Methodology in Phylogenetics

Joseph Crisp

BBSRC funded PhD Student

Rowland Kao
Ruth Zadoks

j.crisp.1@research.gla.ac.uk

Next Generation Sequencing Platforms

- Cheap and fast alternatives to Sanger sequencing methods

- Prone to error production

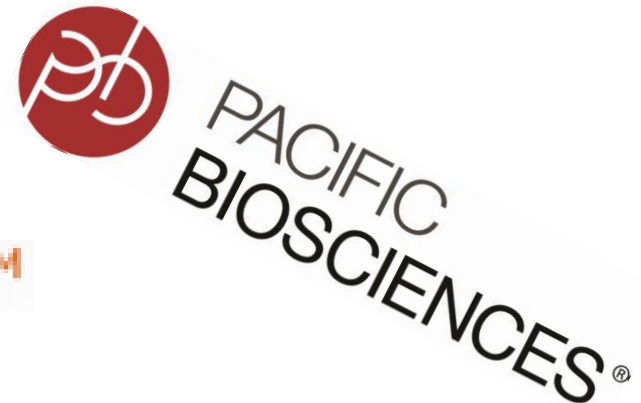- Provides a huge amount of data that can be used to establish the validity of variation observed
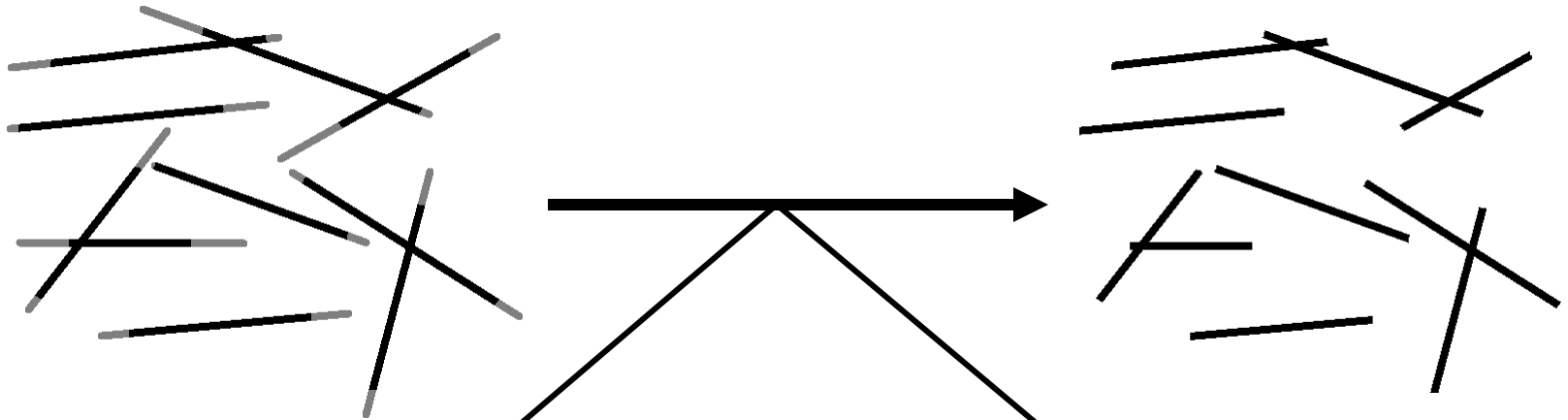
FASTQ File

@InstrumentName : RunId : FlowCellId : Lane : TileNumber : X-coord : Y-coord   PairId : Filtered? : Bit

CGCGTTCATGTGCGGCGCCGGCGCTGGTTTCCATCGGTCGATTCGGACGGCAAACATCGACCGCAGGATCCGCCGTACCATGTCCGACA
GGCGCCCCTTGGGTAGATTGCCGTCGGCGTAGGCGGCACGCAGGCGGACGGAGAGTGCTTCCGAGTGCCACGGCGCTGAATCGATCCG
CGCCGCGCACCTTTGGTCCAGGGCGGCCAGCGCGCACTCCCAGCTGGGTGTTTCGGCCCCATCCGCACTCACCC

+

1>1>>AA1@333AA11AE00AEAE///FH0BFFGB/E/?/>FHH/B//>EEGE0>0BGE0>//E/C//1<C?AFC///<FGF1?1?C?-<FF-@??-
C./CGG0:0C0;C0-.;C-?A@--.:?-@-@B@<@=-9--9---:;-9///;//-----/;/9----9;99///;-9----@;-9-;9-9-;//;B-9/B/9-;9@--@AE@@-;-
;-BB-9/-//;A9-;B-/--------9------//;--

FASTQC

Andrews 2010 – FastQC: A quality control tool for high throughput sequence data

ATGATAGCATAGGGATTTTACATACATACATACAGATACATAGGATCTACAGCGC

CATAGGGATTTTACATACATACATACAGATACATAGGA

Fabbro et al. 2013 – An extensive evaluation of read trimming effects on Illumina NGS data analysis

FASTQC

TGATCATTGGGATATTACATACATACATAGAGATACATAGCATAGT

CATAGGGATTTTACATACATACATACAGATACATAGGA

Li et al. 2008 – Mapping short DNA sequencing reads and calling variants using mapping quality scores
Li et al. 2009 – The sequence alignment MAP format and SAMtools
Garnier et al. 2003 – The complete genome sequence of Mycobacterium bovis

Read Depth

Mapping Quality

High Quality Base Depth

Allele Support

Biek et al. 2012 – Whole genome sequencing reveals local transmission patterns of *Mycobacterium bovis* in sympatric Cattle and Badger populations
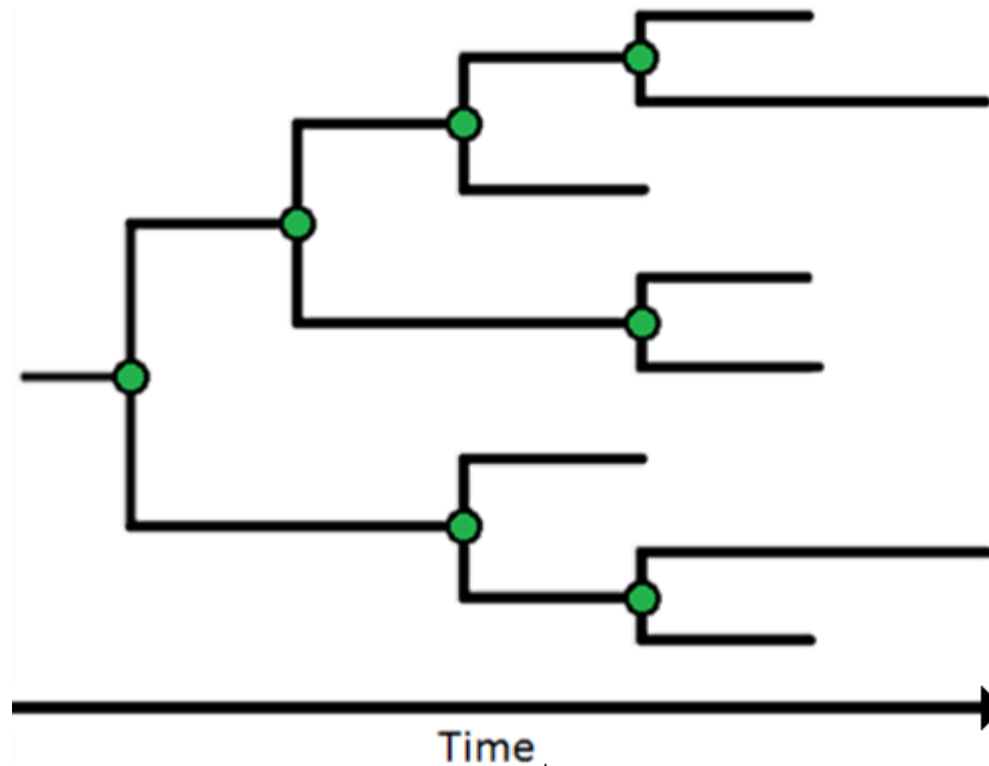Sampson 2011 – Mycobacterial PE/PPE proteins at the host-pathogen interface
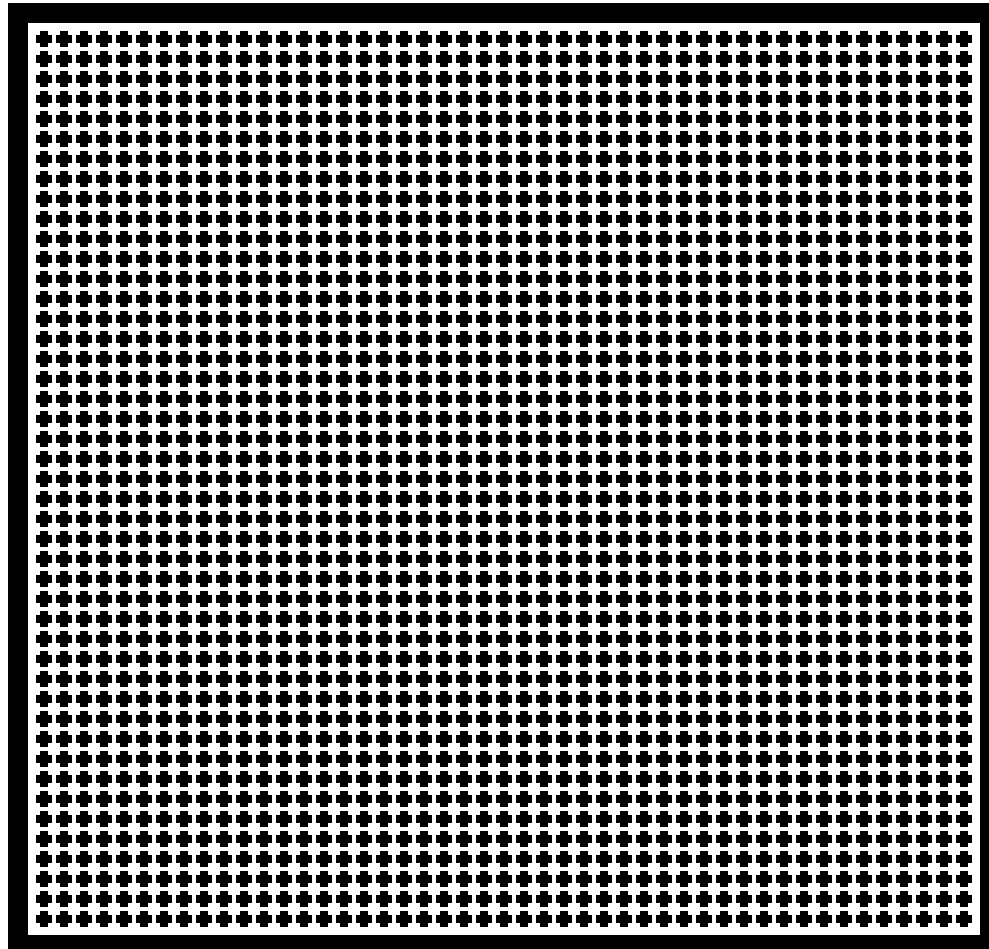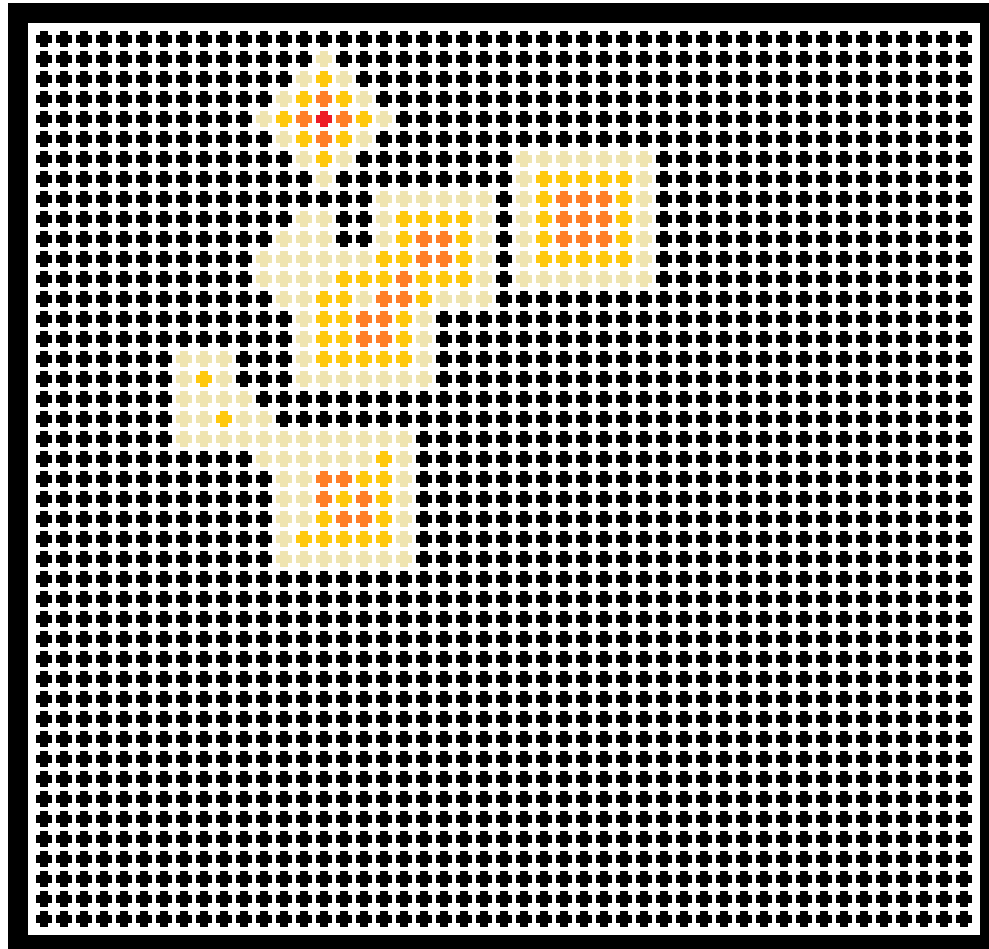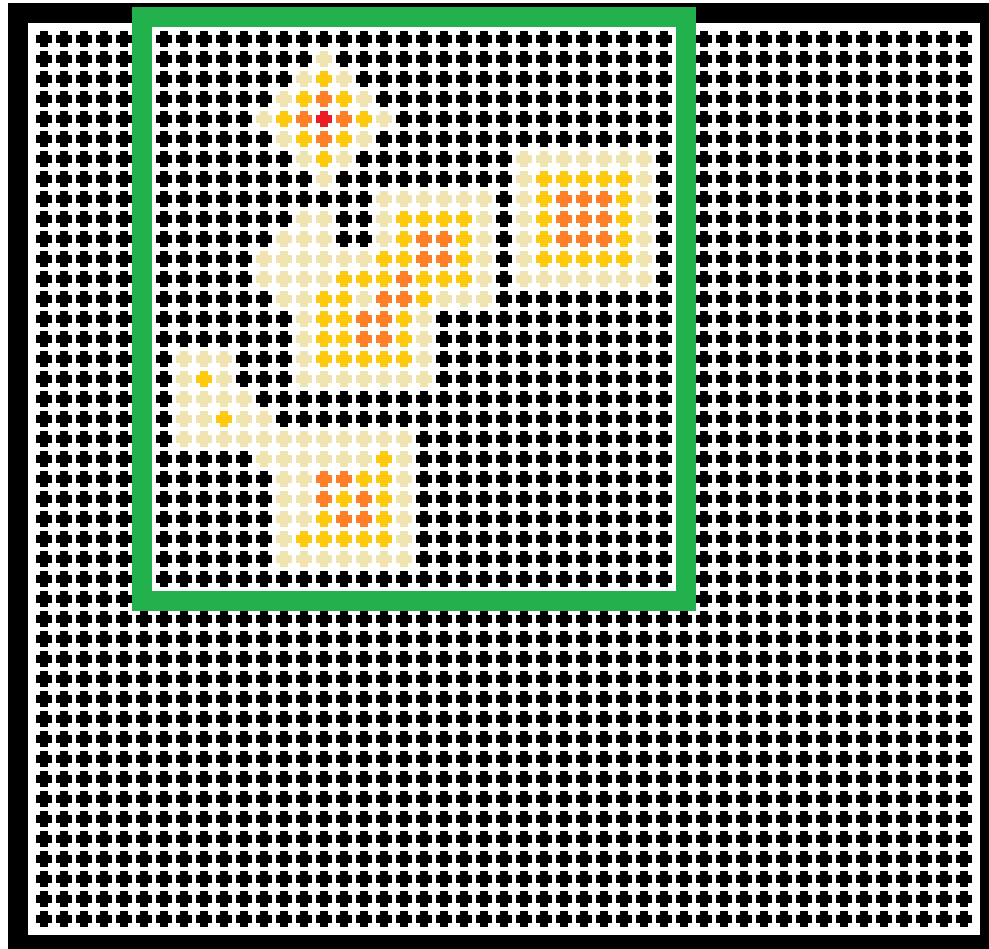
NEXUS File

```
# NEXUS


BEGIN DATA;
DIMENSIONS NTAX=3 NCHAR=40;
FORMAT MISSING=N GAP=- DATATYPE=DNA;
MATRIX;
IsolateId_Year   CGTACGCGTAGCGTGCACAGTGCCTAGTTGTGATCTGTCA
ExampleA_2009    GGTACGCCTTGCGTGGACAGTCCCTAGTTCTGATCAGTCT
ExampleB_2005    CGTACGCGTTGCGTGCACAGTGCCTAGTTCTGATCAGTCA
;
END;
```

Bishop 2014 – Bioinformatics and Data Analysis in Micobiology

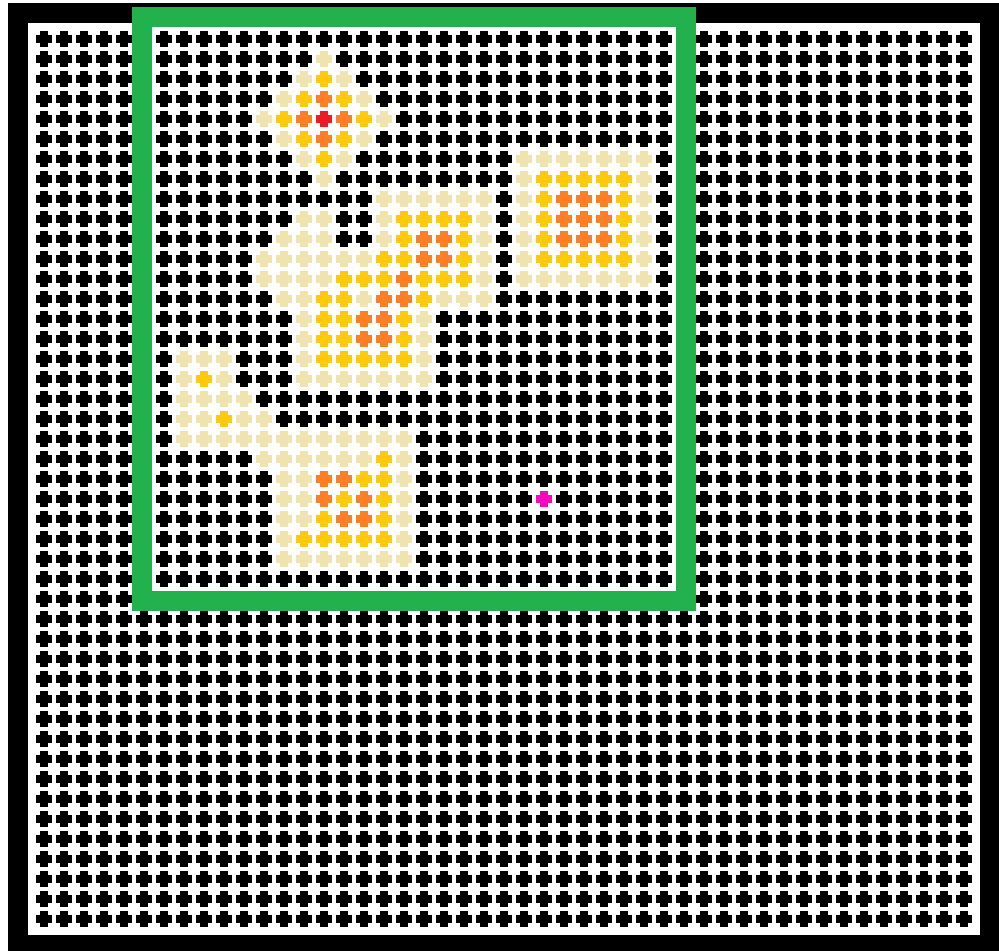Maddison et al. 1997 – NEXUS: an extensible file format for systematic information

Bayesian Evolutionary Analysis by Sampling Trees

Time

Drummond and Rambaut 2007 – BEAST – Bayesian evolutionary analysis by sampling trees
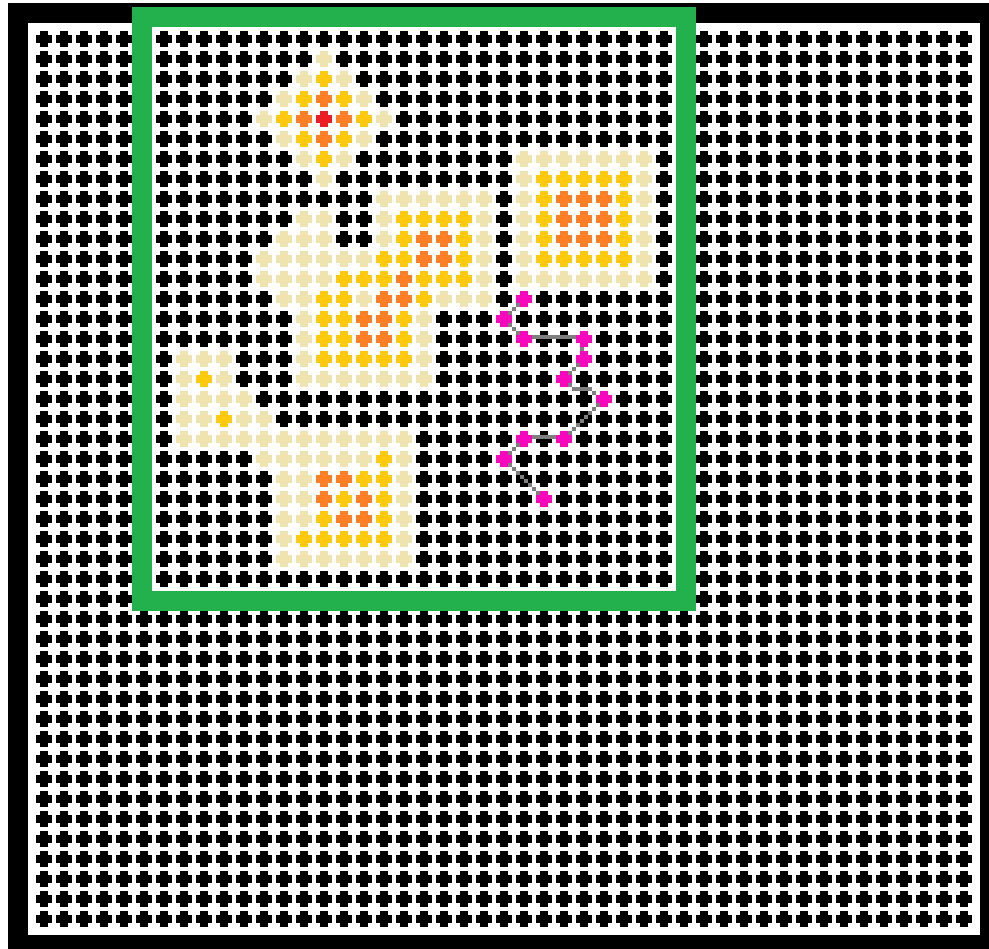Bouckaert et al. 2014 – BEAST2 – A software platform for bayesian evolutionary analyses
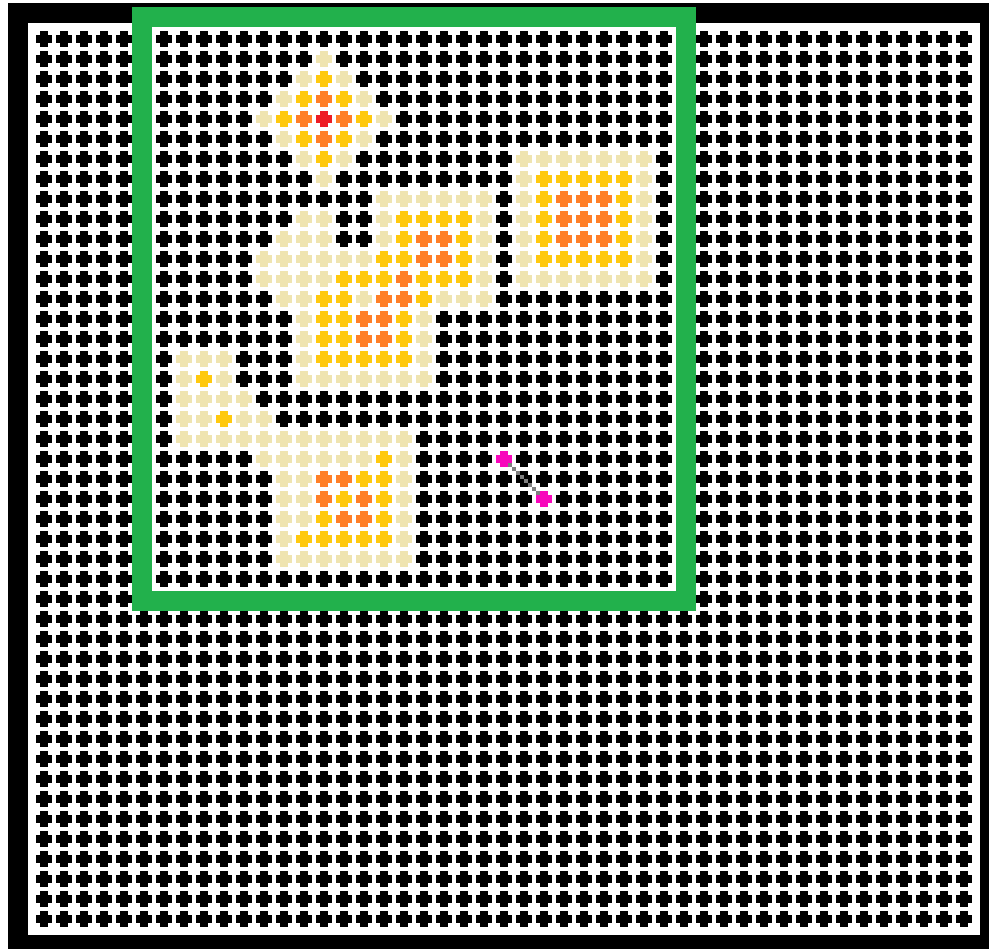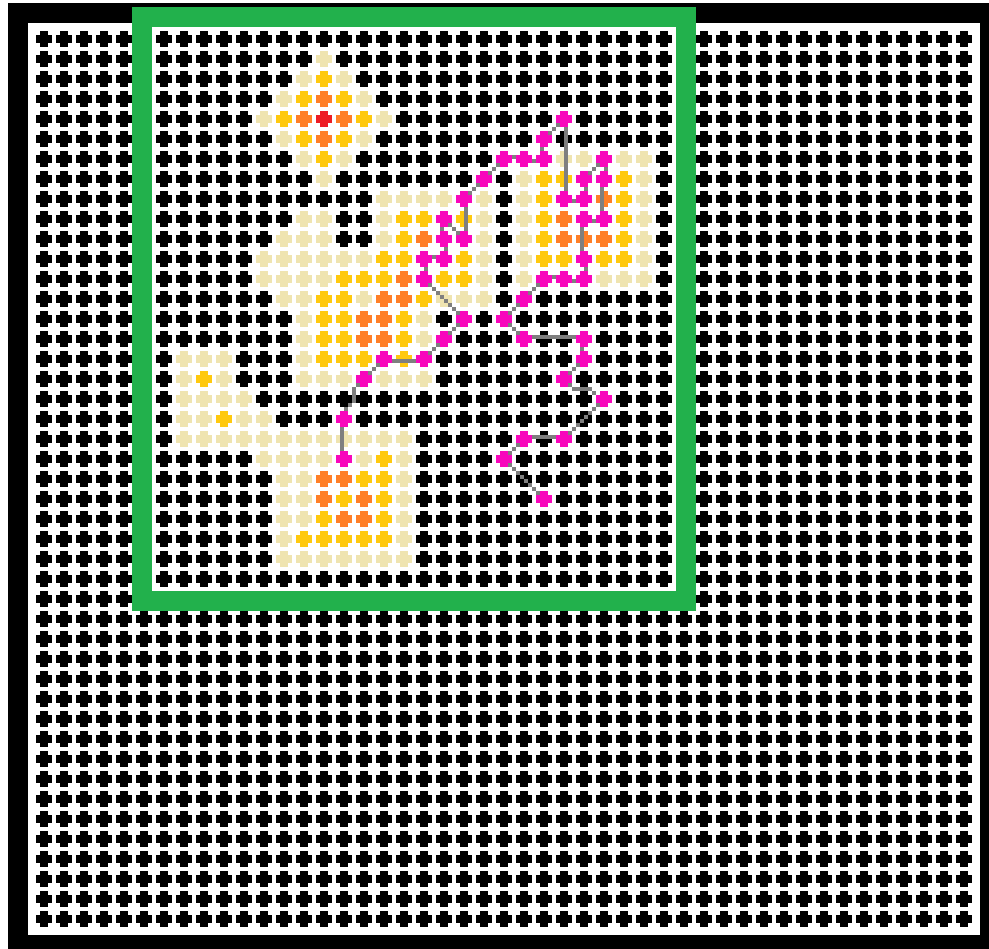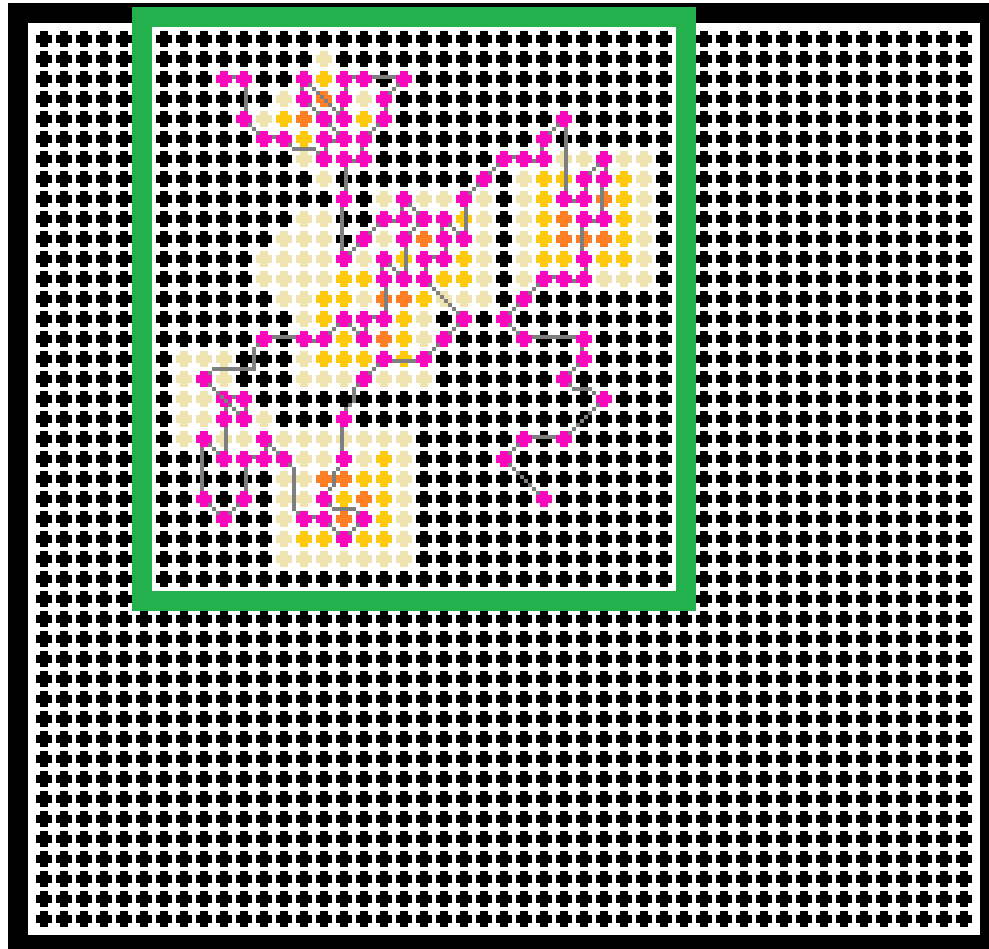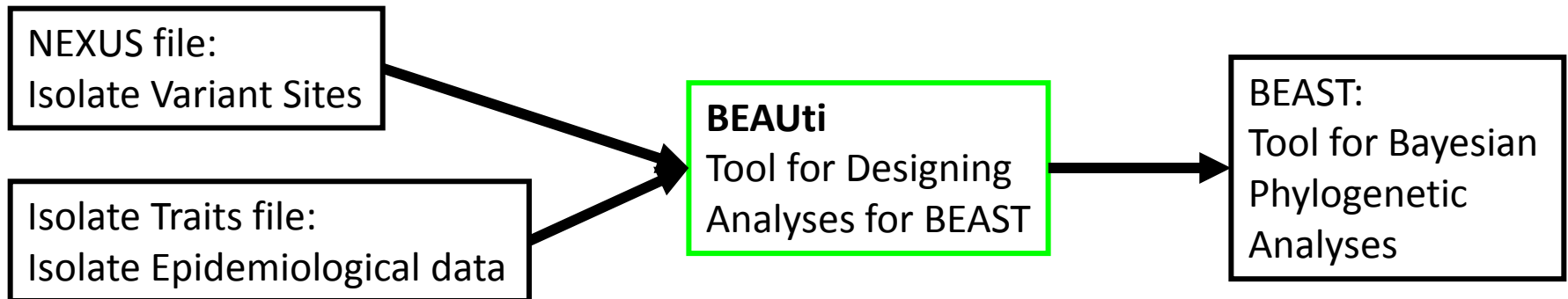
Bayesian Evolutionary Analysis Utility

NEXUS File

```
# NEXUS


BEGIN DATA;
DIMENSIONS NTAX=3 NCHAR=40;
FORMAT MISSING=N GAP=- DATATYPE=DNA;
MATRIX;
IsolateId_Year          CGTACGCGTAGCGTGCACAGTGCCTAGTTGTGATCTGTCA
ExampleA_2009           GGTACGCCTTGCGTGGACAGTCCCTAGTTCTGATCAGTCT
ExampleB_2005           CGTACGCGTTGCGTGCACAGTGCCTAGTTCTGATCAGTCA
;
END;
```

Isolate Traits File

traits	Species	Location	Latitude	Longitude	CATTLE
ExampleA_2009	BOVINE	WHATAROA	-42.345	170.998	BOVINE
ExampleB_2005	POSSUM	WESTPORT	-41.543	171.321	NONBOVINE

ExampleA_2009 **G**GTACGC**C**TTGCGTG**G**ACAGT**C**CCTAGTTCTGATCAGTC**T**
ExampleB_2005 **C**GTACGC**G**TTGCGTG**C**ACAGT**G**CCTAGTTCTGATCAGTC**A**

	A	C	G	T
A		?	?	?
C	?		?	?
G	?	?		?
T	?	?	?	

Transitions
Transversions

Jukes-Cantor	JC	1969	n=0
Hasegawa-Kishino-Yano	HKY	1985	n=4
Tamura-Nei	TN	1993	n=5
Generalised Time-Reversible	GTR	1986	n=8

Jukes & Cantor 1969 – Evolution of protein molecules
Hasegawa et al. 1985 – Dating of the human-ape splitting by a molecular clock of mitochondrial DNA
Tamura & Nei – Estimation of the number of nucleotide substitutions in the control region of the mitochondrial DNA in human and chimpanzees
Tavare 1986 – Some probabilistic and statistical problems in the analysis of DNA sequences
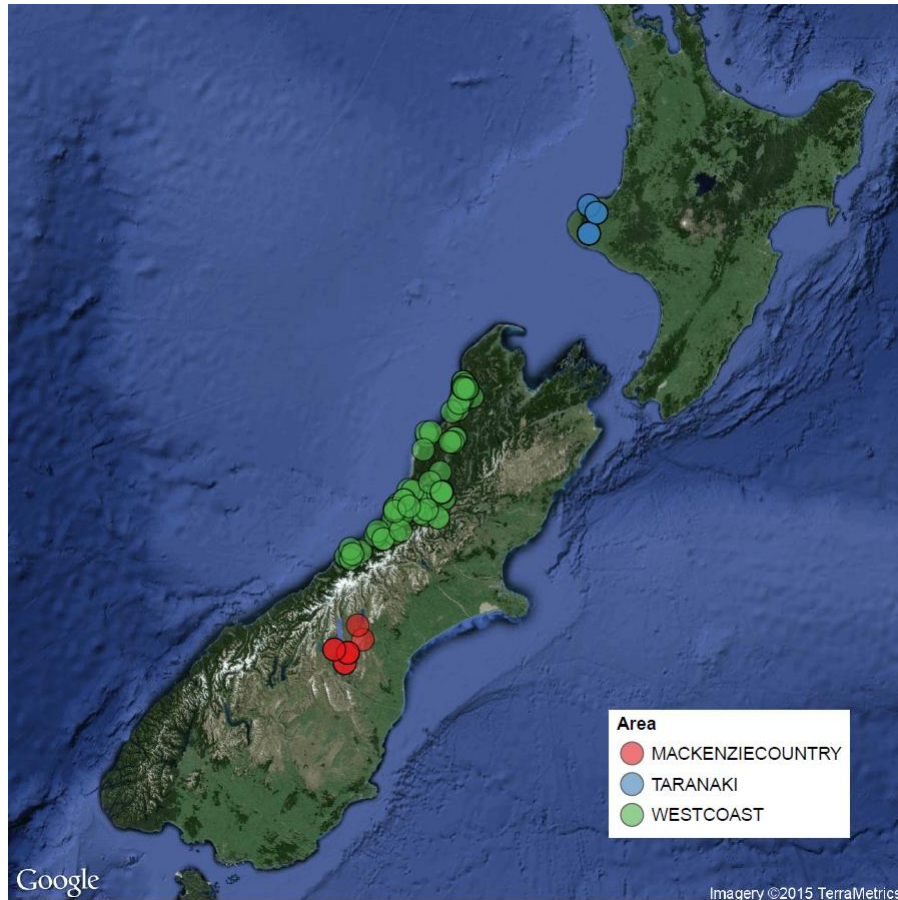
Sampling Range for each Species
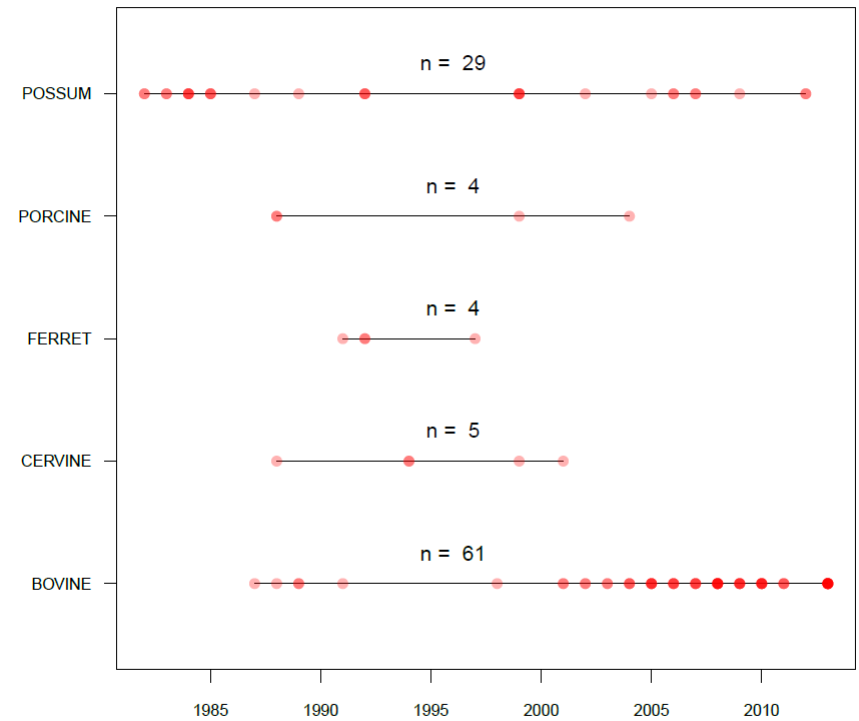
BEAUti

File Edit Help

Partitions | Taxon Sets | Tips | Traits | Sites | **Clocks** | Trees | States | Priors | Operators | MCMC

Clock Model :

Name	Model	Estimate	Rate	Group
ClockModel	Strict clock ▼	☑	1.0	nucleotide_group ▼
CATTLE	Strict clock ▼	☑	1.0	generaldatatype_group ▼

Clock Model Group:

Group Name	Fix Mean	Rate
nucleotide_group		

\+   \−

Data: 103 taxa, 2 partitions; Tip times calibrated in nucleotide_group; Estimate clock rate in generaldatatype_group;

⚙ Generate BEAST File...

Drummond et al. 2006 – Relaxed phylogenetics and dating with confidence

File Edit Help

Partitions | Taxon Sets | Tips | Traits | Sites | Clocks | **Trees** | States | Priors | Operators | MCMC

☑ Link tree prior for all trees

Trees
TreeModel

Tree prior shared by all tree models

Tree Prior: Coalescent: Constant Size ▾

Citation: Kingman JFC (1982) Stoch Proc Appl 13, 235-248 [Constant
Coalescent].
Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002)
Genetics 161, 1307-1320 [Serially Sampled Data].

Tree Model - sequences_11-08-15

◉ Random starting tree

○ UPGMA starting tree

○ User-specified starting tree

Select user-specified tree: no tree loaded ▾

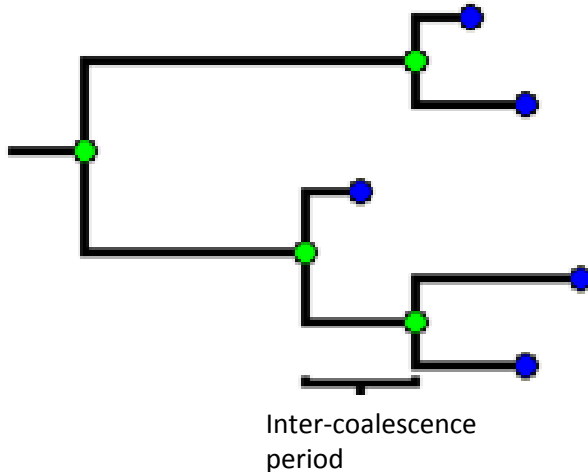Export format for tree: Newick ▾

Import user-specified starting trees from **NEXUS**
format data files using the 'Import Data' menu option.
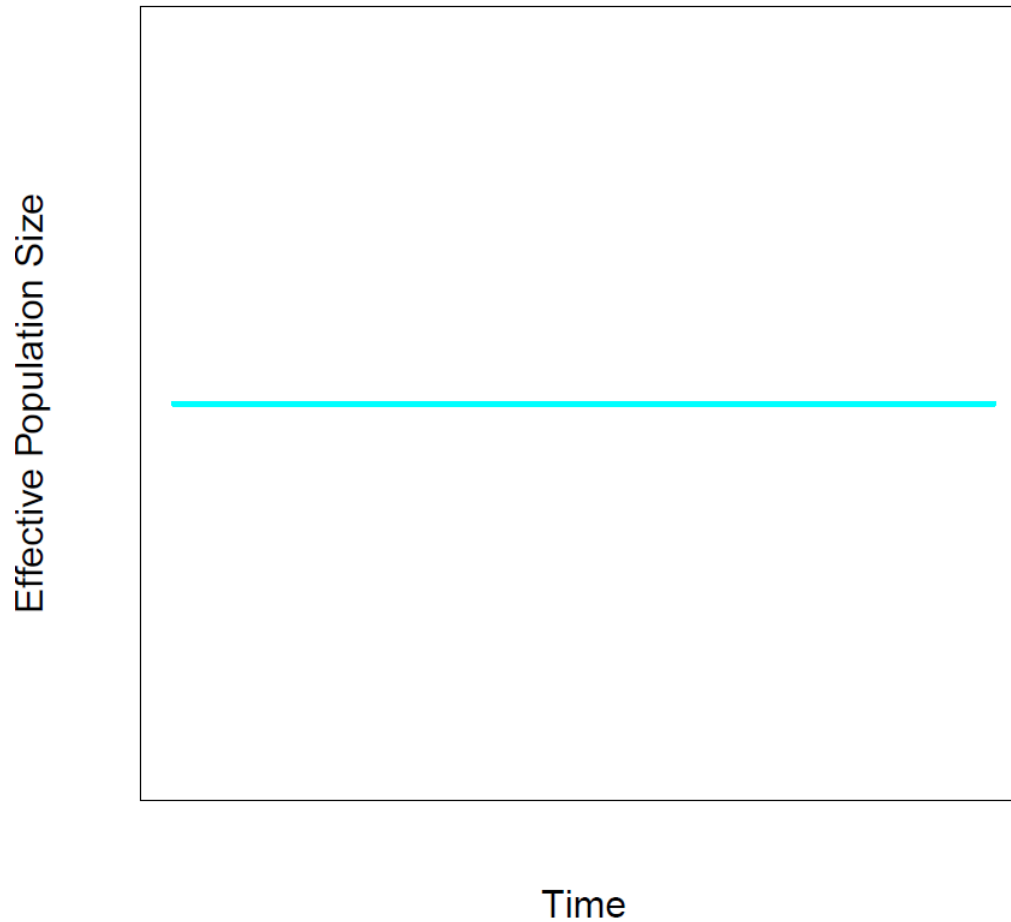Trees must be rooted and strictly bifurcating (binary).

Data: 103 taxa, 2 partitions; Tip times calibrated in nucleotide_group; Estimate clock rate in generaldatatype_group;

⚙ Generate BEAST File...

Inter-coalescence
period
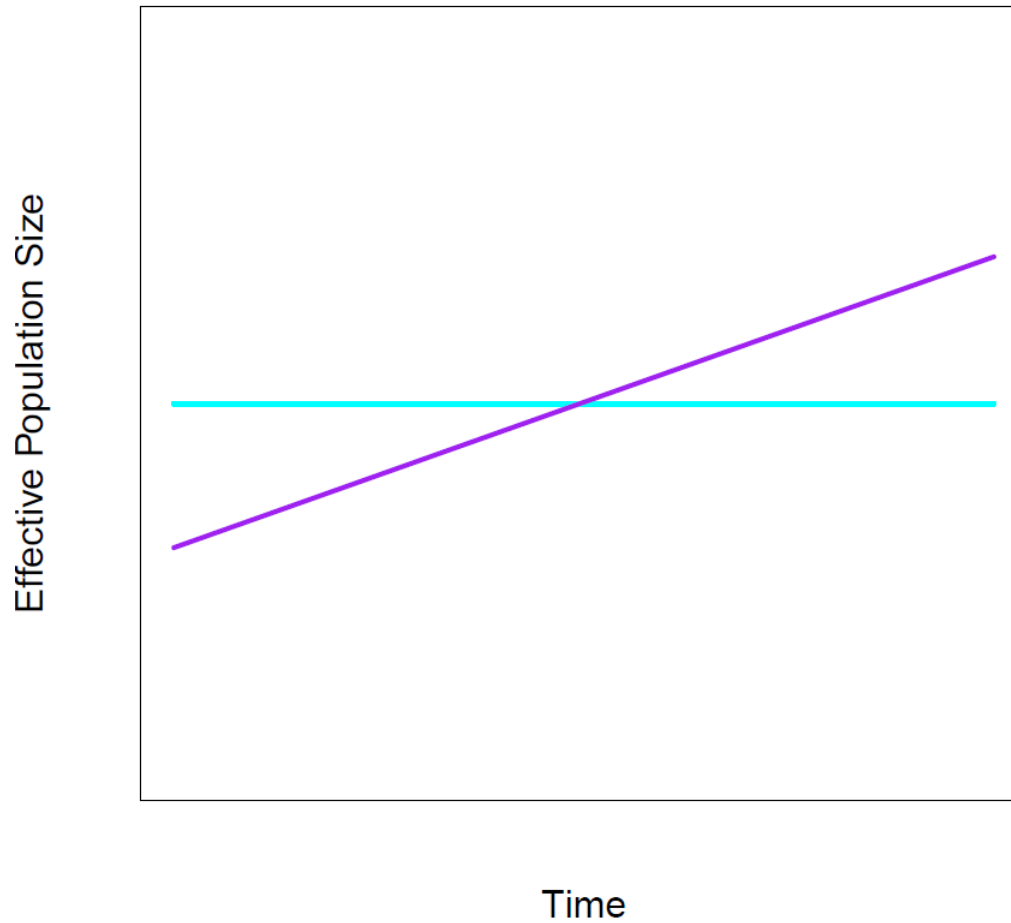
A phylogeny can be thought of as describing a random series of coalescence events.
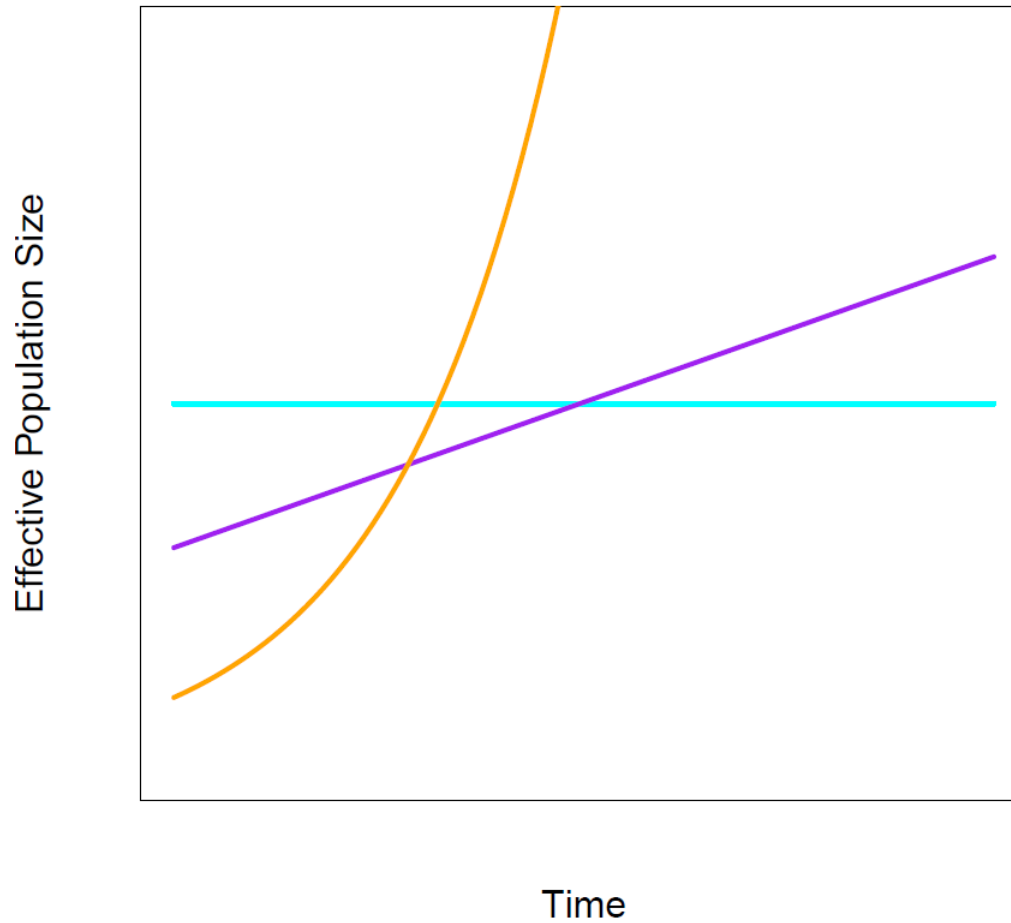
The frequency and timings of coalescence events can be used to infer underlying population dynamics.

Coalescence events are likely to occur more frequently in smaller populations.

37

Volz & Frost 2014 – Sampling through time and phylodynamic inference with coalescent and birth-death models

Ho & Shapiro 2011 – Skyline-plot methods for estimating demographic history from nucleotide sequences

BEAUti

File Edit Help

Partitions | Taxon Sets | Tips | Traits | Sites | Clocks | Trees | States | Priors | Operators | MCMC

Priors for model parameters and statistics:

Parameter	Prior	Bound	Description
kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY transition-transversion parameter
frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	base frequencies
alpha	* Exponential [0.5], initial=0.5	[0, ∞]	gamma shape parameter
ClockModel.uced.mean	? Not yet specified	[0, ∞]	uncorrelated exponential relaxed clock mean
CATTLE.clock.rate	* Approx. Reference Prior, initial=1	[0, ∞]	substitution rate
treeModel.rootHeight	* Using Tree Prior in [306, ∞]	[306, ∞]	root height of the tree
skygrid.precision	* Gamma [0.001, 1000], initial=0.1	[0, ∞]	GMRF Bayesian SkyGrid precision
ClockModel.meanRate	* Indirectly Specified Through Other Parameter	n/a	The mean rate of evolution over the whole tree
ClockModel.covariance	* Indirectly Specified Through Other Parameter	n/a	The covariance in rates of evolution on each lineage with their ancestral lineages
ClockModel.coefficientOfVariation	* Indirectly Specified Through Other Parameter	n/a	The variation in rate of evolution over the whole tree
CATTLE.frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	discrete state frequencies
CATTLE.rates	* Gamma [1, 1], initial=1	[0, ∞]	discrete trait instantaneous transition rates
CATTLE.root.frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	discrete state root frequencies

Link parameters together Link parameters into a hierarchical model | Unlink parameters

* Marked parameters currently have a default prior distribution. You should check that these are appropriate.

Data: 103 taxa, 2 partitions; Tip times calibrated in nucleotide_group; Estimate clock rate in generaldatatype_group;

⚙ Generate BEAST File...

File Edit Help

Partitions | Taxon Sets | Tips | Traits | Sites | Clocks | Trees | States | Priors | Operators | MCMC

Priors for model parameters and statistics:

Parameter	Prior	Bound	Description
ClockModel.uced.mean	? Not yet specified	[0, ∞]	uncorrelated exponential relaxed clock mean
treeModel.rootHeight	* Using Tree Prior in [306, ∞]	[306, ∞]	root height of the tree
ClockModel.meanRate	* Indirectly Specified Through Other Parameter	n/a	The mean rate of evolution over the whole tree
ClockModel.covariance	* Indirectly Specified Through Other Parameter	n/a	The covariance in rates of evolution on each lineage with their ancestral lineages
ClockModel.coefficientOfVariation	* Indirectly Specified Through Other Parameter	n/a	The variation in rate of evolution over the whole tree

Clock Model

Link parameters together Link parameters into a hierarchical model | Unlink parameters

* Marked parameters currently have a default prior distribution. You should check that these are appropriate.

Data: 103 taxa, 2 partitions; Tip times calibrated in nucleotide_group; Estimate clock rate in generaldatatype_group;

✿ Generate BEAST File...

BEAST v1.8.2

BEAST

Bayesian Evolutionary Analysis Sampling Trees
Version v1.8.2, 2002-2015

BEAST XML File: ntial_Skygrid_Cattle-Wildlife.xml   Choose File...

☐ Allow ov

Random number seed: 144104867

Thread pool size: Automatic

☑ Use BEAGLE library if available:

Prefer use of: CPU ▼

☑ Use CPU's S

Prefer precision: Double ▼

Rescaling scheme: Default ▼

☐ Show list of

BEAGLE is a high-performance phyl
additional computational resources
downloaded and installed independ

http://beagle-lib.googlecode

Run   Quit

NZ_1_28-08-15_HKY_Relaxed-Exponential_Skygrid_Cattle-Wildlife.xml

File   Edit   Help

```
Branch rate model used: strictClockBranchRates
Using BEAGLE resource 0: CPU
  with instance flags:  PRECISION_DOUBLE COMPUTATION_SYNCH EIGEN_COMPLEX SCALING_MANUAL SCALERS_RAW VECTOR_NONE THREADING_NONE PROCESSOR_CPU FRAMEW
Ignoring ambiguities in tree likelihood.
With 1 unique site patterns.
Using rescaling scheme : dynamic (rescaling every 100 evaluations)
Creating swap operator for parameter ClockModel.branchRates.categories (weight=10.0)
Optimization Schedule: log

--------------------------------
Creating ctmcScalePrior model.
        If you publish results using this prior, please reference:
                1. Ferreira and Suchard (2008) for the conditional reference prior on CTMC scale parameter prior;
Constructing a cache around likelihood 'null', signal = CATTLE.rates
Likelihood computation is using an auto sizing thread pool.
Creating the MCMC chain:
  chainLength=500000000
  autoOptimize=true
  autoOptimize delayed for 5000000 steps
# BEAST v1.8.2, r6687
# Generated Tue Sep 01 09:03:43 BST 2015 [seed=1441048672824]
state   Posterior       Prior          Likelihood       rootHeight   ClockModel.uced.mean    CATTLE.clock.rate
0       -55636.7037     -2779.8602     -52856.8435      392.682      5E-3                    1.00000          -
50000   -14778.5369     -850.3737      -13928.1632      138.025      3.67378E-4              2.3569E-3        0.21 hours/million states
100000  -14133.7440     -829.6907      -13304.0533      143.620      2.35725E-4              2.95104E-3       0.2 hours/million states
```

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In
NZ_1_13-08-15_GT...	46900000	4690000

+ −

Traces:

Statistic	Mean	ESS	...
posterior	-16267.392	68	R
prior	-3087.49	50	R
likelihood	-13179.902	232	R
treeModel.rootHeight	224.027	83	R
skyline.popSize1	59.582	385	R
skyline.popSize2	144.068	162	R
skyline.popSize3	364.45	170	R
skyline.popSize4	602.427	197	R
skyline.popSize5	683.913	201	R
skyline.popSize6	661.806	166	R
skyline.popSize7	578.859	230	R
skyline.popSize8	381.102	216	R
skyline.popSize9	253.149	115	R
skyline.popSize10	109.631	92	R
skyline.groupSize1	6.231	224	R
skyline.groupSize2	4.844	292	R
skyline.groupSize3	8.599	137	R
skyline.groupSize4	13.005	201	R
skyline.groupSize5	12.908	209	R
skyline.groupSize6	12.842	200	R
skyline.groupSize7	12.304	250	R
skyline.groupSize8	11.608	218	R
skyline.groupSize9	11.241	202	R
skyline.groupSize10	8.417	87	R
kappa	5.499	361	R
frequencies1	0.193	424	R
frequencies2	0.309	424	R
frequencies3	0.311	424	R
frequencies4	0.188	265	R
alpha	13.996	424	R
uced.mean	3.218E-4	64	R
meanRate	2.51E-4	51	R
coefficientOfVariation	0.923	257	R
covariance	-1.79E-5	327	R
LatLongs.precision.col11	0.208	347	R
LatLongs.precision.col12	-0.147	351	R
LatLongs.precision.col21	-0.147	351	R
LatLongs.precision.col22	0.203	378	R
correlation	0.712	337	R
LatLongs.varCovar.LatLon...	10.27	356	R
LatLongs.varCovar.LatLon...	7.434	370	R
LatLongs.varCovar.LatLon...	7.434	370	R
LatLongs.varCovar.LatLon...	10.54	417	R
LatLongs.diffusionRate	1.53	47	R

Data type: (R)eal   (I)nt   (C)at

Tabs: Ⅱ Estimates | ⚠ Marginal Prob Distribution | 📈 Joint-Marginal | 〽 Trace

Y-axis label: NZ_1_13-08-15_GTR_Relaxed-Exp_Skyline_Brownian.log.txt

X-axis label: State
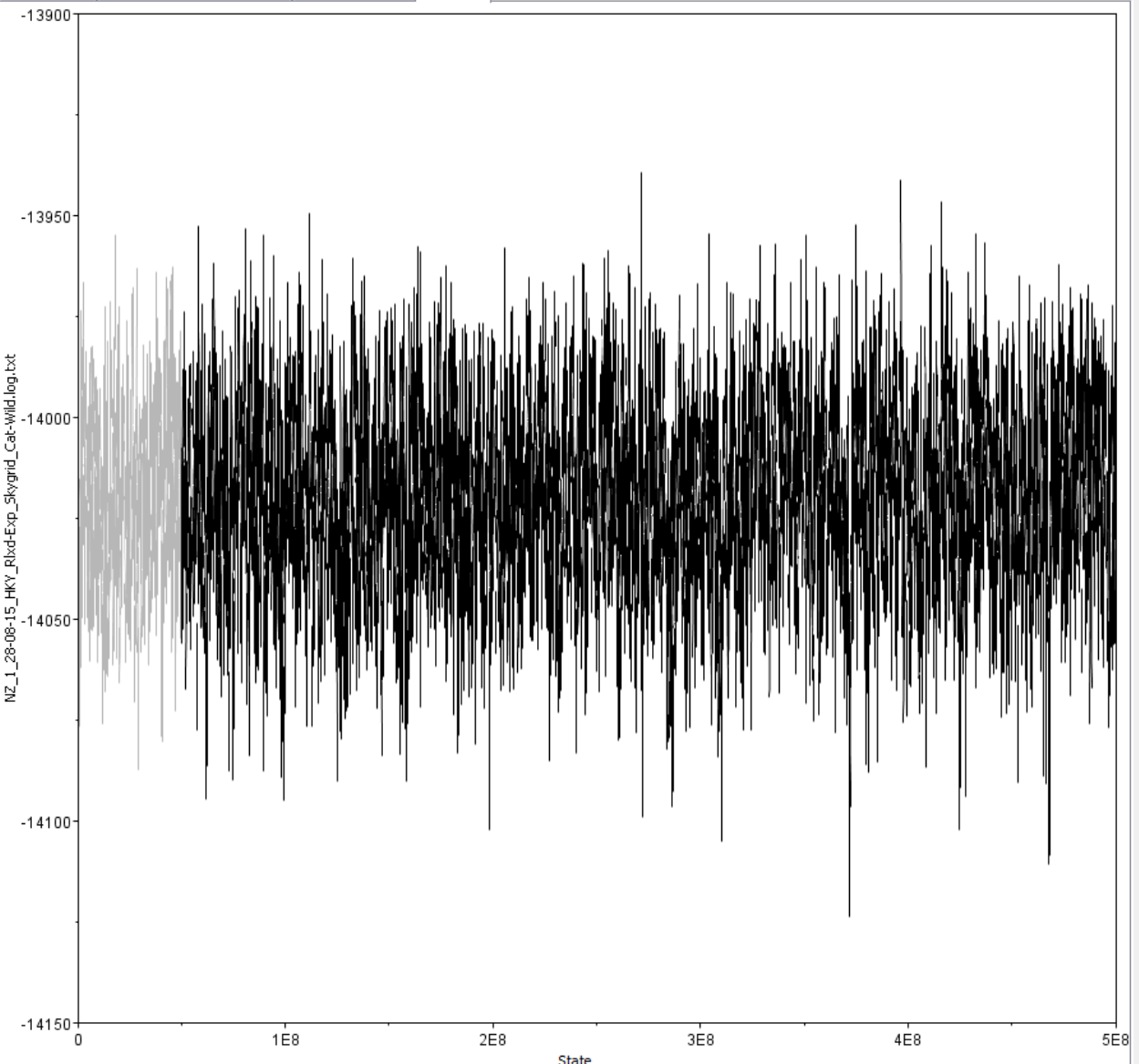
Axes... ☑ Show Burn-in   ☐ Sample only   ☑ Draw line plot   Legend: None ▼   Colour by: Trace ▼   Listen
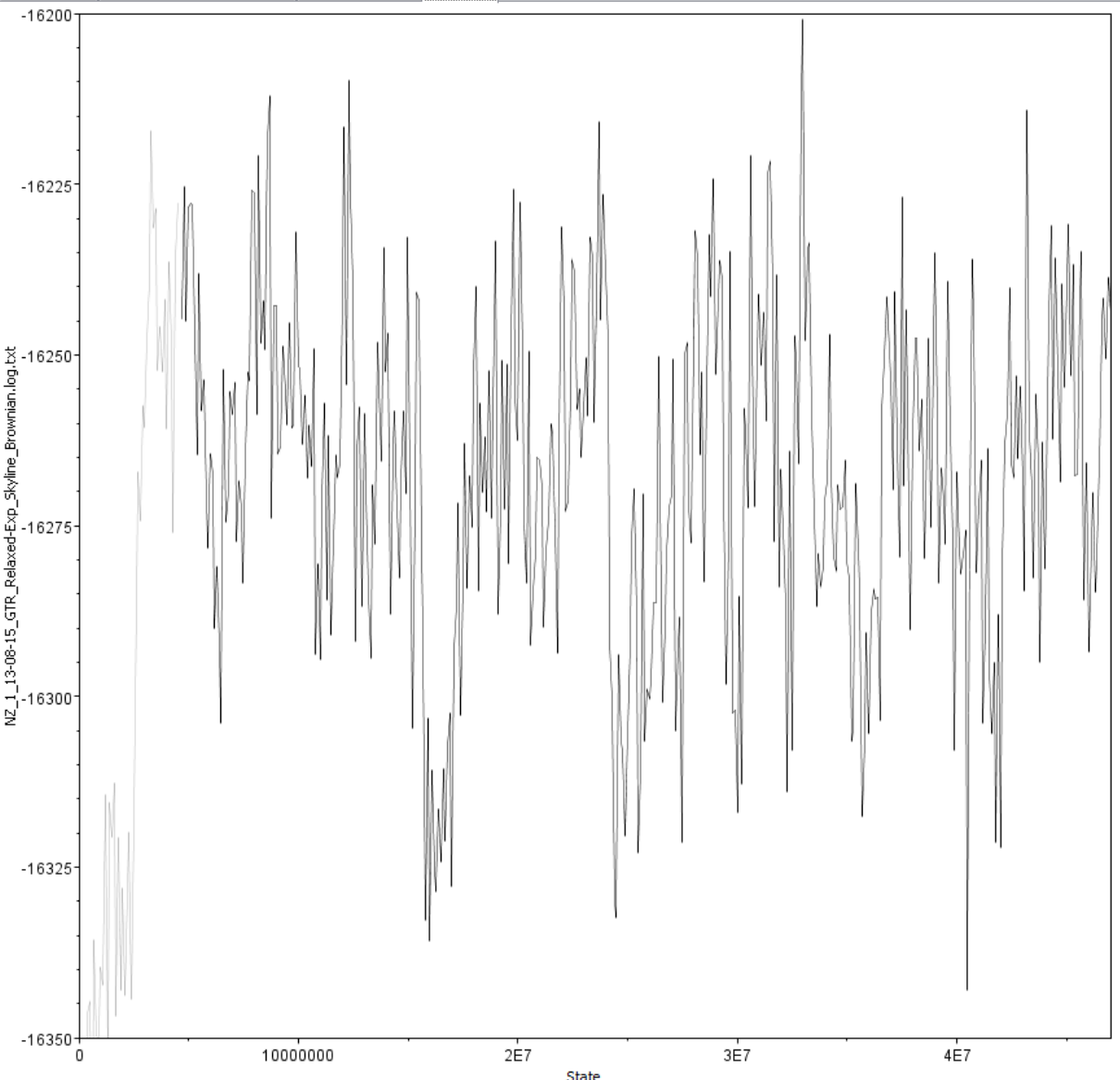
Mean Clock Rate Posterior Distribution

Mean = 0.89
Lower = 0.64
Upper = 1.17
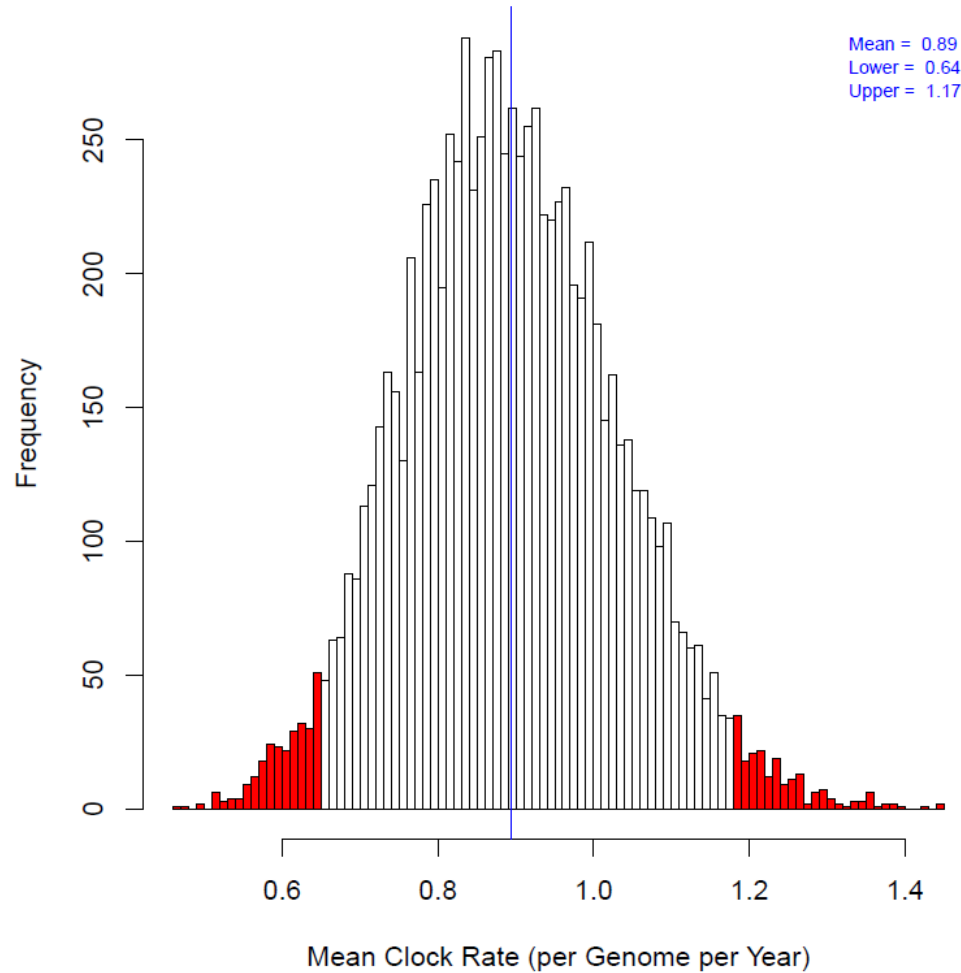
M.bovis Transition Rate Posterior Distributions
Cattle <-> Non-Cattle
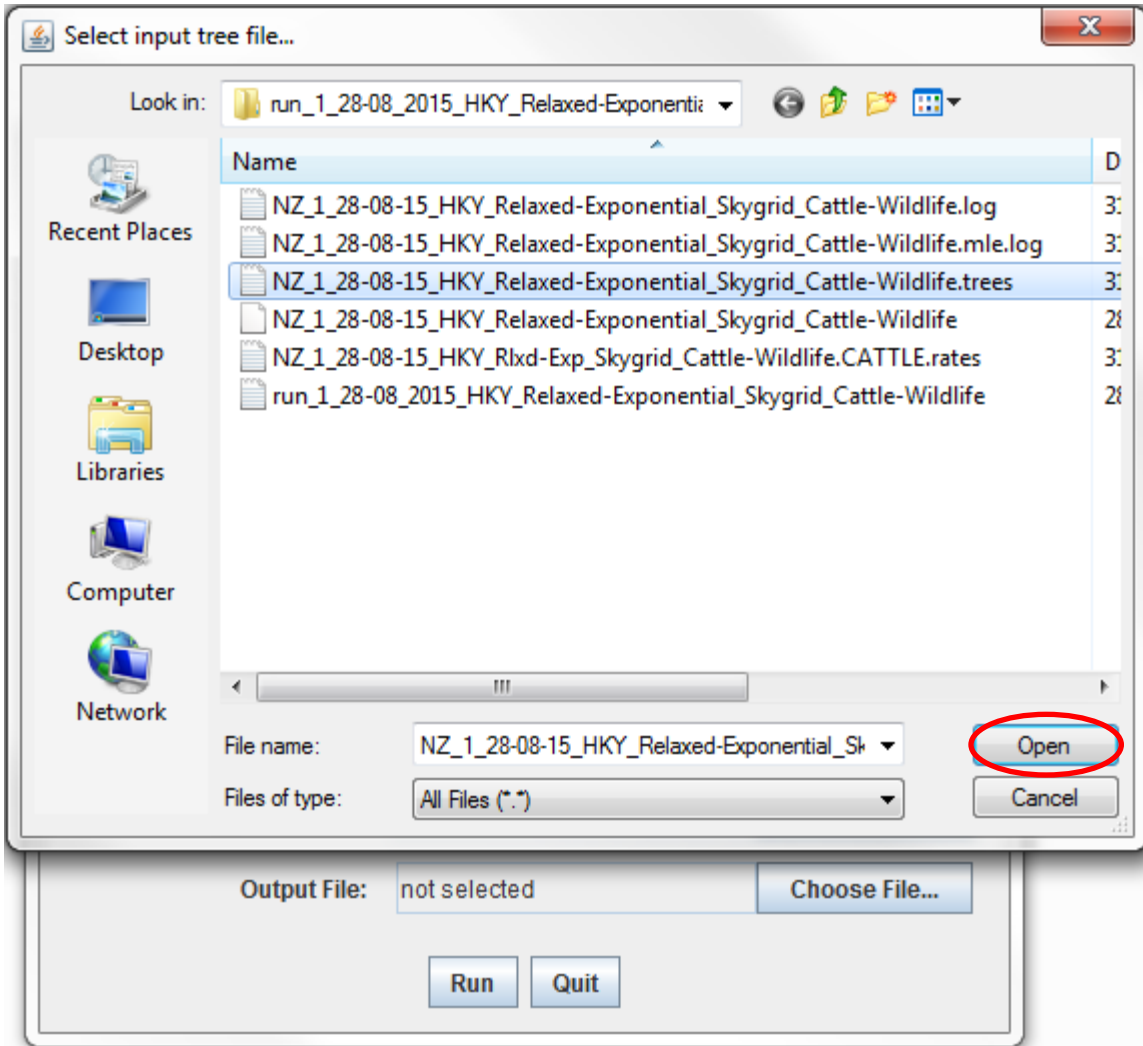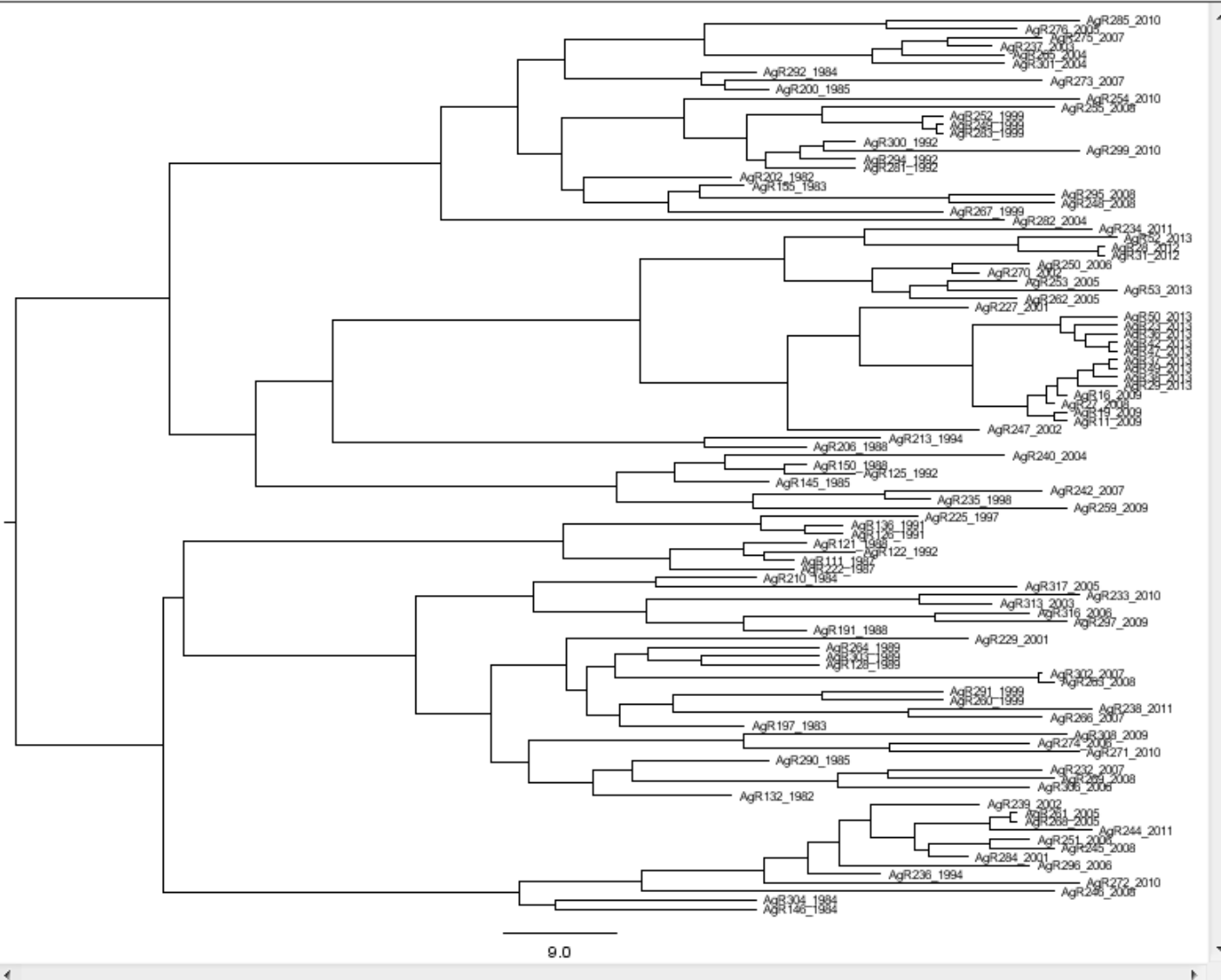
NON-BOVINE
BOVINE

Rowland Kao

Ruth Zadoks

Marian Price-Carter

Paul Livingstone

Maree Joyce

Geoffe de Lisle

Mark Neill

Desmond Collins

Roman Biek

Colette Mair

Patrick Brock   Joseph Hughes   Sibylle Mohr

Nicola Kerbyson

Mikhail Churakov   Liliana Salvador

Daniel Balaz

Sam Lycett   Joanna Morrell   Michael Deason

Thomas Doherty

Thorsten Stefan   Hannah Trewby

Jessica Enright   Mary Parmiter