Online Phylodynamic Inference

January 19, 2020

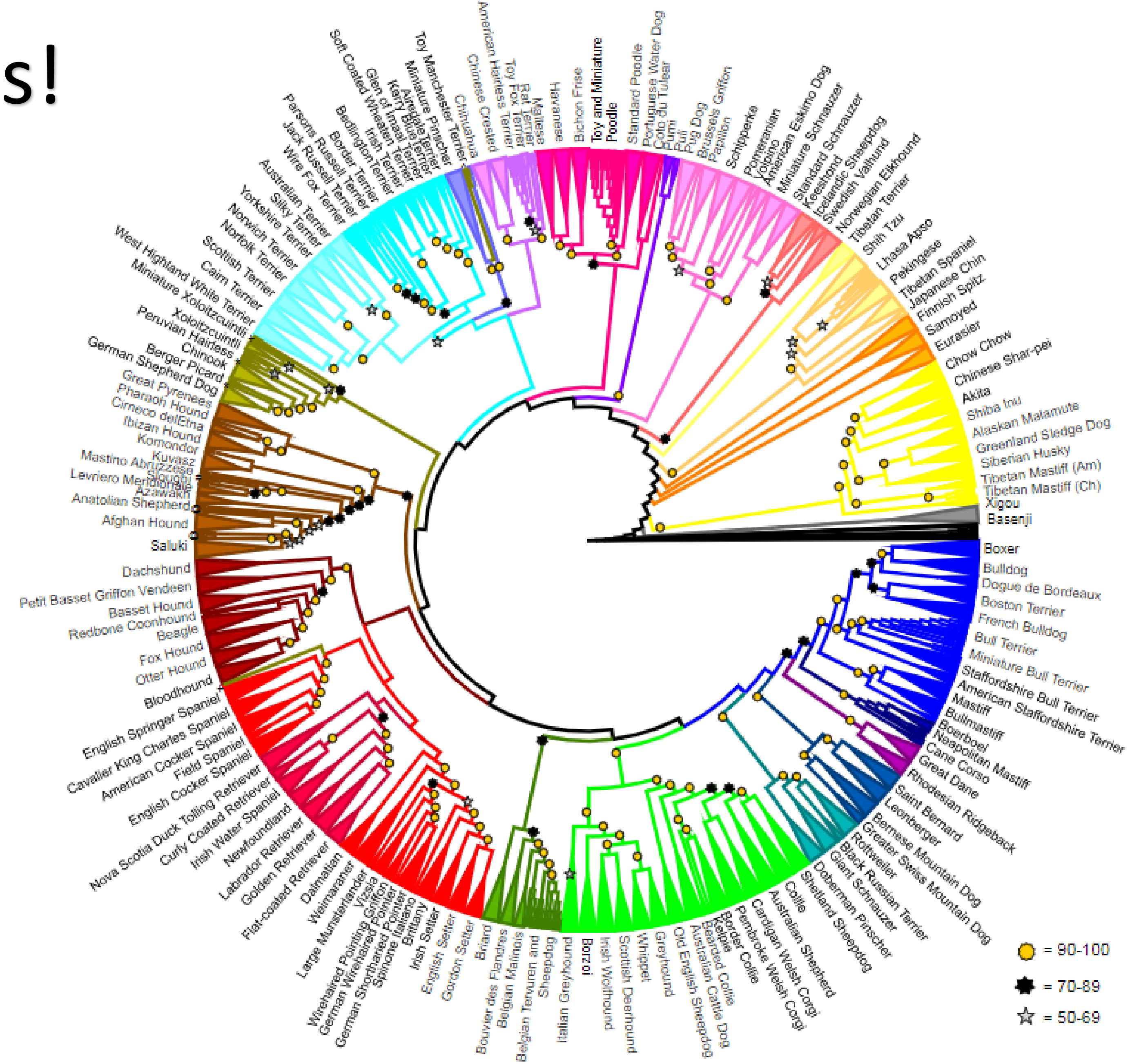
Seth Temple

Based on:

“Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10” (2018)

“Online Bayesian Phylodynamic Inference in BEAST with Application to Epidemic Reconstruction” (2020)

Dog Breeds!



Parker et al. (2017)

Estimating phylogenies for fast-evolving organisms in real time

AACTGAACTG

AATTGAACTG

AACCGAACTG

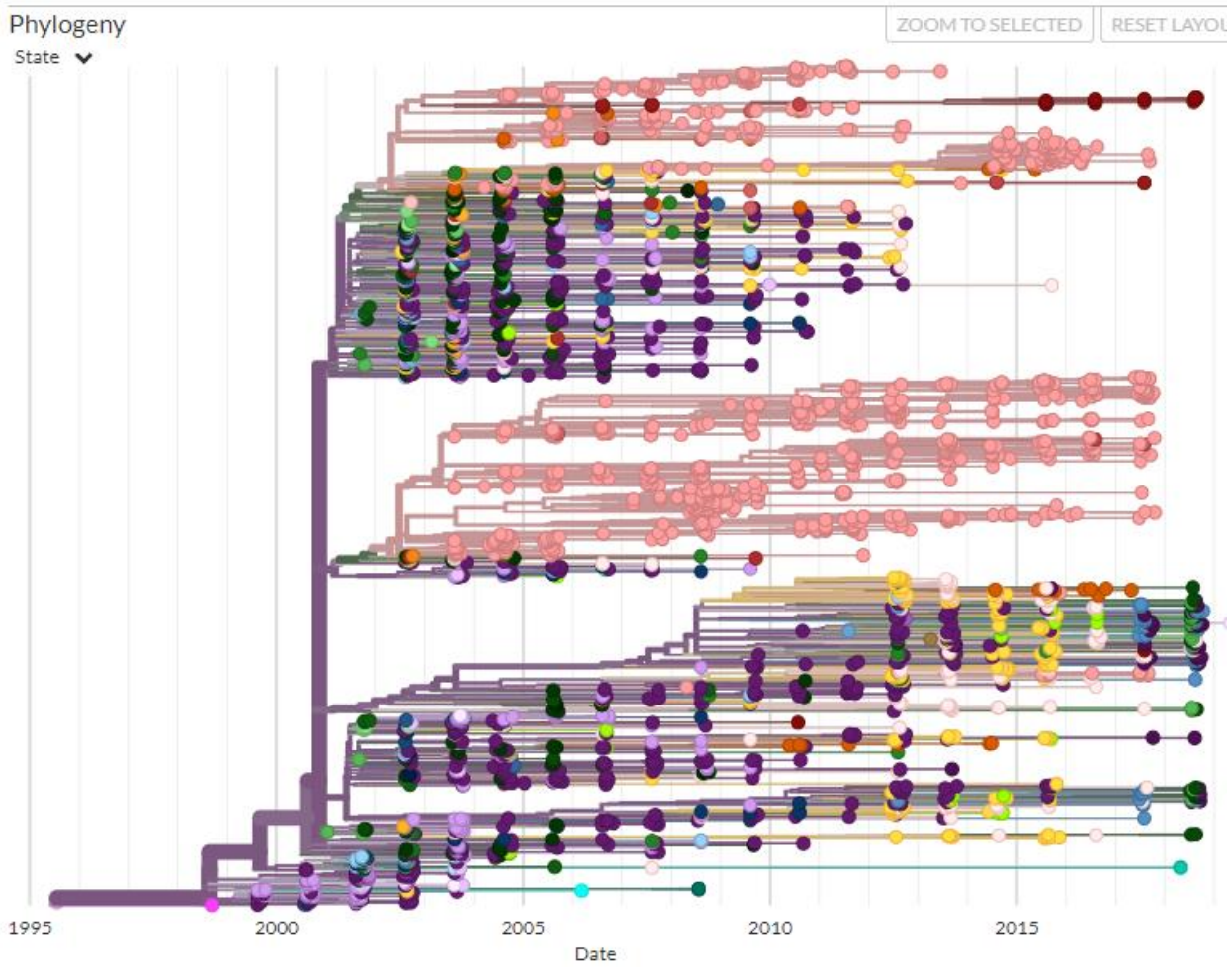
TACTGAACTG

TACTTAACTG

TACGGAACTG

CCCTGAACTG

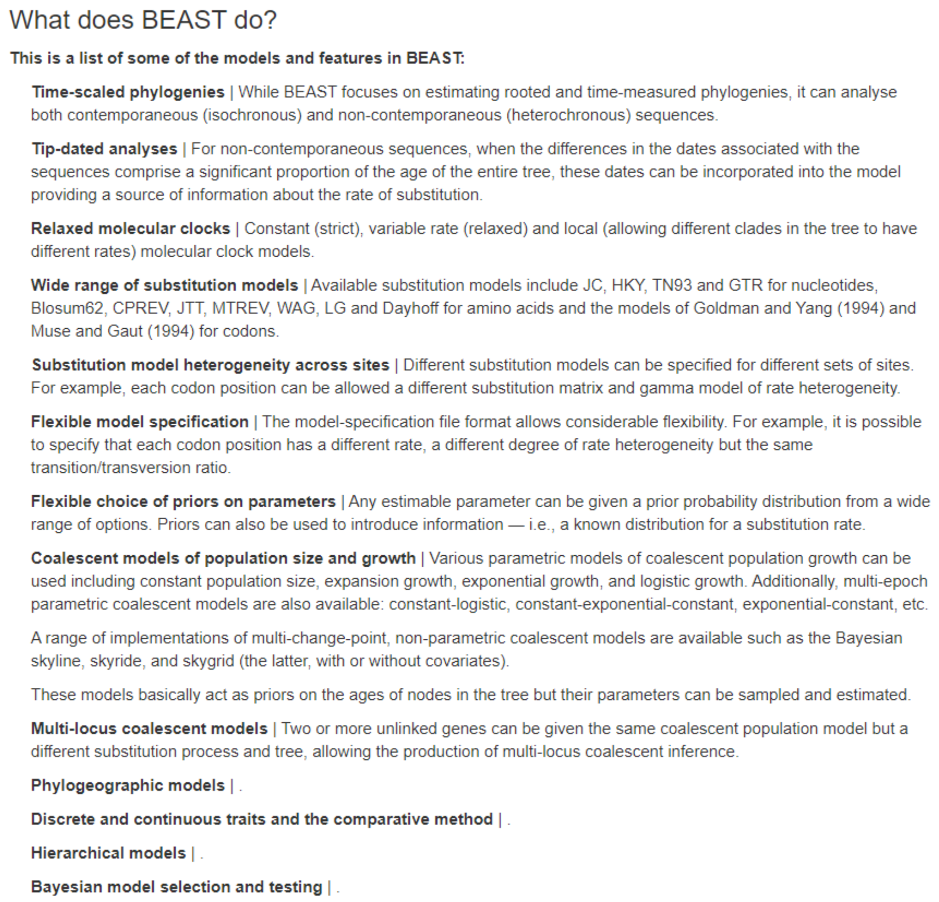
CACTGAACTG



https://nextstrain.org/WNV/NA

BEAST Software Package

- Explores tree space via MCMC and draws sample tree phylogenies that are likely based on available sequence data
- It takes a very long time for the exploration to converge to a reasonable subset of trees!
- Version 10 improves design and efficiency, but most importantly enables new covariates like continuous geographic variables and case counts



What does BEAST do?

This is a list of some of the models and features in BEAST:

Time-scaled phylogenies | While BEAST focuses on estimating rooted and time-measured phylogenies, it can analyse both contemporaneous (isochronous) and non-contemporaneous (heterochronous) sequences.

Tip-dated analyses | For non-contemporaneous sequences, when the differences in the dates associated with the sequences comprise a significant proportion of the age of the entire tree, these dates can be incorporated into the model providing a source of information about the rate of substitution.

Relaxed molecular clocks | Constant (strict), variable rate (relaxed) and local (allowing different clades in the tree to have different rates) molecular clock models.

Wide range of substitution models | Available substitution models include JC, HKY, TN93 and GTR for nucleotides, Blosum62, CPREV, JTT, MTREV, WAG, LG and Dayhoff for amino acids and the models of Goldman and Yang (1994) and Muse and Gaut (1994) for codons.

Substitution model heterogeneity across sites | Different substitution models can be specified for different sets of sites. For example, each codon position can be allowed a different substitution matrix and gamma model of rate heterogeneity.

Flexible model specification | The model-specification file format allows considerable flexibility. For example, it is possible to specify that each codon position has a different rate, a different degree of rate heterogeneity but the same transition/transversion ratio.

Flexible choice of priors on parameters | Any estimable parameter can be given a prior probability distribution from a wide range of options. Priors can also be used to introduce information — i.e., a known distribution for a substitution rate.

Coalescent models of population size and growth | Various parametric models of coalescent population growth can be used including constant population size, expansion growth, exponential growth, and logistic growth. Additionally, multi-epoch parametric coalescent models are also available: constant-logistic, constant-exponential-constant, exponential-constant, etc.

A range of implementations of multi-change-point, non-parametric coalescent models are available such as the Bayesian skyline, skyride, and skygrid (the latter, with or without covariates).

These models basically act as priors on the ages of nodes in the tree but their parameters can be sampled and estimated.

Multi-locus coalescent models | Two or more unlinked genes can be given the same coalescent population model but a different substitution process and tree, allowing the production of multi-locus coalescent inference.

Phylogeographic models | .

Discrete and continuous traits and the comparative method | .

Hierarchical models | .

Bayesian model selection and testing | .

Share what you know about a concept in the word list.
Or, express interest in an explanation from a peer.

- Splits/clades
- Substitution models
- Jukes and Cantor ‘69 model
- Felsenstein ‘81 model
- Transition vs transmission
- Site-specific rates
- Codon models
- Molecular clock
- Phylogeography
- Split frequencies

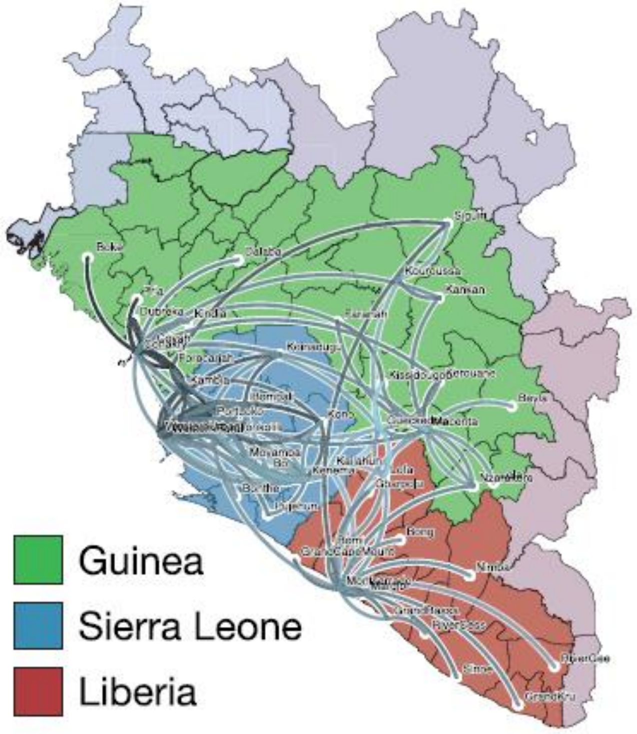
- Bayesian perspective
- Bayes factor
- How to move through tree space
- Priors for lengths
- Priors for proportions
- Sliding window proposal
- Stepping-stone
- Burn-in or ESS (effective sample)
- Stationary distribution

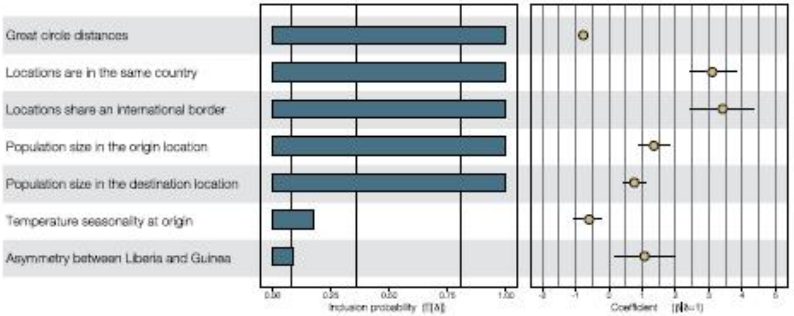
See lectures 76-79 from phyloseminar.org for more information.

Suchard *et al.* (2018) Questions

“[C]ontinuous models have most frequently been applied to diffusion on a geographical landscape with traits representing coordinates and the phylogeny reconstructing the epidemiological process within the host population”.

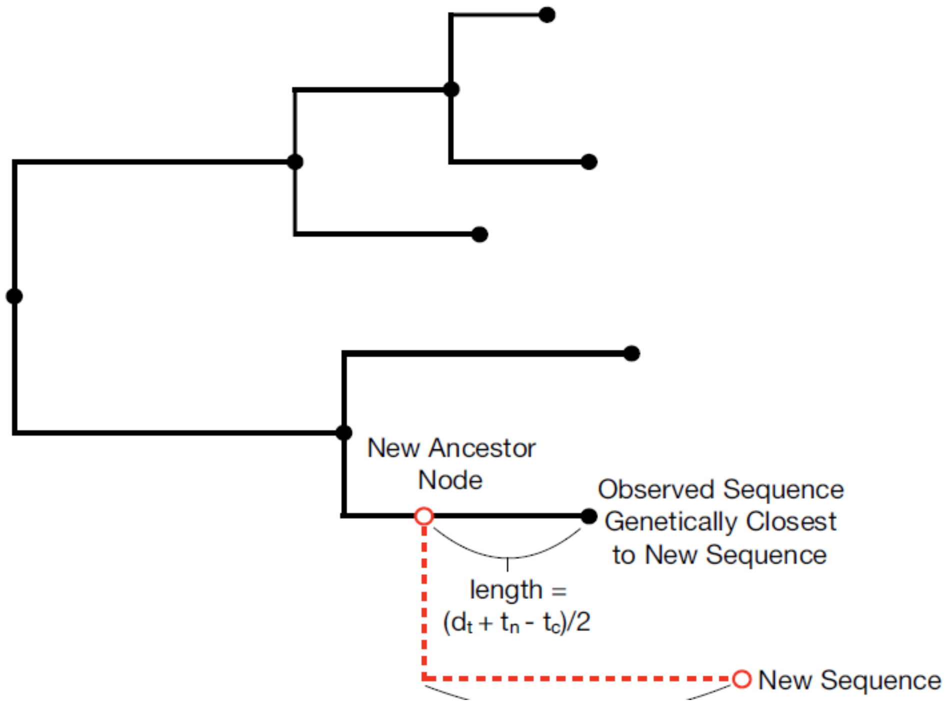
- When may incorporating geographic location assist in a phylogenetic reconstruction of epidemic spread? When may it be inappropriate?
- How does West Africa differ from other geographic regions?





The online method

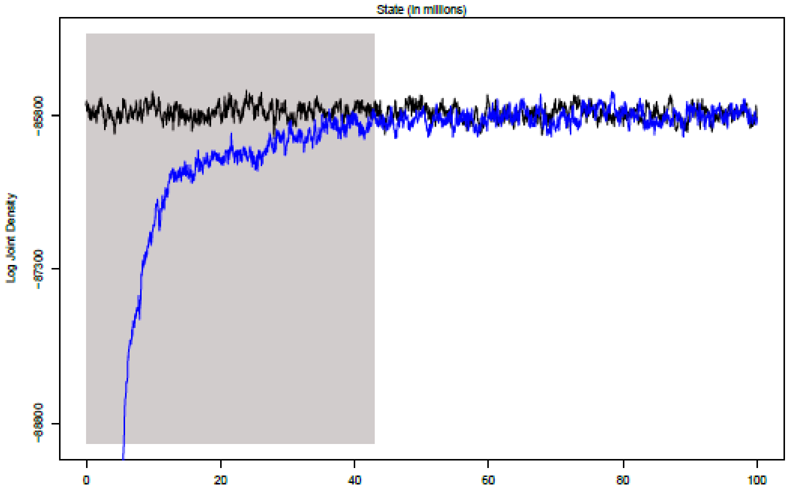
- Samples from the current posterior have already learned from the data and are good initial values for a new run of tree fitting given new data.
- Place the new sequences near to the most similar sequence already in the tree.

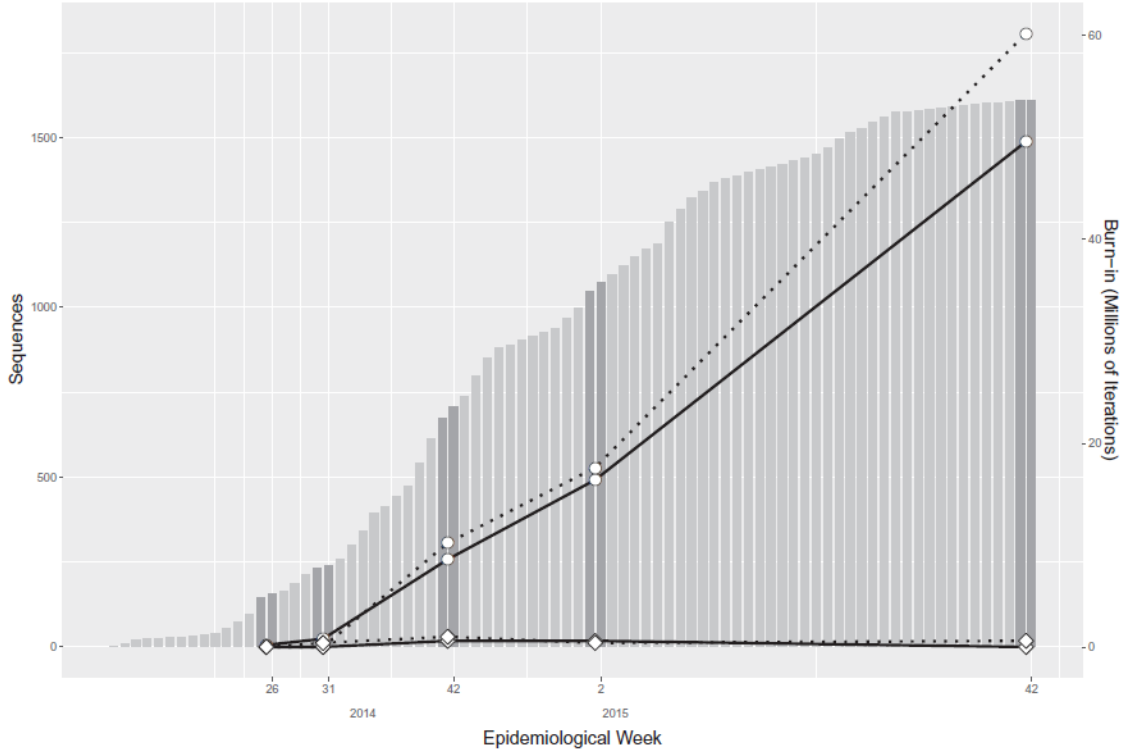


Online method reduces burn-in time

“Our online inference approach leads to higher computation time savings as the complexity of the data increases, with up to 600 h being saved on average on a modern multi-core processor.”

Gill *et al.* (2020)





Gill *et al.* (2020) Questions

- “[T]here is no need to completely terminate the chain … if it is interrupted to incorporate new data because the … chain can be resumed after the interruption, and the … simulation for the expanded data set can be started as an independent process.” Do we have any concerns about this interruption handling?
- When should we add new sequences and refit the model?
- The paper discusses Dunn tests about time savings in the language of *p*-values. How do we discuss significance in the context of runtime efficiency?