink Trees	
Add Morphological Data	1	Sites	Data Typ	e Site Mod	el Clock l	Mo Tree	e Ambiguit	
Import Alignment								
Template	•	Standa	rd					
Set working dir	•	StarBe	st2					
Manage Packages		Species	TreeUCED					
Clear Class Path		Species	TreeRLC					
Launch Apps		AIM						
Save	Ctrl+S	Species	TreeUCLN					
Save As		StarBe	ast3					
Close	Ctrl+W	Other ⁻	emplate					
Exit	Ctrl+Q							
				No.pa	rtitions los	idad vat		
				No pa	10013108	lueu yet		

Figure 3: Selecting a StarBEAST2 template

3.1 Importing alignments

First up, create a new folder somewhere with a sensible name like "CanisPhylogeny". Relaunch BEAUti and load up the strict clock template by selecting "StarBeast2" from the Templates submenu of the File menu (Figure 3). Now we will import multiple sequence alignments of the 16 nuclear loci sequenced by Lindblad-Toh et al. 2005. Select Import Alignments from the File menu, and navigate to the "data" subfolder of the tutorial. Select all 16 FASTA files, and click OK to import (Figure 4). Do not select "morphology-canis.nexus" which is for the total evidence study.

After you click OK, you will be asked what type of sequences are in the files. As all 16 files are MSAs of nucleotide sequences, select "all are nucleotide" and click OK again to continue. There should now be 16 partitions listed in BEAUti.

3.2 Linking models

As a general rule, clock models should be linked when using the strict clock StarBEAST2 template. Rates between loci are still allowed to vary and the mean rate among all sites is fixed at 1. Therefore by linking the clocks, the average clock rate can be fixed at an *a priori* value to calibrate the analysis, or estimated when node-dating or tip-dating is used. Select all 16 loci and click the "Link Clock Models" button. Now all loci should have "APOBS1" as the clock model name. The site model, clock model and tree names are text fields that you can click on to edit. Change the name of the clock model to "molecular" and it should

ink Site Moo	dels Unlin	k Site Mode	els Link Cloc	k Models	Unlink Clock N	Andels Link	Trees			
Name	File	Taxa		K MOUCIS	OTHINK CIOCK P			Inlink Irees		
Name	File	Taxa								
		IdXd	Sites	Data Type	Site Model	Clock Mo	Tree	Ambiguit		
		Cano	cel		Load Alig	gnment File		٩	Open	
		ං Hor	me	< 🌣 huey	IdeaProjects	BEAST2 st	arbeast2	tutorial data	>	
		Des .	sktop	Name		~ Size	Туре	Mod	lified	
		Do	cuments	APOBS1.	fasta	11.7 kB	Text	Yest	erday	
		Dov		APOBS2.		10.6 kB			erday	
		Ric		BDNF.fas		8.3 kB			erday	
			eos	BRCA1S1	1.fasta	11.7 kB			erday	
		□ bea	ist	BRCA1S2	2.fasta	12.3 kB			erday	
				Ch14.fas	ta	15.2 kB	Text		erday	
				Ch21.fas	ta	10.1 kB	Text	Yest	erday	
				Ch24.fas	ta	12.1 kB	Text		erday	
				CHST12.	fasta	11.7 kB	Text	Yest	erday	
				CMKOR1	.fasta	12.2 kB	Text	Yest	erday	
				FGFR3.fa		8.5 kB	Text	Yest	erday	
				GHRex09	9.fasta	12.2 kB	Text	Yest	erday	
				morphol	ogy_canis.ne:	x 9.0 kB	Text	Yest	erday	
				RAG1.ras	sta	12.3 kB	lext	Yest	erday	
				tip_dates	s_canis.txt	577 byte	s lext	Yest	erday	
					.rasta	10.3 kB	Text	Yest	erday	
						9.2 KB		Yest	erday ocday	
				• vvvF.rast	d	12.2 KB	Text	Yest	erday	

Figure 4: Importing *Canis* multiple sequence alignments

look like Figure 5.

Mixing nuclear and mitochondrial loci

StarBEAST2 can be used with a combination of mitochondrial and nuclear loci, but this requires careful thought when specifying the model. One possibility is to link the nuclear loci as above, but not to link the mitochondrial loci. Then when editing the site models, untick "estimate" for the substitution rate of each mitochondrial locus. Now separate clock rates will be used for each mitochondrial locus, in addition to the overall clock rate which applies to nuclear loci.

If the species tree is calibrated using an *a priori* nuclear or mitochondrial molecular clock rate, you can set that rate in the Clock Model panel. If the mitochondrial rate is fixed, the nuclear rate should probably be estimated and *vice versa*. If the species tree is calibrated using node or tip dating, then all clock rates should probably be estimated.

Make sure appropriate priors are used for all estimated clock rates. This could be a 1/X prior for rates where you genuinely have no prior knowledge. However in most circumstances you will have a general idea of the clock rate, e.g. within an order of magnitude. In that case, you can set a broad prior with a mean based on your prior knowledge. Broad priors include Log Normal with a large standard deviation (up to about 2), or a Gamma distribution with a small α (set "alpha" to 1 or 2).

						BEAUt	i 2: StarBeast2					- • ×
File N	1ode View	/ Help										
Partition	s Taxon se	ts Tip D	Dates G	iene Ploidy F	opulation M	odel Site Mo	del Clock Model	Priors	мсмс			
Link Site	Models	Jnlink Sit	e Mode	Link Cloc	k Models	Unlink Clock N	Iodels Link Tre	es Unli	ink Trees			Â
Name	File	Таха	Sites	Data Type	Site	Model	Clock Mo	del	Tree		Ambiguities	
APOBS1	APOBS1		702	nucleotide	APOBS1	-	molecular	-	APOBS1	-		
APOBS2	APOBS2	16	633	nucleotide	APOBS2	•	molecular	•	APOBS2	•		
BDNF	BDNF	16	489	nucleotide	BDNF	-	molecular	-	BDNF	-		
BRCA1S1	BRCA1S1	16	705	nucleotide	BRCA1S1	-	molecular	-	BRCA1S1	-		
BRCA1S2	BRCA1S2	16	741	nucleotide	BRCA1S2	-	molecular	-	BRCA1S2	-		
Ch14	Ch14	16	921	nucleotide	Ch14	-	molecular	-	Ch14	-		
Ch21	Ch21	16	601	nucleotide	Ch21	-	molecular	-	Ch21	-		
Ch24	Ch24	16	730	nucleotide	Ch24	-	molecular	-	Ch24	-		
CHST12	CHST12	16	705	nucleotide	CHST12	-	molecular	-	CHST12	•		
CMKOR1	CMKOR1	16	735	nucleotide	CMKOR1	-	molecular	-	CMKOR1	-		
FGFR3	FGFR3	16	503	nucleotide	FGFR3	-	molecular	-	FGFR3	-		
GHRex09	GHRex09	16	736	nucleotide	GHRex09	•	molecular	-	GHRex09	-		
RAG1	RAG1	16	741	nucleotide	RAG1	•	molecular	-	RAG1	-		
TMEM20	TMEM20	16	615	nucleotide	TMEM20	•	molecular	-	TMEM20	-		
VANGL2	VANGL2	16	546	nucleotide	VANGL2	-	molecular	•	VANGL2	•		
VWF	VWF	16	732	nucleotide	VWF	•	molecular	-	VWF	-		
												> ×

Figure 5: Linking clock models

Linking site models

When clock models are linked as is recommended by this tutorial, linking site models will result in every gene having exactly the same clock rate. As this is unrealistic, generally site models should **not** be linked using BEAUti when setting up a StarBEAST2 analysis.

3.3 Specifying species names

In the FASTA files, each sequence has a name like "Canis_anthus_a". Here *Canis anthus* is the binomial name, and "a" is a haplotype. For the data supplied with this tutorial, each species has two haplotypes "a" and "b", both from the same diploid individual. However in other studies multiple individuals may be sequenced, so an arbitrary number of haplotypes are available per species.

To assign haplotypes to species, select the Taxon Sets tab in BEAUti, then click the Guess button. To assign the correct names, keep "use everything" selected, but change "after first" to "before last". Leave the underscore in the text box and click OK (Figure 6). This way everything before the last underscore in the haplotype name will be used as the species name, so "Canis_anthus" will be the species name for "Canis_anthus_a".

Now each haplotype (Taxon) should have a corresponding species (Species/Population). Based on how we assigned the species names, there should be two haplotypes "a" and "b" for each species (Figure 7).

		BEAUti 2: 1	StarBeast2		-	- ×
File Mode View Help						
Partitions Taxon sets Tip Dates	Gene Ploidy Population Mod	el Site Model	Clock Model	Priors	МСМС	
filter:						
					Current of (Densulation	
Canis latrans h	Taxon		c		Species/Population	-
Canis simensis a			c			
Canis simensis b						
Canis latrans a					- <u> </u>	
Canis_mesomelas_b						
Cuon_alpinus_b	use everything	oefore last	· _			
Cuon_alpinus_a						
Canis_mesomelas_a	🔿 enlit en skavester		a 10 d	taka aw		
Canis_lupus_b	Split on character		anu	take gro	oup(s):	
Canis_lupus_a						
Canis_anthus_a	Ouse regular expression					
Canis_adustus_a						
Canis_anthus_b	read from file			Bre	rowse ?	
Canis_adustus_b						
Lycaon_pictus_b						
Lycoon nictus a						`
Fill down Guess						
					Cancel	
< [

Figure 6: Guessing species names from haplotype names

3.4 Gene Ploidy and Population Model

Ignore the Tip Dates tab, which is for tip-dating and will be covered in section 5 of the tutorial. Open the Gene Ploidy tab, and observe that the default value for all loci is 2.0. This is because there are two copies of a locus in each individual for diploid populations, so we scale the effective population sizes by 2.0. If you use any mitochondrial or Y/W chromosomal loci, you should change their ploidy to 0.5, because there is effectively only half as many copies of those molecules as there are individuals (male mitochondrial genomes do not count, as they are not inherited by offspring). The ploidy of X/Z chromosomal loci should be set to 1.5 for the same reason.

Select the Population Model tab. By default analytical integration is used, which is slightly faster but does not produce estimates of per-species population sizes. Change the model to "Constant Populations", which will add effective population sizes to the species tree output (Figure 8).

anis_mesomelas_b	Taxon	Species/Population	
anis_mesomelas_b anis_mesomelas_a	Taxon	Species/Population	
anis_mesomelas_b anis_mesomelas_a		Canis_mesomelas	
anis_mesomelas_a			
		Canis_mesomelas	
anis_anthus_a		Canis_anthus	
anis_anthus_b		Canis_anthus	
anis_lupus_b		Canis_lupus	
anis_lupus_a		Canis_lupus	
anis_latrans_b		Canis_latrans	
anis_latrans_a		Canis_latrans	
ycaon_pictus_b		Lycaon_pictus	
ycaon_pictus_a		Lycaon_pictus	
uon_alpinus_b		Cuon_alpinus	
uon_alpinus_a		Cuon_alpinus	
anis_simensis_a		Canis_simensis	
anis_simensis_b		Canis_simensis	
anis_adustus_a		Canis_adustus	
anic aductue h		Capie aductue	

Figure 7: The assignment of haplotypes to species

					BEAUti 2: :	StarBeast2			
File Mo	de View	Help							
artitions	Taxon sets	Tip Dates	Gene Ploidy	Population Model	Site Model	Clock Model	Priors	мсмс	
Analytical	Population S	Size Integrat	tion 🔻						
Analytical	Population S	ize Integrat	ion						
Constant I	Populations								
Linear Wit	th Constant R	loot Populat	ions						
Uniform P	opulations								

Figure 8: Changing the population model

3.5 Site Model

Select the Site Model tab, and you will see the site model for the first partition displayed. Change "JC69" to "HKY", and then set the frequencies to empirical (Figure 9). HKY is more flexible because it allows nucleotide transitions to have a different rate relative to transversions (Hasegawa et al. 1985).

		BEAUti 2: StarBeast2		- • ×
File Mode View Help				
Partitions Taxon sets Tip Date	es Gene Ploidy Population Mode	I Site Model Clock Model Priors	MCMC	
Partition	Gamma Site Model 💌 🕐			î
APOBS1	Substitution Pate	2 Q v ostimate		
APOBS2	Substitution Nate			
BDNF	Gamma Category Count	0		
BRCA1S1				
BRCATS2	Proportion Invariant	0.0 0.0 estimate		
Ch21	Subst Model	НКҮ 👻 🕖		
Ch24		Kanna	2 e V estimate	
CHST12		2		
CMKOR1		Frequencies	stimated 👻 🕐 💿	
FGFR3		E	stimated	
GHRex09		E	mpirical N	
RAG1		A	ll Equal	
TMEM20				
VANGL2				
VWF				
	<			> ×

Figure 9: Setting the site model to HKY with empirical base frequencies

Now to change the site model for *all* loci to HKY with empirical base frequencies, select all of the partitions in the left hand column with the shift key. Choose to clone the site model you just configured (presumably for APOBS1), then click "OK" to apply the same site model used for the first locus to everything else (Figure 10).

In a more serious analysis, you might consider setting the number of gamma categories to 4. This allows for different sites to evolve at different rates, an obviously more realistic model (Yang 1994). However the phylogenetic likelihood must be calculated once for each category, so 4 categories will be $4 \times$ slower. That likelihood calculation is a major part of the StarBEAST2 algorithm, so more gamma rate categories will require more computer time.

3.6 Clock Model

Hugall et al. 2007 estimated that the molecular clock rate for the RAG-1 nuclear coding gene in mammals is approximately 10^{-3} substitutions per site per million years. While we don't know the average rate across all loci in our data within *Canis*, we can use it as a **very approximate** calibration. Open the clock model

								BE	EAUti 2:	StarBeast2					×
File N	Node	View	Help												
Partition	ns Tax	on sets	Tip D	ates	Gene Pl	oidy	Population Model	Sit	e Model	Clock Mode	Priors	мсмс			
Partition				Clo	ne from	APO	BS1	-	ОК						
APOBS1			ŕ												
APOBS2															
BDNF															
BRCA1S1															
BRCA152	2														
Ch14															
Ch21															
Ch24															
CHST12															
CMKOR1															
FGFR3															
GHRex09	9														
RAG1															
TMEM20															
VANGLZ															
VVVF			~												

Figure 10: Cloning the site model

panel and set the rate to "0.001", to match the *a priori* estimate (Figure 11).

You can also use the scientific notation shorthand, 1e-3, which is equivalent to 0.001.

3.7 Priors and MCMC

Open the Priors panel and change "Yule Model" to "Birth Death Model". These models are identical, except the birth-death model allows for a non-zero rate of extinction.

StarBEAST2 is an MCMC method. These kind of methods do better at estimating the posterior distributions (of trees or other parameters) the longer they are run, although after a point there are diminishing returns. The default chain length in StarBEAST2 is 10 million states, but for this analysis we need a bit more for good estimates; change the chain length to 40 million (Figure 12)

Save your model as an XML file – in the folder you created before importing alignments – by clicking Save As in the File menu. Navigate to the folder and give your XML file a name like "CanisPhylogeny.xml".

3.8 Running BEAST

You can run BEAST from the command line or using the GUI. To run your XML from the command line, first navigate to the folder you saved the XML file into. Then run "/path/to/beast/bin/beast CanisPhylogeny.xml". Make sure to change /path/to/beast to match the folder where BEAST is installed on your

			BEAUti 2: S	StarBeast2			9	
File Mode View Hel	þ							
Partitions Taxon sets Tip	Dates Gene Ploidy	Population Model	Site Model	Clock Model	Priors	MCMC		
Strict Clock	•							
Clock.rate	0.001	🤊 💿 📄 estimate	2					
	•							

Figure 11: Using an *a priori* clock rate for calibration

computer.

BEAST should take about 5 to 10 minutes to finish the chain. Sampled statistics and various parameters will be saved to "starbeast.log", species trees to "species.trees", and separate gene tree files will be created for each locus. The command line output should start off looking something like what follows:

```
BEAST v2.7.5 Prerelease, 2002-2023
Bayesian Evolutionary Analysis Sampling Trees
Designed and developed by
Remco Bouckaert, Alexei J. Drummond, Andrew Rambaut & Marc A. Suchard
Centre for Computational Evolution
University of Auckland
r.bouckaert@auckland.ac.nz
alexei@cs.auckland.ac.nz
Institute of Evolutionary Biology
University of Edinburgh
a.rambaut@ed.ac.uk
David Geffen School of Medicine
```

			BEAUti 2: S	tarBeast2			۰	×
File Mode View Help								
Partitions Taxon sets Tip Date	s Gene Ploidy	Population Model	Site Model	Clock Model	Priors	MCMC		
Chain Length	40000000	I						Î
Store Every	5000	0						
Pre Burnin	0	0						
Num Initialization Attempts	10							
t racelog								
speciesTreeLogger								
screenlog								
treelog.t:APOBS1								
treelog.t:Ch14								
treelog.t:RAG1								
treelog.t:TMEM20								
treelog.t:Ch21								
treelog.t:BRCA1S2								
treelog.t:FGFR3								
treelog.t:BDNF								
treelog.t:BRCA1S1								
treelog.t:CHST12								~

Figure 12: Setting a longer MCMC chain length of 40 million iterations

University of California, Los Angeles msuchard@ucla.edu

> Downloads, Help & Resources: http://beast2.org/

Source code distributed under the GNU Lesser General Public License: http://github.com/CompEvol/beast2

BEAST developers:

Alex Alekseyenko, Trevor Bedford, Erik Bloomquist, Joseph Heled, Sebastian Hoehna, Denise Kuehnert, Philippe Lemey, Wai Lok Sibon Li, Gerton Lunter, Sidney Markowitz, Vladimir Minin, Michael Defoin Platel, Oliver Pybus, Tim Vaughan, Chieh-Hsi Wu, Walter Xie

> Thanks to: Roald Forsberg, Beth Shapiro and Korbinian Strimmer

Random number seed: 1691867441007

File: Canis-Phylogeny.xml seed: 1691867441007 threads: 1

Loading package BEAST.app v2.7.5 SA v2.1.1 MM v1.2.1 ORC v1.1.2 starbeast2 v1.0.0 BEAST.base v2.7.5 BEASTLabs v2.0.1 starbeast3 v1.1.7

At the end of the output will be statistics describing the performance of all the MCMC operators. This will look something like below, except I needed to replace some of the output with ellipses so it would fit in this document:

Operator	• • •	Tuning	#accept	#rejec
 beast.base.inference.operator.UpDownOperator(clockUpDownOperator		0.80611	18039	5211
ScaleOperator(TreeScaler.t:Ch14)		0.81774	18037	5174
ScaleOperator(TreeRootScaler.t:Ch14)		0.52148	14234	5571
Uniform(UniformOperator.t:Ch14)		-	121376	22855
SubtreeSlide(SubtreeSlide.t:Ch14)		0.65112	117291	23259
Exchange(Narrow.t:Ch14)		-	21860	32686
Exchange(Wide.t:Ch14)		-	804	34796
WilsonBalding(WilsonBalding.t:Ch14)		-	1394	34782
<pre>beast.base.inference.operator.kernel.BactrianDeltaExchangeOperat</pre>		0.39628	86822	28432
ScaleOperator(KappaScaler.s:APOBS1)		0.18949	6789	1665
ScaleOperator(KappaScaler.s:APOBS2)		0.21016	7722	1577
ScaleOperator(KappaScaler.s:BDNF)		0.18970	7620	1565
ScaleOperator(KappaScaler.s:BRCA1S1)		0.18986	7245	1581
ScaleOperator(KappaScaler.s:BRCA1S2)		0.19130	6763	1659
ScaleOperator(KappaScaler.s:Ch14)		0.30988	6979	1629
ScaleOperator(KappaScaler.s:Ch21)		0.25278	6882	1641
ScaleOperator(KappaScaler.s:Ch24)		0.23802	7303	1601
ScaleOperator(KappaScaler.s:CHST12)		0.18101	7118	1618
ScaleOperator(KappaScaler.s:CMKOR1)		0.18940	7070	1624
ScaleOperator(KappaScaler.s:FGFR3)		0.20670	7454	1577
ScaleOperator(KappaScaler.s:GHRex09)		0.19213	7583	1583
ScaleOperator(KappaScaler.s:RAG1)		0.22845	7414	1582
ScaleOperator(KappaScaler.s:TMEM20)		0.21491	7438	1592
ScaleOperator(KappaScaler.s:VANGL2)		0.19990	7590	1588
ScaleOperator(KappaScaler.s:VWF)		0.21476	7301	1609
<pre>starbeast2.NodeReheight2(Reheight.t:Species)</pre>	• • •	-	98775	143875
<pre>starbeast2.CoordinatedUniform(coordinatedUniform.t:Species)</pre>	• • •	-	354356	41500
$\verb+starbeast2.CoordinatedExponential(coordinatedExponential.t:Speci$		0.07335	453543	31525
<pre>beast.base.inference.operator.UpDownOperator(updownAll:Species)</pre>		0.66322	74117	23371
<pre>starbeast2.RealCycle(constPopSizesSwap.Species)</pre>		2.00000	30178	12398
<pre>ScaleOperator(constPopSizesScale.Species)</pre>		0.28757	37835	11563
<pre>ScaleOperator(constPopMeanScale.Species)</pre>	• • •	0.44194	16611	3449
${\tt ScaleOperator(netDiversificationRateScale.t:Species)}$	• • •	0.20702	13692	3787
<pre>ScaleOperator(ExtinctionFractionScale.t:Species)</pre>	• • •	0.11933	5801	1994
$\verb+base.inference.operator.UniformOperator(ExtinctionFractioFractionFractionFractionF$	• • •	-	13197	1214
SubtreeSlide(bdSubtreeSlide.t:Species)	• • •	0.06648	236813	53210
WilsonBalding(bdWilsonBalding.t:Species)	• • •	-	495	77058
Exchange(bdWide.t:Species)	• • •	-	4258	76440

Exchange(bdNarrow.t:Species)	 -	38229	730568
Uniform(bdUniformOperator.t:Species)	 -	59042	710796
<pre>ScaleOperator(bdTreeRootScaler.t:Species)</pre>	 0.87763	7185	147379
<pre>ScaleOperator(bdTreeScaler.t:Species)</pre>	 0.98773	27551	126232

Tuning: The value of the operator's tuning parameter, or '-' if the operator can't be optimiz #accept: The total number of times a proposal by this operator has been accepted.

#reject: The total number of times a proposal by this operator has been rejected.

Pr(m): The probability this operator is chosen in a step of the MCMC (i.e. the normalized we Pr(acc|m): The acceptance probability (#accept as a fraction of the total proposals for this operator)

```
Total calculation time: 375.843 seconds
End likelihood: -16950.572066013454
Done!
```

3.9 Checking the log file

To check parameters other than species tree topologies and branch values, or to verify that the chain has been run long enough to reliably represent the posterior distribution, we will use Tracer. Start the Tracer app and then open the "starbeast.log" file. The first statistic that will be displayed is a histogram of the log posterior probability (Figure 13).



Figure 13: Opening a log file in Tracer

The posterior probability is the sum of the likelihood (which is the sum of log phylogenetic likelihoods for all sites for all loci), the prior probability (which is the sum of log prior probabilities for all parameters), and the speciescoalescent (which is the sum of log coalescent probabilities for all gene trees). TreeHeight.Species is the height of the root node of the species tree, and TreeLength.Species is the sum of all branch lengths in the species tree.

Tracer computes effective sample sizes (ESS) for each logged statistic and parameter. As a rule, ESS values should be at least 200, particularly for the important statistics just noted and for parameters of interest. Because of the stochastic nature of MCMC algorithms, your values **will be different** to those in Figure 13.

3.10 Checking the species trees

Start the DensiTree app (included with BEAST2), and then open the "species.trees" file. Under the Show panel, enable the Root Canal tree to get an idea of the most plausible species tree. Then open the Grid panel, and enable the full grid (Figure 14). If you want to check the clade posterior probabilities, select "View clade toolbar" from the Window menu.



Figure 14: Viewing the species trees in DensiTree

You can see that using a fixed clock rate of 10^{-3} substitutions per site per year, the split between *Canis latrans* (coyotes) and *Canis lupus* (wolves) is probably less than 500,000 years ago. The age of the most recent common ancestor (MRCA) of extant *Canis* taxa is approximately 2.5 million years ago, more recent than the split between humans and chimpanzees (Prado-Martinez et al. 2013).

3.11 Generating a summary tree

Summary trees are a way of reducing a posterior distribution of tree topologies and times to a single tree. This is often more readable than the cloud of trees that DensiTree displays, but researchers should be cautious not to give it too much emphasis because for most clades and data sets there is substantial uncertainty in the posterior distribution of topologies and times.

To generate a summary tree, open the Tree Annotator app included with BEAST2. Set the burnin

percentage to 10 and choose the "species.trees" file generated by StarBEAST2 as the input file. Specify the output file to be something sensible like "summary.tree" and click Run to generate the summary tree (Figure 15).

	TreeAnnotator 2.7.5	- 0	×
Burn in percentage:	10		
Posterior probability limit:	0.0		
Target tree type:	Maximum clade credibility tree 🔹		
Node heights:	Common Ancestor heights 🔹		
Target tree file:		Choose file	
Input Tree File:	/home/huey/IdeaProjects/BEAST2/starbea	Choose file	
Output File:	;t2/tutorial/CanisPhylogeny/summary.tree	Choose file	
Low memory:			
	Quit	Run	

Figure 15: Running Tree Annotator

Load http://icytree.org/ in your web browser and open the summary tree file using the "Load from file" option in the File menu. In the Style menu, under "Node height error bars," choose "CAheight_95%_HPD" to display 95% highest posterior density (HPD) intervals of species divergence times on each branch. Also in the Style menu, under "Internal node text," choose "dmv1" to display estimates of ancestral effective population size. Enable the time axis by typing the letter "a" or by going through the Style menu again (Figure 16).

In StarBEAST2 and many other BEAST packages, effective population sizes are scaled by generation time. That means if generation times (the average number of years from zygote to zygote) vary between species in your analysis, the effective population sizes of different branches cannot be directly compared.

If the average generation time is 5 years, and our tree is scaled in millions of years, then the generation time is 5×10^{-6} . To get effective population sizes in numbers of individuals, divide each value by the generation time. So if the scaled effective population size is 1, that will correspond to 200,000 individuals.

4 Estimating per-species clock rates

We can estimate molecular clock rates separately for each species using a relaxed clock model, for example the uncorrelated log-normal (UCLN) model. First create a new folder for this analysis with a sensible name, something like "CanisUCLN". Relaunch BEAUti, and select "SpeciesTreeUCLN" from the Templates submenu of the File menu (Figure 17).

Import the same FASTA files as in subsection 3.1 (Figure 4). For any of the species tree relaxed clock templates, do **NOT** link the clock models. This will break the inference of per-species clock rates.

Assign the same species names as in subsection 3.3 (Figure 6, 7). This time we will keep the default setting for the population model (Analytical Integration), so that our analysis will run slightly faster. Set all the site models to HKY with empirical frequencies as in subsection 3.5 (Figure 9, 10).



Figure 16: Running IcyTree. This tree been made more legible by increasing font size and limiting numbers to two significant figures, further options under IcyTree's "Style" menu.

Open the Clock Model tab, and manually set the Clock.rate to 0.001 (or equivalently 1e-3) for **every** locus (Figure 18). This is necessary because for technical reasons we cannot link the clock models for the species tree relaxed clock templates.

Open the Priors tab and change the species tree prior from Yule to Birth-Death to allow for extinction. Finally, open the MCMC tab and change the length of the chain to 40 million, as in subsection 3.7 (Figure 12). Save the file in the folder you created previously for this analysis, and give it a sensible name like "Canis-UCLN.xml".

Run the MCMC chain by opening the XML file in BEAST, or running it from the command line as in subsection 3.8. This will take about twice as long (10 to 20 minutes) as the fixed clock analysis, because relaxed clocks are more computationally intensive.

After BEAST has finished, open the "starbeast.log" file in Tracer to check that the important statistics have ESS values of at least 200. Select the branchRatesStdev.Species parameter (Figure 19).

This parameter models the spread of molecular clock rates among branches in the species tree. The default prior for this parameter in StarBEAST2 is a log-normal distribution with a mean of 1, but the posterior distribution of this parameter has a mean of about 0.2. This means that the data is pulling this parameter lower, probably because there is very little variation in clock rates between species. Indeed the mode of the posterior distribution is about zero, suggesting that a strict clock may be more appropriate (Figure 19).

New	Ctrl+N	el Clo	ck Model	Priors M	ИСМС			
Load	Ctrl+O	Link C	ock Models	Unlink Cloc	k Models	Link Trees	Unlink Trees	
Add Morphological Data		Sites	Data Typ	e Site Mode	Clock N	Mo Tree	Ambiguit	
Import Alignment								
Template	•	Standard						
Set working dir	•	StarBeas	t2					
Manage Packages		SpeciesT	reeUCED					
Clear Class Path		SpeciesT	reeRLC					
Launch Apps		AIM						
Save	Ctrl+S	SpeciesT	regUCLN					
Save As		StarBeas	t3					
Close	Ctrl+W	Other Te	mplate					
Exit	Ctrl+Q							
				No par	titions loa	ded vet		

Figure 17: Starting a UCLN relaxed clock analysis

5 Total evidence tip-dating

StarBEAST is able to combine tip-dating using the fossilized birth death (FBD) process with multispecies coalescent (MSC) inference of species and gene trees. We call this integrative model "FBD-MSC". To begin your analysis, creat a new folder for this purpose with a sensible name, something like "CanisFBD".

When setting up an FBD-MSC analysis in BEAUti, it is necessary to import morphological data <u>after</u> specifying the taxon sets and the tree model, but <u>before</u> specifying the tip dates. The following steps will be consistent with that ordering. First, repeat section 3 from subsection 3.1 to and including 3.3. Leave the population model set at the default (Analytical Integration) to save time, then set all site models to HKY with empirical frequencies as in subsection 3.5.

Before adding the morphological data, go to the Priors panel and change the prior on the species tree from Yule to FBD (Figure 20). This is necessary because the inclusion of fossil taxa means our species tree will be *serially sampled* rather than *ultrametric*. StarBEAST2 gives you the option of various birth-death models – Yule, Calibrated Yule, Birth Death and FBD – but only FBD is valid for serially sampled trees.

Now go back to the Partitions tab, and from the File menu select "Add Morphology to Species Tree" (Figure 21). Navigate to the data folder of the tutorial and import the "morphology_canis.nex" file.¹

 $^{^{1}}$ This file contains recorded states for 50 morphological characters of *Canis* species and some related species, taken from Slater 2015.

				BEAUti 2: SpeciesTreeUCLN	×
File Mo	de View	Help		· · · · · · · · · · · · · · · · · · ·	_
Partitions	Taxon sets	Tip Dates	Gene Ploidy	y Population Model Site Model Clock Model Priors MCMC	
Partition Partition APOBS1 BRCA151 BRCA152 Ch14 Ch21 Ch24 CH5T12 CMKOR1 FGFR3 GHRex09 RAG1 TMEN20 VANGL2 VWF		× v bates	Clock.rate	• Openation model Determine 0.001 • estimate mean clock rate (defaults to 1.0)	

Figure 18: Manually setting every clock rate



Figure 19: Checking the species tree relaxed clock analysis in Tracer

-											
File	Mode View H	elp	r		ĩ		(
Partitio	ons Taxon sets T	ip Dates G	iene Ploidy	Population Mo	odel	Site Model	Clock Model Pr	iors M	ICMC		
•	Tree.t:Species		Yule Mo	del	•	?					
•			Yule Moo	iel							
	kappa.s:APOBS1		Calibrate	d Yule Model		ini 🕐	tial = [2.0] [0.0,∞]		HKY tra	nsition-transversion parameter of partition	on s:AP
•			Birth Dea	ath Model							
•	kappa.s:APOBS2		FBDMod	el 🕟		²⁾ ini	tial = [2.0] [0.0,∞]				
▶						?)	ti-l - 12 01 10 01				
	Kappa.s.bbNr		LOG NOT	mai[1.0,1.25]	•		tiai – [2.0] [0.0,∞]				
	kappa.s:BRCA1S1		Log Nor	mal[1.0,1.25]	•	² ini	tial = [2.0] [0.0,∞]				
	kappa.s:BRCA1S2		Log Nor	mal[1.0,1.25]	•	🤊 🚺 ini	tial = [2.0] [0.0,∞]				
•											
•	kappa.s:CHST12		Log Nor	mal[1.0,1.25]	•	²⁾ ini	tial = [2.0] [0.0,∞]				
▶	kappa c/CMKOB1					2)	ti-l - 12 01 10 01				
	карра.з.сыкокт		LOG NOF	mai[1.0,1.25]	•		tiai = [2.0] [0.0,∞]				
	kappa.s:Ch14		Log Nor	mal[1.0.1.25]	•	² ini	tial = [2.0] [0.0,∞]				
•	kappa.s:Ch21		Log Nor	mal[1.0,1.25]	•) ini	tial = [2.0] [0.0,∞]				
•	kappa.s:Ch24		Log Nor	mal[1.0,1.25]	•	²⁾ ini	tial = [2.0] [0.0,∞]				
						2					
•	kappa.s:FGFR3		Log Nor	mal[1.0,1.25]	•	ini	tial = [2.0] [0.0,∞]				

Figure 20: Change the tree model from Yule to FBD.

							BEA	٩Uti	2: StarBeast2					- (
File N	lode View	Hel	р												
🗋 New	/			Ctrl+N	y F	Population l	Model Site	Mo	del Clock Model	Priors	MCMC				
🖻 Load	d			Ctrl+O	Cloc	k Models	Unlink Clo	ck N	lodels Link Tree	ur Ur	nlink Trees				
Add	Morphologi	cal Dat	a		pe	Sit	e Model		Clock Mod	el	Tree		Ambiguities		
4 Imp	ort Alignmer	nt			de	APOBS1		•	molecular	•	APOBS1	•			
Add	Morphology	to Spe	ecies Tree	•	de	APOBS2		•	molecular	-	APOBS2	-			
E Tem	iplate working dir			•	de	BDNF		•	molecular	-	BDNF	-			
E Mar	nage Package	25		,	de	BRCA1S1		•	molecular	•	BRCA1S1	•			
E Clea	r Class Path				de	BRCA152		•	molecular	-	BRCA1S2	•			
C Laui	nch Apps				de	Ch14		•	molecular	•	Ch14	•			
c 🖪 Save	2			Ctrl+S	de	Ch21		•	molecular	-	Ch21	-			
Save	e As				de	Ch24		•	molecular	-	Ch24	-			
Clos	e			Ctrl+W	de	CHST12		•	molecular	-	CHST12	-			
CIVINONI	CIVINONI	10	100	пистеот	de	CMKOR1		•	molecular	-	CMKOR1	-			
FGFR3	FGFR3	16	503	nucleot	ide	FGFR3		•	molecular	-	FGFR3	-			
GHRex09	GHRex09	16	736	nucleot	ide	GHRex09) .	•	molecular	-	GHRex09	•			
RAG1	RAG1	16	741	nucleot	ide	RAG1		•	molecular	-	RAG1	-			
TMEM20	TMEM20	16	615	nucleot	ide	TMEM20		•	molecular	-	TMEM20	•			
VANGL2	VANGL2	16	546	nucleot	ide	VANGL2		•	molecular	-	VANGL2	-			
VWF	VWF	16	732	nucleot	ide	VWF		•	molecular	-	VWF	-			
															>

Figure 21: The option to add morphological data to the species tree

Select Yes to condition on recording variable characters only (Mkv). Mkv is used when only characters that vary between species have been included in the character matrix, as is the case for this tutorial's data set. You should now see four morphology partitions (Figure 22), one for each number of character states. For example, morphology_canis2 is for binary characters.

					BEAUti 2: StarBeas	t2						- • ×
File Mode Vie	w Help											
Partitions Taxon s	ets Tip Dates	Gene Ploidy	Popula	tion Model	Site Model Clock M	odel	Priors	МСМС				
Name	File	Таха	Sites	Data Type	Site Model			Clock Model		Tree		Ambiguit
BDNF	BDNF	16	489	nucleotide	BDNF	•	mo	lecular	•	BDNF	-	
BRCA1S1	BRCA1S1	16	705	nucleotide	BRCA1S1	•	mo	lecular	•	BRCA1S1	•	
BRCA1S2	BRCA1S2	16	741	nucleotide	BRCA1S2	•	mo	lecular	•	BRCA1S2	-	
Ch14	Ch14	16	921	nucleotide	Ch14	•	mo	lecular	-	Ch14	-	
Ch21	Ch21	16	601	nucleotide	Ch21	•	mo	lecular	•	Ch21	-	
Ch24	Ch24	16	730	nucleotide	Ch24	•	mo	lecular	•	Ch24	-	
CHST12	CHST12	16	705	nucleotide	CHST12	•	mo	lecular	-	CHST12	•	
CMKOR1	CMKOR1	16	735	nucleotide	CMKOR1	•	mo	lecular	-	CMKOR1	-	
FGFR3	FGFR3	16	503	nucleotide	FGFR3	•	mo	lecular	-	FGFR3	•	
GHRex09	GHRex09	16	736	nucleotide	GHRex09	•	mo	lecular	-	GHRex09	•	
RAG1	RAG1	16	741	nucleotide	RAG1	•	mo	lecular	•	RAG1	•	
TMEM20	TMEM20	16	615	nucleotide	TMEM20	•	mo	lecular	•	TMEM20	•	
VANGL2	VANGL2	16	546	nucleotide	VANGL2	•	mo	lecular	•	VANGL2	•	
VWF	VWF	16	732	nucleotide	VWF	•	mo	lecular	•	VWF	•	
morphology_canis2	morphology_ca	inis 24	29	standard	morphology_canis	•	mo	rphology_cani:	•	Species	•	
morphology_canis3	morphology_ca	anis 24	16	standard	morphology_canis	•	mo	rphology_cani:	•	Species	•	
morphology_canis4	morphology_ca	anis 24	11	standard	morphology_canis	•	mo	rphology_cani:	•	Species	•	
morphology_canis5	morphology_ca	anis 24	8	standard	morphology_canis	•	mo	rphology_cani:	•	Species	•	
												<u>></u>
+ - r Split												, ``

Figure 22: After morphological data has been added to the species tree

Open the Tip Dates panel and enable "Use tip dates". Change "Since some time in the past" to "before the present". Click on Auto-configure and select "read from file". Choose the "tip_dates_canis.dat" file in the data folder of the tutorial. Click OK, and your tip dates should look like Figure 23.

Go to the Clock Model panel, and enable "estimate" for the molecular clock (by default called APOBS1 after the first locus, although this can be changed in the Partitions tab). Then select the morphological data partition and enable "estimate" so that the morphological clock rate will also be estimated (Figure 24). Make sure you have enabled "estimate" for **both** clocks, since we are using fossil data to calibrate our tree.

Next, go back to the Priors panel, scroll down to the strictClockRate.c:molecular parameter, and expand that prior. Change the mean to 0.001, the *a priori* molecular clock rate estimate from section 3. Leave the standard deviation set at 1, a moderately informative prior that will still allow the rate to be guided by the fossil data. Then change the prior distribution for the strictClockRate.c:morphology_canis prior to 1/X (Figure 25). This is an improper, uninformative prior so the inferred morphological clock rate will be estimated solely from the data.

	BEAUti 2: StarBeast2		- • •
File Mode View Help			
Partitions Taxon sets Tip Dates Gene Ploidy Population Model	Site Model Clock Model Priors	s MCMC	
✓ Use tip dates			
Before the present	Auto-configure Clear		
Name	Date	Height	
Canis_mesomelas	0.000000	0.0	
Canis_anthus	0.000000	0.0	
Canis_lupus	0.000000	0.0	
Canis_latrans	0.000000	0.0	
Lycaon_pictus	0.000000	0.0	
Cuon_alpinus	0.000000	0.0	
Canis_simensis	0.000000	0.0	
Canis_adustus	0.000000	0.0	
Canis_antonii	1.200000	1.2	
Canis_armbrusteri	0.300000	0.3	
Canis_arnensis	1.500000	1.5	
Canis_chihliensis	1.200000	1.2	
Canis_dirus	0.012000	0.012	
Canis_edwardii	0.300000	0.3	
Canis_etruscus	1.700000	1.7	
Canis_falconeri	1.500000	1.5	
Canis_ferox	3.500000	3.5	
Canis_lepophagus	1.800000	1.8	
Canis_mosbachensis	0.500000	0.5	
Canis_palmidens	1.500000	1.5	
Canis_thooides	1.800000	1.8	
Canis_variabilis	0.500000	0.5	
Eucyon_davisi	4.900000	4.9	

Figure 23: Configured tip dates

					BEAUti 2: S	StarBeast2			•	
File Mo	ode View	Help								
Partitions	Taxon sets	Tip Dates	Gene Ploidy	Population Model	Site Model	Clock Model	Priors	MCMC		
Partition			Strict Clock	•	?					
morpholo	av canis		Clock.rate		1.0	🤊 💿 🗸 e	stimate			
	<i></i>					- T	Estimate	value of this parameter in the MCMC chain		

Figure 24: Estimating the clock rates



Figure 25: Configuring the clock prior distributions

The FBD-MSC model is quite complex and requires very long chains for reliable sampling. Open the MCMC tab and change the chain length to 200 million. We will need to reduce the sampling rate so that the output files remain managably small, so change the Store Every value to 50,000. Expand the tracelog, speciesTreeLogger, screenlog and every gene tree log, and set the Log Every value to 50,000 for each of them. This will limit the number of samples to 4,000 in each log file (Figure 26).

	BEAUti 2: StarBeast2	•	×
File Mode View Help			
Partitions Taxon sets Tip Dates	Gene Ploidy Population Model Site Model Clock Model Priors MCMC		
Chain Length	20000000		Î
Store Every	50000		
Pre Burnin	0 0		
Num Initialization Attempts	10 💿		
▼ tracelog			
File Name	starbeast.log		
Log Every	50000 I @		
Mode	Number of the samples logged autodetect		
Sort	smart v		
Sanitise Headers 💿			
posterior			
likelihood			
prior			
speciescoalescent			
popMean.Species			
TreeHeight.Species			
TreeLength.Species			~

Figure 26: Changing the sampling rates after setting a chain length of 200 million

Save the XML file in the folder you created for this analysis, and name it something like "CanisFBD.xml". Run the MCMC chain by opening the XML file in BEAST, or running it from the command line as in subsection 3.8. The chain should take between 1 and 2 hours to finish, depending on the speed of your computer.

Editing BEAST XML files

BEAUti is a wonderful tool, but sometimes it can be faster to edit XML files directly rather than using BEAUti to configure every aspect of your model. Setting the sampling rate for gene tree log files one at a time is a pain, especially if you have 100 or more loci in your analysis! Instead of changing the sampling rates in BEAUti, you can leave them at their default values, and use a text editor to change them all at once. After opening your XML file in the editor, find-andreplace all occurences of the sampling rate attribute and its default value, e.g. logEvery="5000",

with the value changed to the one desired, e.g. logEvery="50000".

You may even want to design a model in a way that is impossible to configure with BEAUti, in which case you *must* edit the XML file directly. However, that sort of thing is beyond the scope of this tutorial.

After BEAST has finished, open the "starbeast.log" file in Tracer to check on the ESS values. These values are likely insufficient given a chain length of only 200 million for such a complex model, and a real analysis might require around 2 billion states, possibly spread across multiple chains.

Select the strictClockRate.c:molecular parameter (Figure 27). The 95% HPD interval for this parameter spans approximately 0.0005 to 0.0013 substitutions per site per million years, in agreement with the *a priori* rate of 0.001.



Figure 27: The estimated posterior distribution of the molecular clock rate

Now open the "Trace" tab to see how this value wandered around over the course of the MCMC chain. With only 200 million states this is unlikely to look like the ideal "hairy catepillar," and will instead exhibit slow transitions between modes of the distribution (Figure 28).



Figure 28: The MCMC trace of the molecular clock rate

Start the DensiTree app again, and then open the "species.trees" file. Under the Show panel, enable the Root Canal tree to get an idea of the most plausible species tree. Then open the Grid panel, and enable the full grid. Compared with the fixed clock analysis (Figure 14), the divergence time between *Canis lupus* and *Canis latrans* is a little older, and the age of the MRCA of extant species is a little younger (Figure 29).



Figure 29: Cloud of FBD-MSC species trees

Open the Tree Annotator app included with BEAST2, and set the burnin percentage to 10. Trees inferred by the FBD-MSC model as implemented in StarBEAST2 can include sampled ancestors (Gavryushkina et al. 2014), which are incompatible with Common Ancestor heights (Heled and Bouckaert 2013) as implemented in Tree Annotator. So in order to generate a summary tree, change "Node heights" to "Mean heights" (Figure 30). Choose the "species.trees" file as the input file, and specify the output file to be something sensible like "summary.tree", then click Run.

Go back to IcyTree (http://icytree.org/) in your web browser and open the summary tree file using the "Load from file" option in the File menu. In the Style menu, under "Node height error bars," choose "height_95%_HPD" to display 95% highest posterior density (HPD) intervals of species divergence times on each branch. This time show the posterior support of each clade by choosing "posterior" in "Internal node text" from the Style menu. Enable the time axis again through the Style menu or by typing the letter "a" (Figure 31).

The tree topology is similar to the fixed clock analysis (Figure 16), except that *Cuon alpinus* and *Lycaon pictus* are now a clade with 100% support. A disagreement between molecular data which contradicts this clade and morphological data which supports it has been reported previously (Zrzavý and Řičánková 2004). This shows how morphological characters and fossil data can influence species tree estimates.

	TreeAnnotator 2.7.5	- • ×
Burn in percentage:	10	
Posterior probability limit:	0.0	
Target tree type:	Maximum clade credibility tree 🔹	
Node heights:	Common Ancestor heights 🔹	
	Common Ancestor heights	
Target tree file:	Median heights	Choose file
	Mean heights	
Input Tree File:	Keep target heights	Choose file
Output File:	/home/huey/IdeaProjects/BEAST2/starbea	Choose file
Low memory:		
	Quit	Run

Figure 30: Running Tree Annotator for sampled ancestor trees



Figure 31: Showing posterior support for clades in IcyTree

6 Wrapping up

That concludes the tutorial. One final note; when running any MCMC method, it is possible than the chain has gotten stuck in a region of parameter space with high local probability. So for any kind of analysis intended for publication, it is a good idea to run at least two MCMC chains. Open both chains in Tracer, and check that the traces for the logged statistics and parameters have the same distributions.

This tutorial was written by Huw A. Ogilvie for Taming the BEAST and is licensed under a Creative Commons Attribution 4.0 International License.

Version dated: November 2, 2023

Relevant References

- Bouckaert, RR. 2010. Densitree: making sense of sets of phylogenetic trees. *Bioinformatics* 26: 1372–1373.
- Gavryushkina, A, D Welch, T Stadler, and AJ Drummond. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology* 10: e1003919.
- Hasegawa, M, H Kishino, and T Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22: 160–174.
- Heled, J and RR Bouckaert. 2013. Looking for trees in the forest: summary tree from posterior samples. BMC Evolutionary Biology 13: 221.
- Hugall, AF, R Foster, MSY Lee, and M Hedin. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. Systematic Biology 56: 543–563.
- Lindblad-Toh, K, CM Wade, TS Mikkelsen, EK Karlsson, DB Jaffe, M Kamal, M Clamp, JL Chang, EJ Kulbokas, MC Zody, et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438: 803–819.
- Ogilvie, HA, RR Bouckaert, and AJ Drummond. 2017. Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution* 34: 2101–2114.
- Ogilvie, HA, J Heled, D Xie, and AJ Drummond. 2016. Computational performance and statistical accuracy of *BEAST and comparisons with other methods. *Systematic Biology* 65: 381–396.
- Prado-Martinez, J et al. 2013. Great ape genetic diversity and population history. Nature 499: 471–475.
- Scornavacca, C and N Galtier. 2017. Incomplete lineage sorting in mammalian phylogenomics. Systematic Biology 66: 112–120.
- Slater, GJ. 2015. Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution. *Proceedings of the National Academy of Sciences* 112: 4897–4902.
- Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39: 306–314.
- Zrzavý, J and V Řičánková. 2004. Phylogeny of recent Canidae (Mammalia, Carnivora): relative reliability and utility of morphological and molecular datasets. *Zoologica Scripta* 33: 311–333.