NC STATE UNIVERSITY

Phylodynamics in a world of rapidly adapting pathogens

David Rasmussen
Dept. of Entomology and Plant Pathology
Bioinformatics Research Center
NC State University

7th Annual Disease Modeling Symposium
Bellevue, WA
April 15th, 2019

Introduction

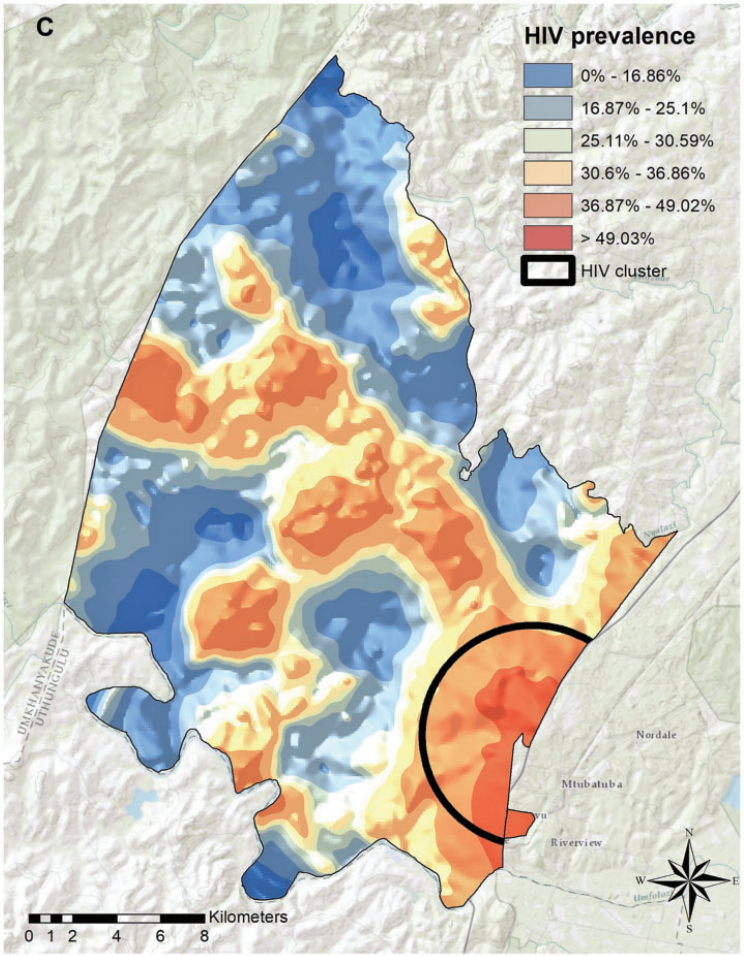
- **Phylodynamics:** the study of how ecological and evolutionary processes act or interact to shape the phylogenetic history of pathogens (Grenfell *et al.*, 2004; Volz *et al.*, 2014)

Introduction

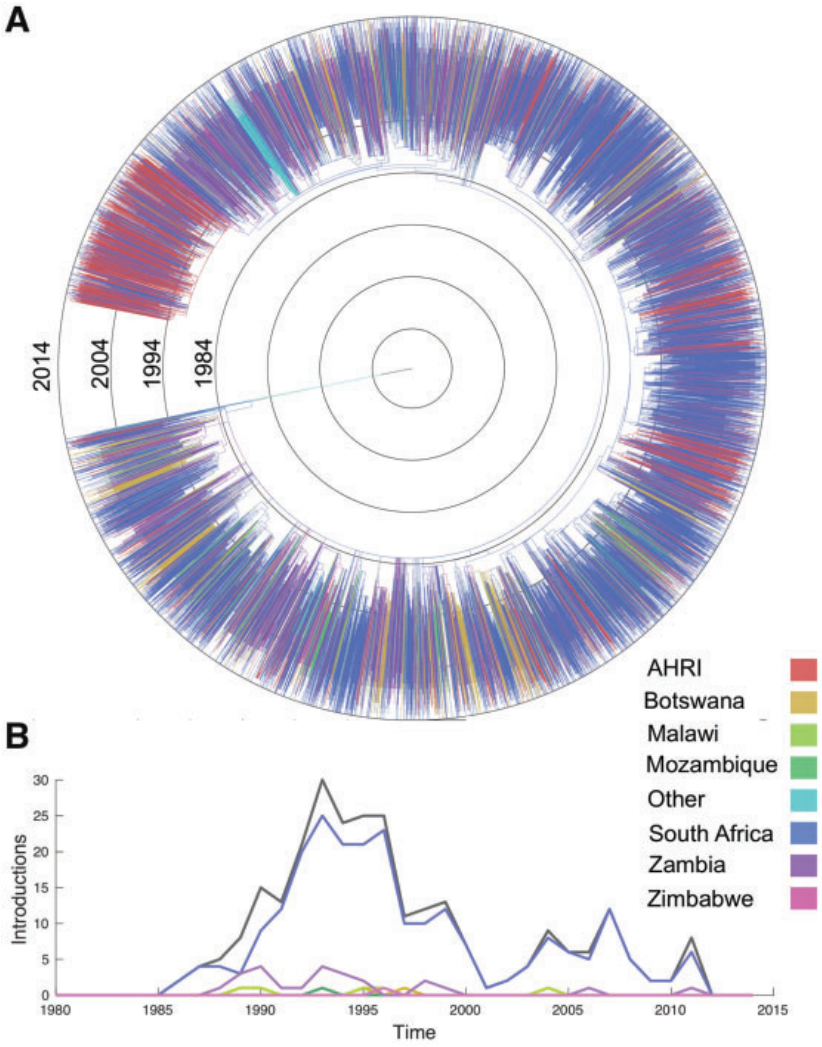
- **Phylodynamics:** the study of how ecological and evolutionary processes act or interact to shape the phylogenetic history of pathogens (Grenfell *et al.*, 2004; Volz *et al.*, 2014)

- **Phylodynamic inference:** the statistical practice of inferring ecological/evolutionary dynamics from phylogenetic trees.

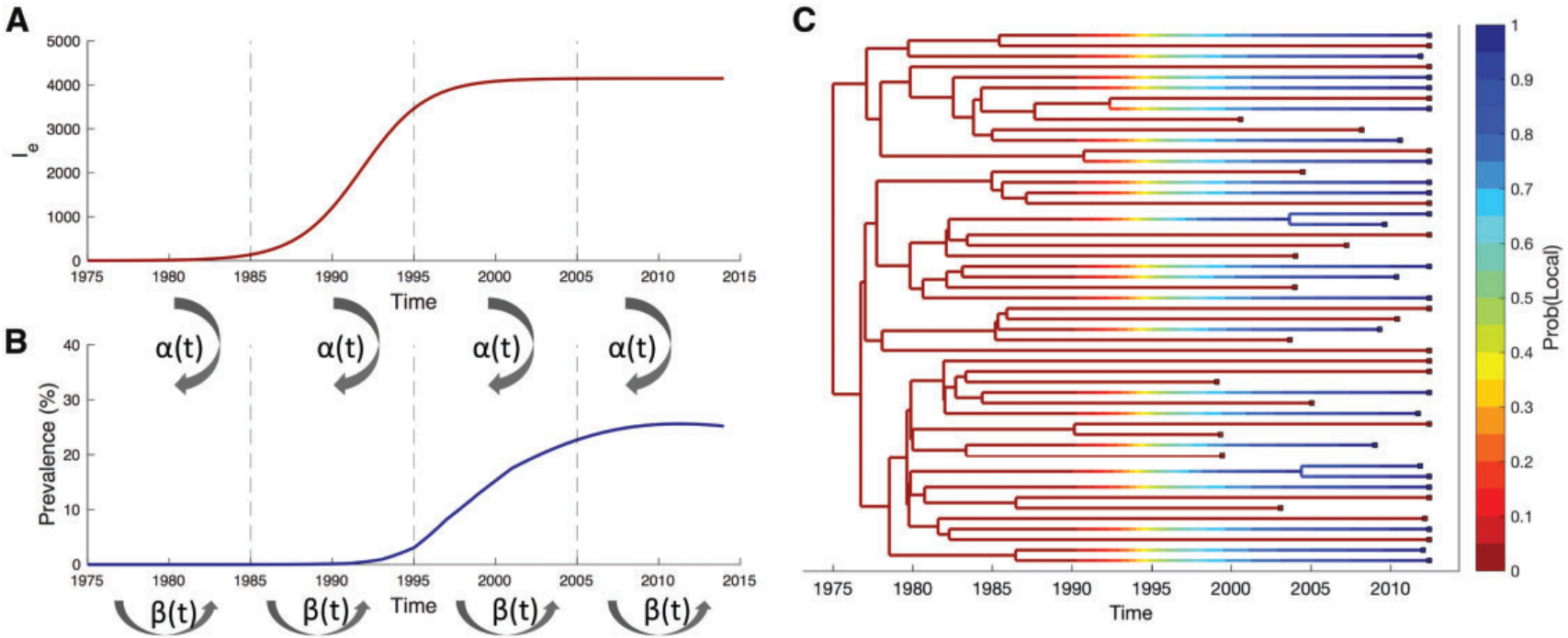
HIV in rural KwaZulu-Natal



HIV in rural KwaZulu-Natal

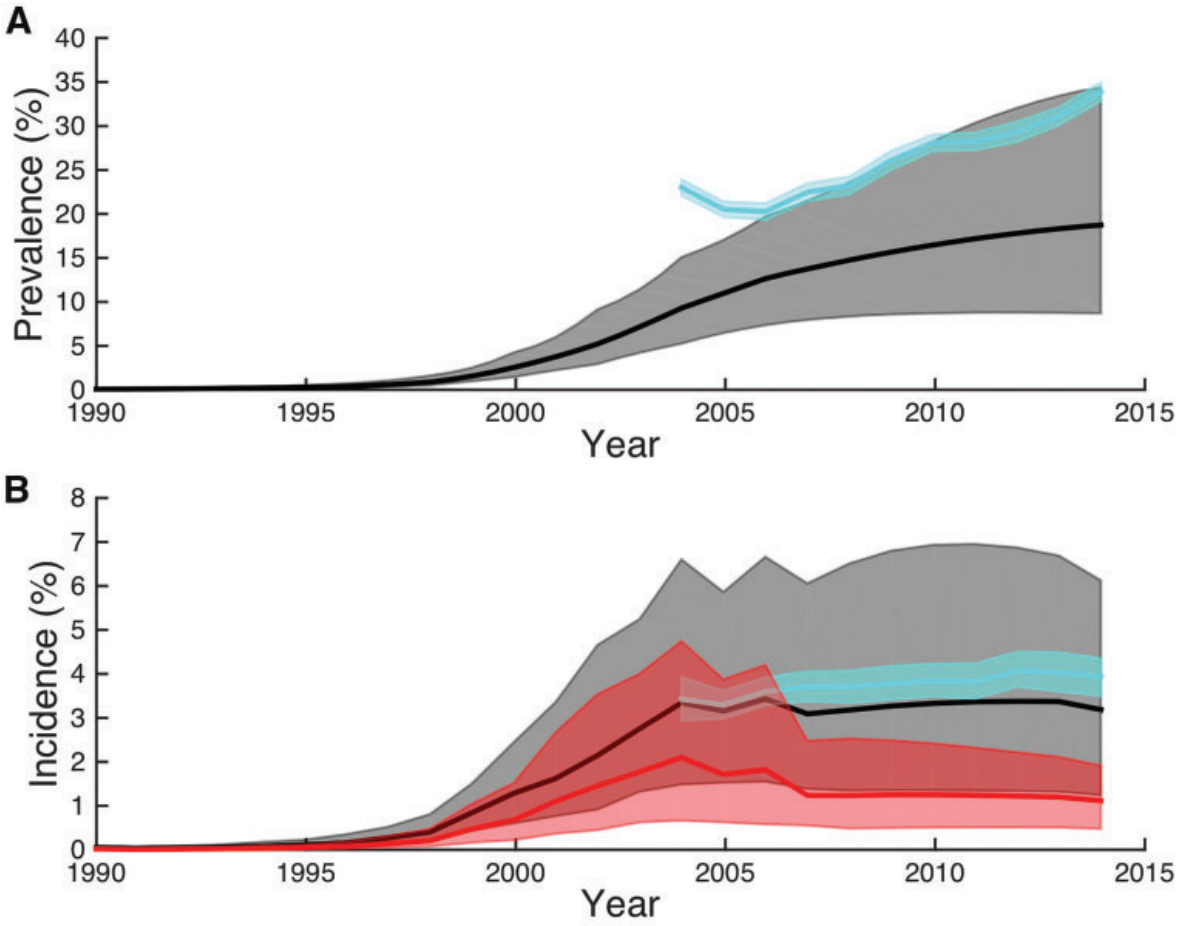


Phylodynamic model for HIV in KZN



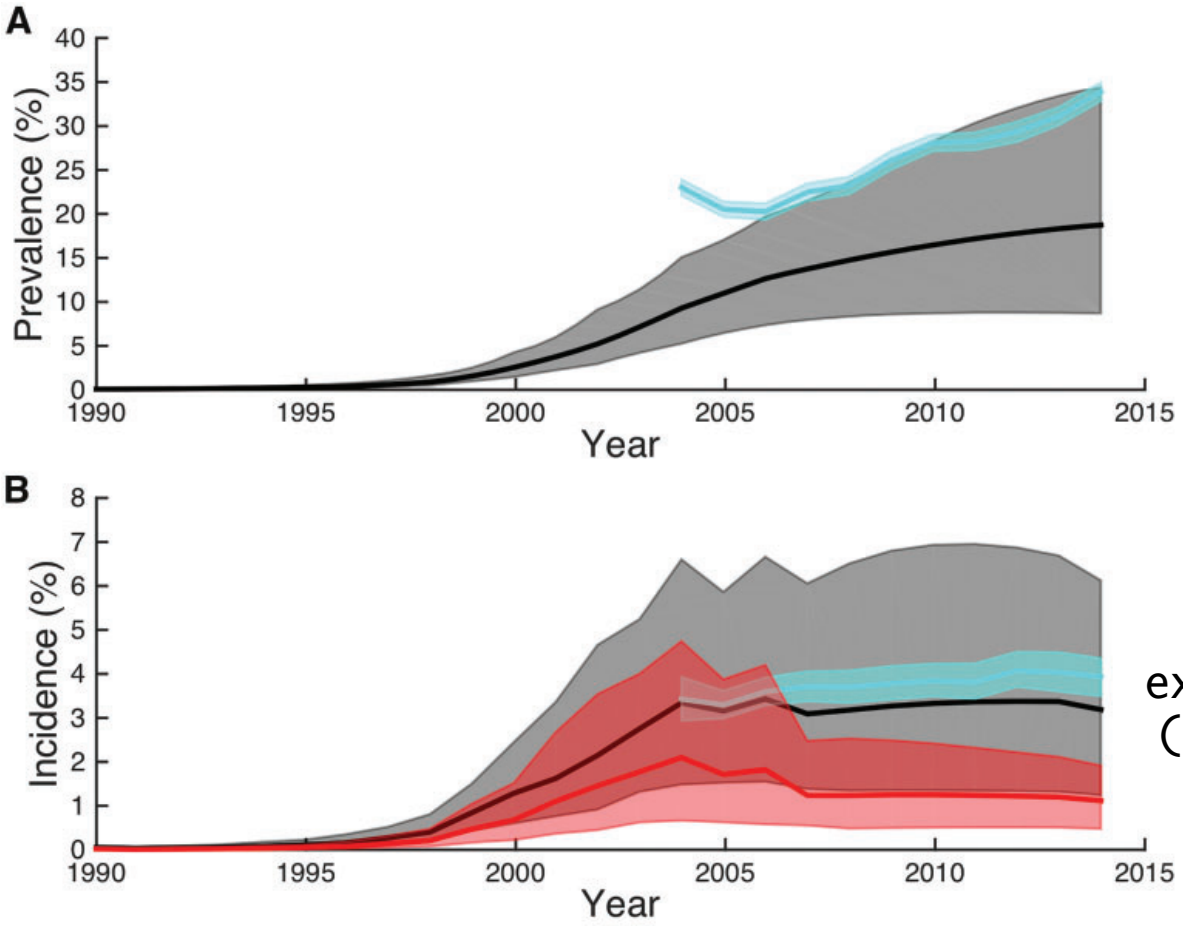
Rasmussen *et al.*, Virus Evolution (2018)

Phylodynamic estimates for HIV in KZN



Rasmussen *et al.*, Virus Evolution (2018)

Phylodynamic estimates for HIV in KZN



Fraction external: 35% (CI: 20-60%)

Rasmussen *et al.*, Virus Evolution (2018)

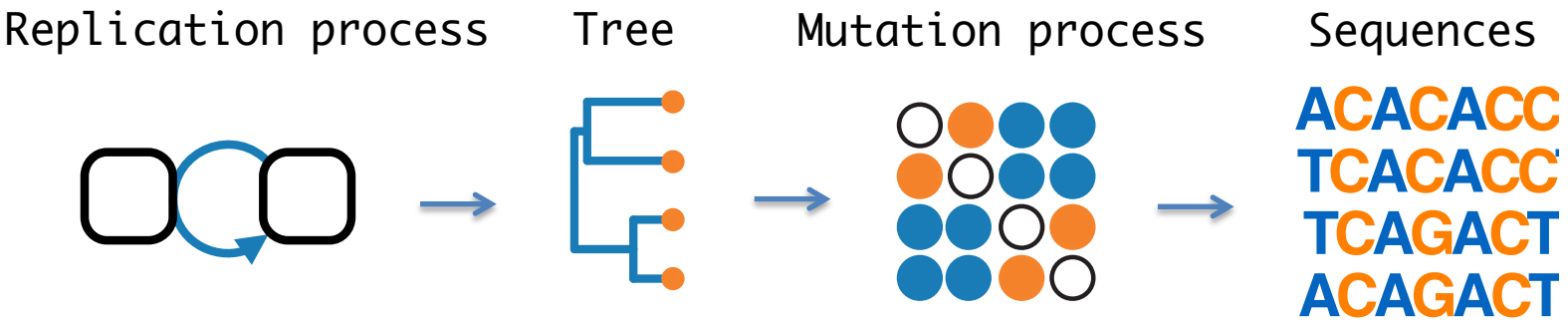
What current phylodynamic methods do well:

- Accurately reconstruct historical and recent epidemic dynamics
- Accommodate geographic and other forms of host population structure – allowing us to infer sources of transmission
- Account for incomplete or biased sampling of sequence data

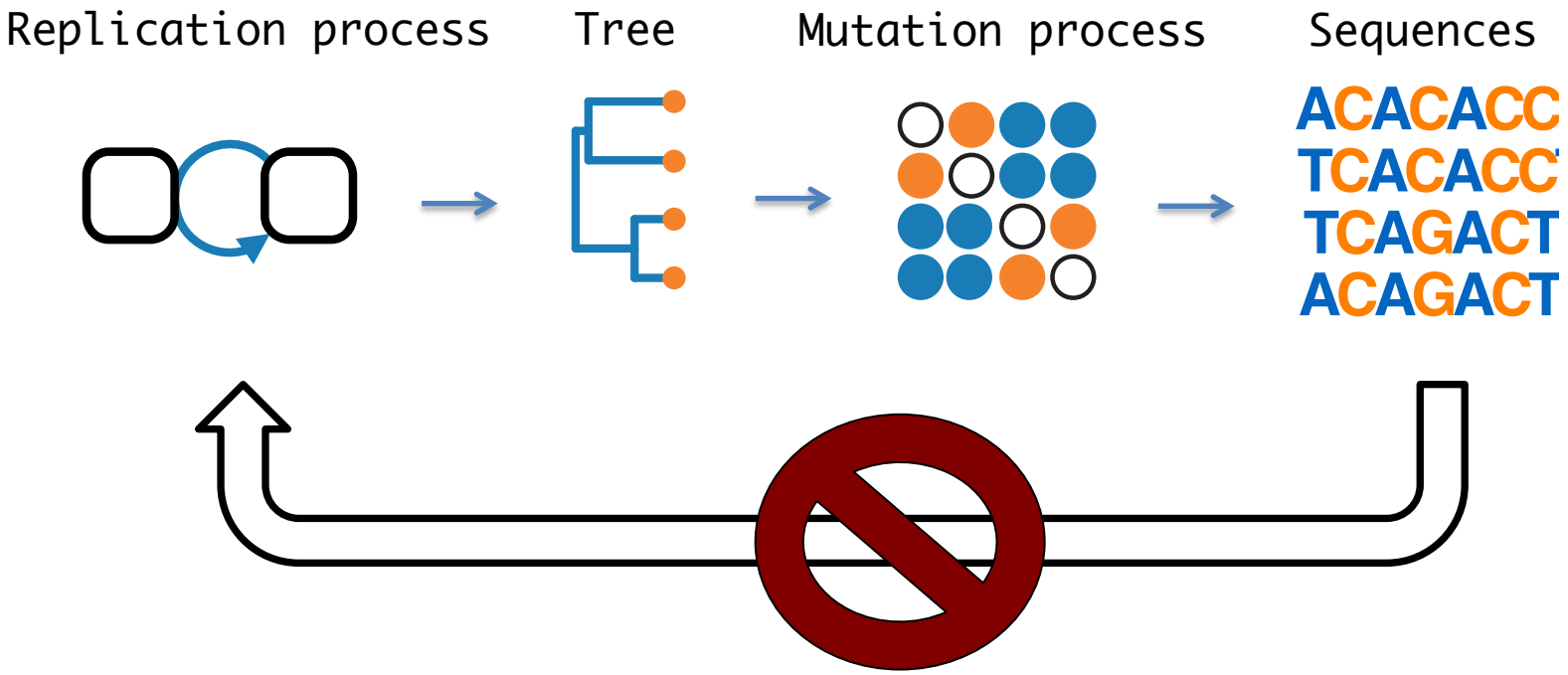
What current methods do not do well:

- Consider non-neutral genetic variation in pathogen fitness and therefore differences in the epidemic potential of pathogen lineages.

The big assumption of phylogenetic models



The big assumption of phylogenetic models

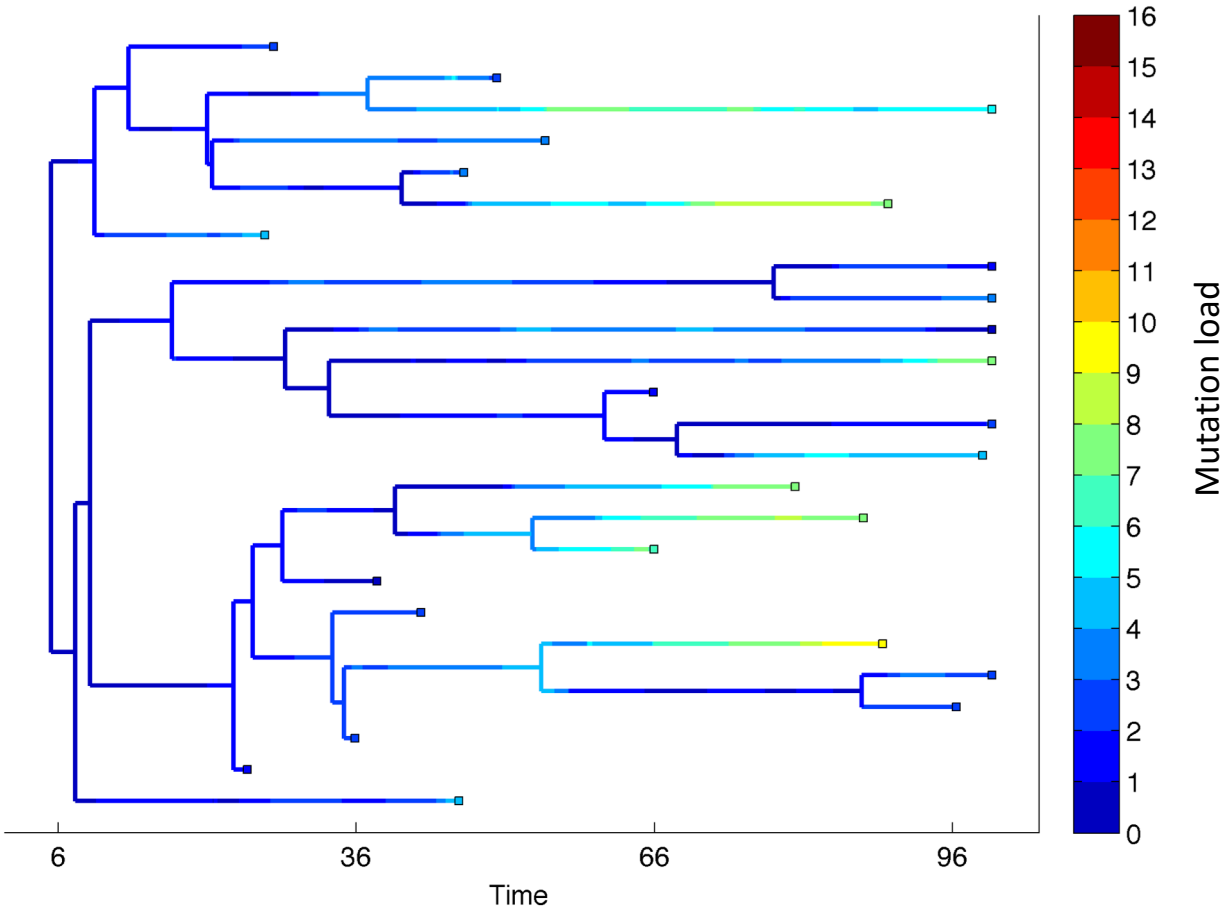


The independence assumption

- This independence assumption allows us to factor the joint likelihood of a tree $\mathcal{T}$ and sequence data $\mathcal{S}$ into terms we can easily compute:

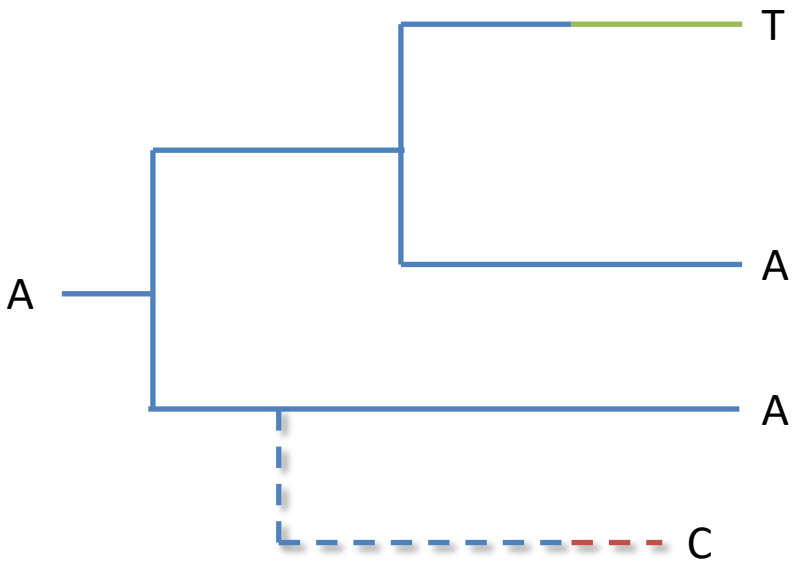
$$L(\mathcal{S}, \mathcal{T}|\mu, \theta) = L(\mathcal{S}|\mathcal{T}, \mu)p(\mathcal{T}|\theta).$$

Example: deleterious mutation load



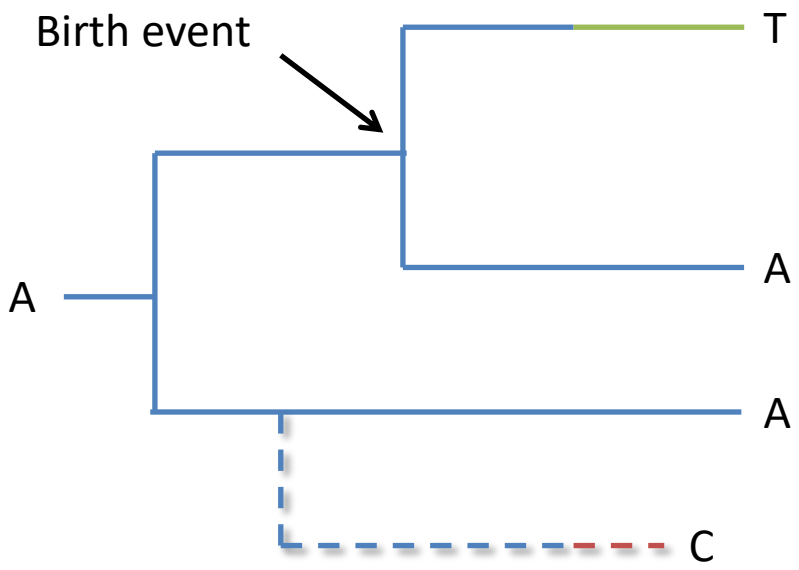
The multi-type birth-death model

- Allows us to compute the joint likelihood that a tree and genotype data at a single loci evolved under a non-neutral model (Stadler & Bonhoeffer, 2013).

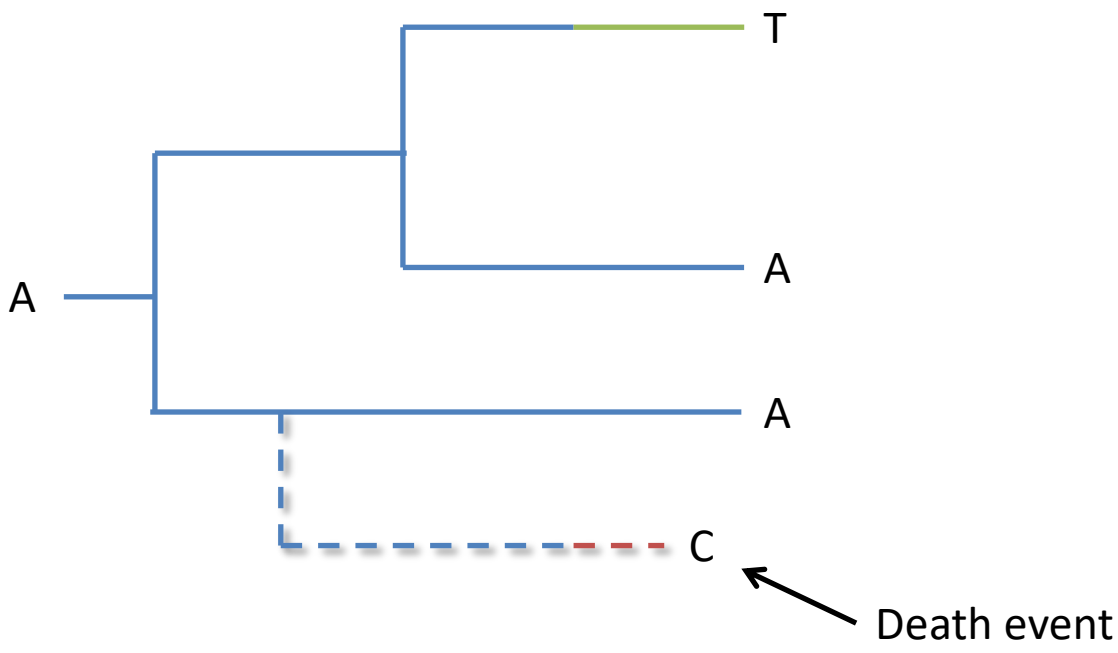


Tanja Stadler

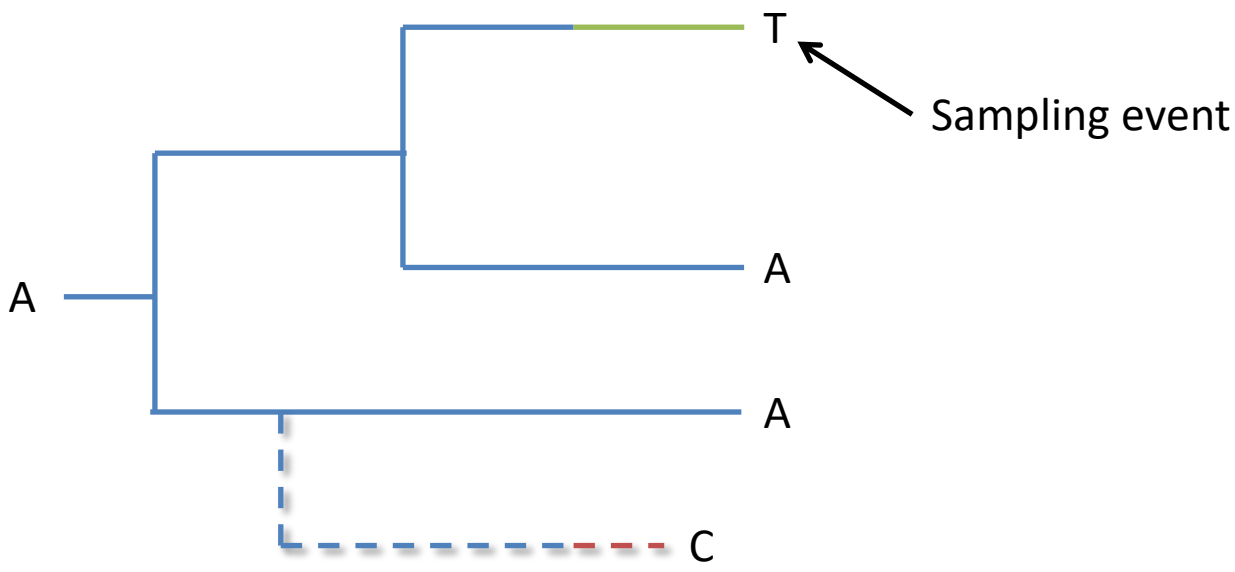
The multi-type birth-death model



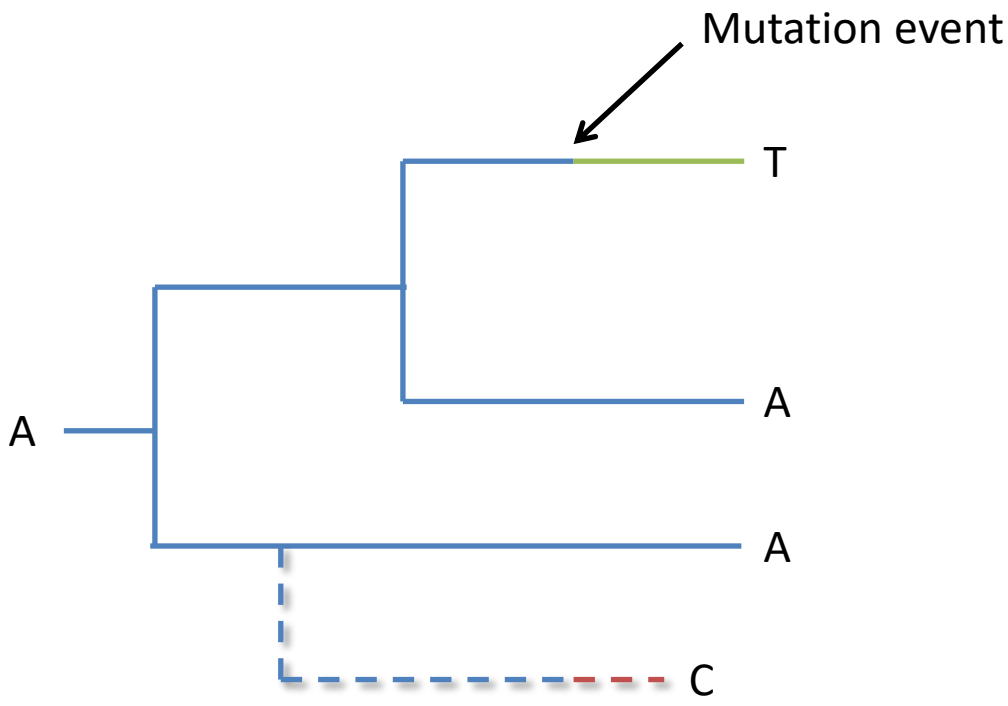
The multi-type birth-death model



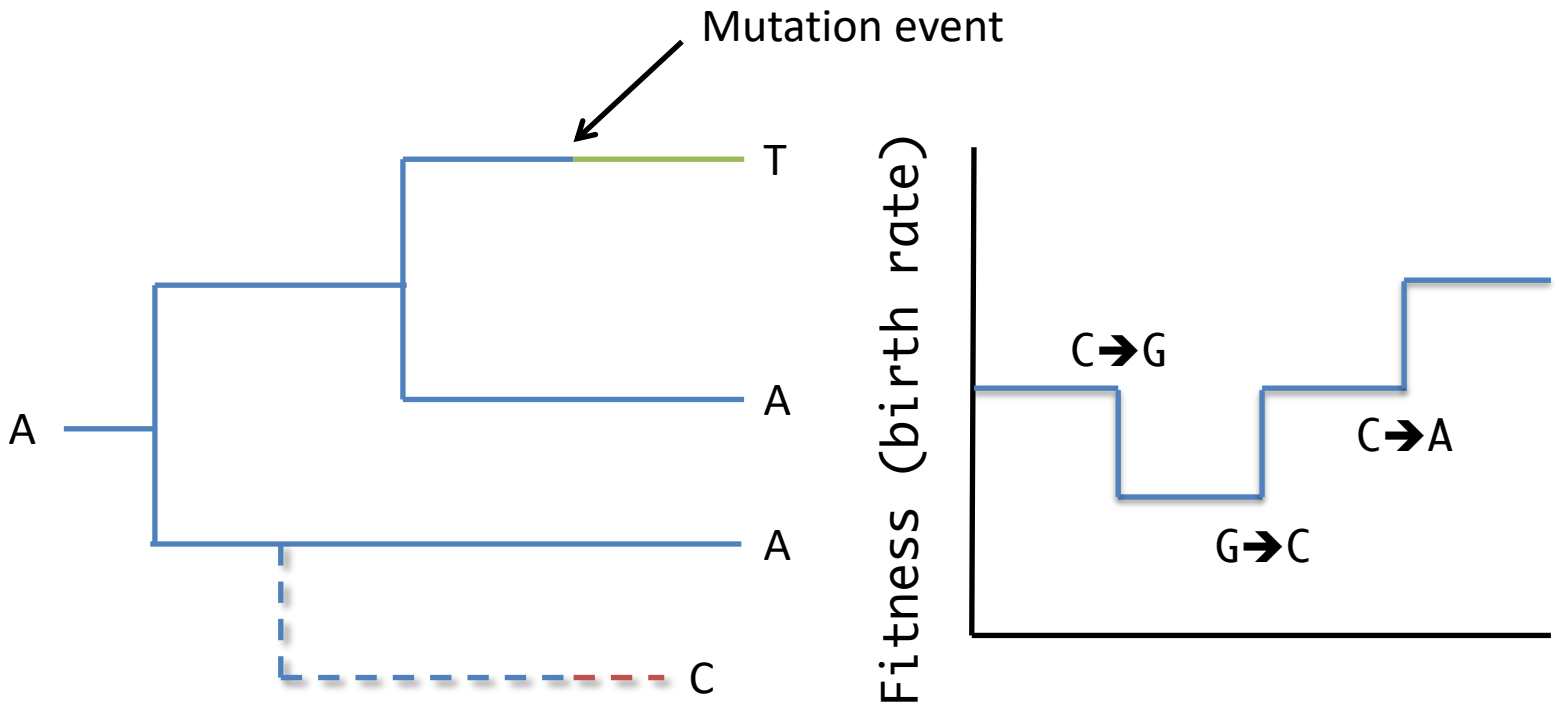
The multi-type birth-death model



The multi-type birth-death model

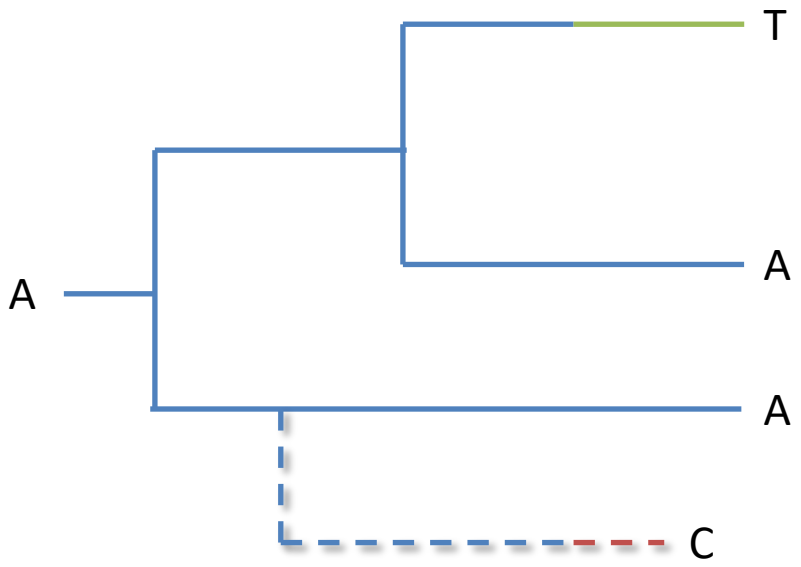


The multi-type birth-death model



The multi-type birth-death model

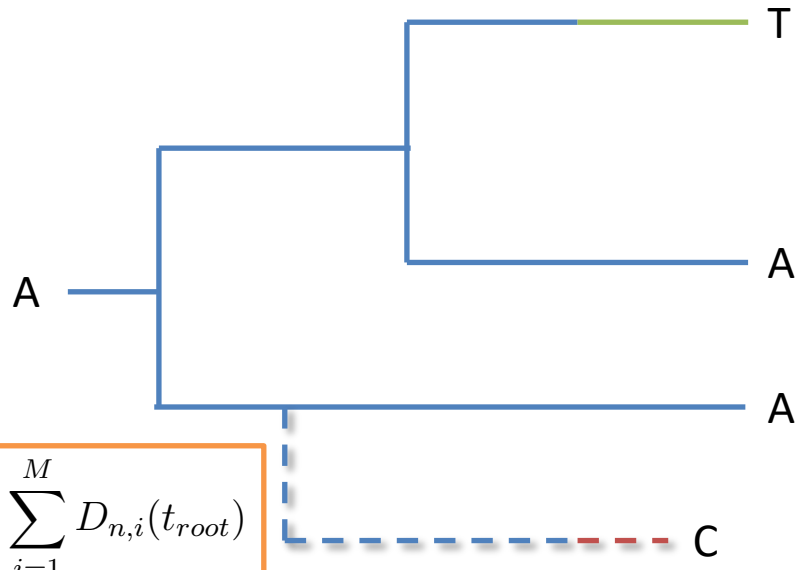
- At a single evolving site, we can compute the joint likelihood of a tree and the 'sequence' at each tip using a multi-type birth-death model (Stadler & Bonhoeffer, 2013).



$$\frac{d}{dt}D_{n,i}(t) = -(\lambda_i + \sum_{j=1}^{M}\gamma_{i,j} + d_i)D_{n,i}(t) + 2\lambda_i E_i(t)D_{n,i}(t) + \sum_{j=1}^{M}\gamma_{i,j}D_{n,j}(t).$$

The multi-type birth-death model

- At a single evolving site, we can compute the joint likelihood of a tree and the 'sequence' at each tip using a multi-type birth-death model (Stadler & Bonhoeffer, 2013).



$$L(\mathcal{S},\mathcal{T}|\theta) = \sum_{i=1}^{M} D_{n,i}(t_{root})$$

$$\frac{d}{dt}D_{n,i}(t) = -(\lambda_i + \sum_{j=1}^{M}\gamma_{i,j} + d_i)D_{n,i}(t) + 2\lambda_i E_i(t) D_{n,i}(t) + \sum_{j=1}^{M}\gamma_{i,j} D_{n,j}(t).$$

The inevitable problem…

- We need to track **all possible genotypes** in the state space of the model, which increases exponentially with the number of sites L (e.g. 4^L for a nucleotide model)

- MTBD becomes prohibitively computationally expensive for anything more than just a few evolving sites.

The marginal fitness birth-death model

- We track molecular evolution at each site, computing the marginal site probability ω that a site is in particular state.

The marginal fitness birth-death model

- We track molecular evolution at each site, computing the marginal site probability $\boldsymbol{\omega}$ that a site is in particular state.

- We approximate the probability of a lineage being in any genotype based on the marginal site probabilities; e.g.:

$$\hat{\omega}_{n,\mathrm{ACT}} = \omega_{n,1,\mathrm{A}} \times \omega_{n,2,\mathrm{C}} \times \omega_{n,3,\mathrm{T}}$$

The marginal fitness birth-death model

- We track molecular evolution at each site, computing the marginal site probability $\boldsymbol{\omega}$ that a site is in particular state.

- We approximate the probability of a lineage being in any genotype based on the marginal site probabilities; e.g.:

$$\hat{\omega}_{n,\mathrm{ACT}} = \omega_{n,1,\mathrm{A}} \times \omega_{n,2,\mathrm{C}} \times \omega_{n,3,\mathrm{T}}$$

- We then sum, or marginalize, over the fitness of each genotype weighted by its approximate genotype probability to compute the expected fitness of a lineage:

$$f_n \approx \sum_{g \in \mathcal{G}} f_g \hat{\omega}_{n,g}$$

The marginal fitness birth-death model

- We now have a new system of ODEs for tracking the probability $D_{n,k,i}$ that a lineage evolved exactly as observed at each site k:

$$\begin{aligned}\frac{d}{dt}D_{n,k,i}(t) = &-\left(\hat{f}_{n,k,i}\lambda_0 + \sum_{j=1}^{M}\gamma_{i,j} + d\right)D_{n,k,i}(t) \\ &+ 2\hat{f}_{n,k,i}\lambda_0 E_u(t) D_{n,k,i}(t) \\ &+ \sum_{j=1}^{M}\gamma_{i,j}D_{n,k,j}(t).\end{aligned}$$

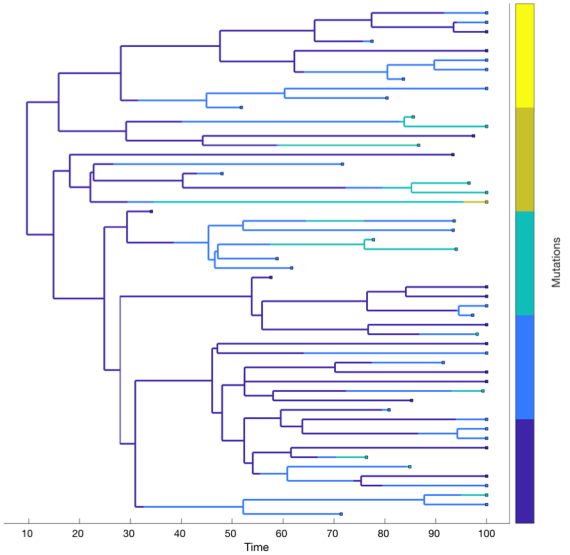
The marginal fitness birth-death model

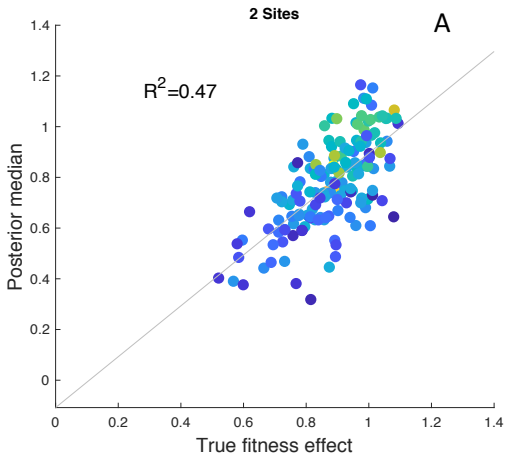
- We now have a new system of ODEs for tracking the probability that a lineage evolved exactly as observed at each site *k*:

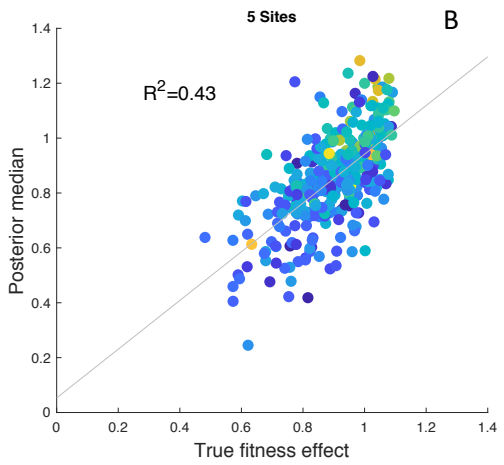
$$\begin{aligned}\frac{d}{dt}D_{n,k,i}(t) = &-\left(\hat{f}_{n,k,i}\lambda_0 + \sum_{j=1}^{M}\gamma_{i,j} + d\right)D_{n,k,i}(t) \\ &+ 2\hat{f}_{n,k,i}\lambda_0 E_u(t) D_{n,k,i}(t) \\ &+ \sum_{j=1}^{M}\gamma_{i,j}D_{n,k,j}(t).\end{aligned}$$

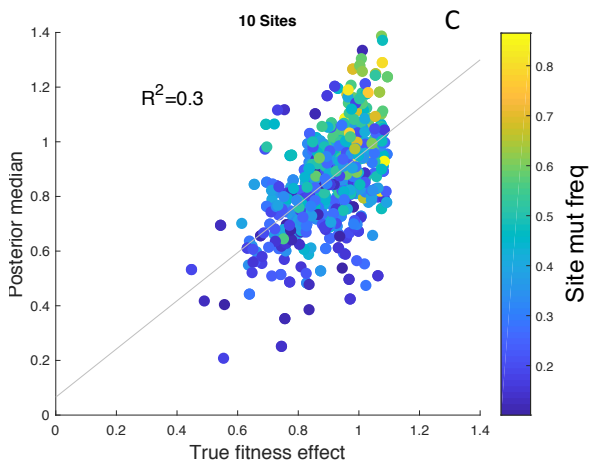
- The important part: This allows us to consider how selection shapes sequence evolution at multiple sites while considering how mutations act together to shape the fitness of a lineage.

Results: quantifying site-specific effects

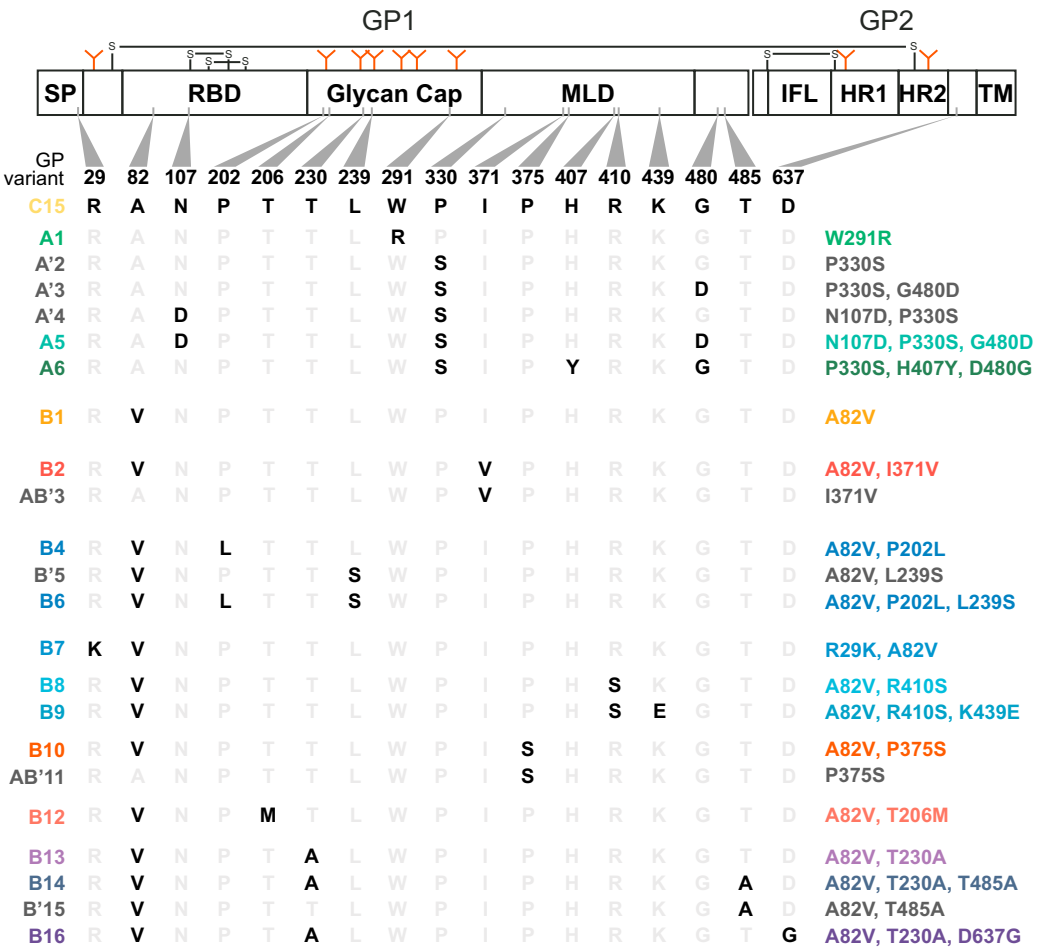


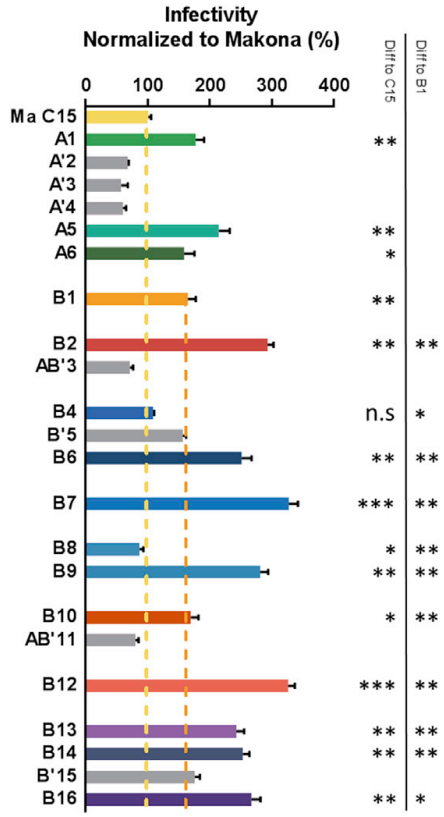






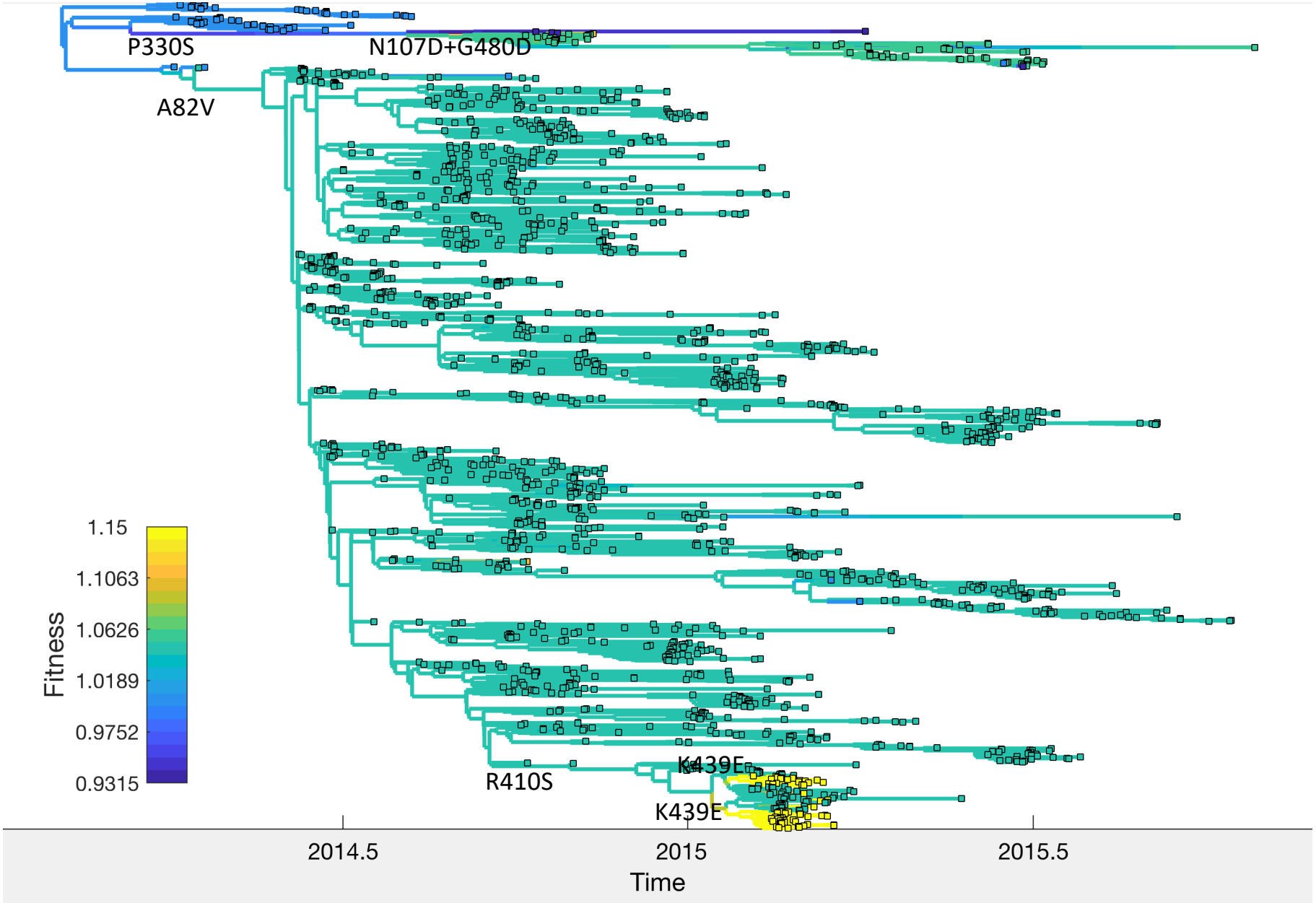
Adaptation of Ebola virus to humans



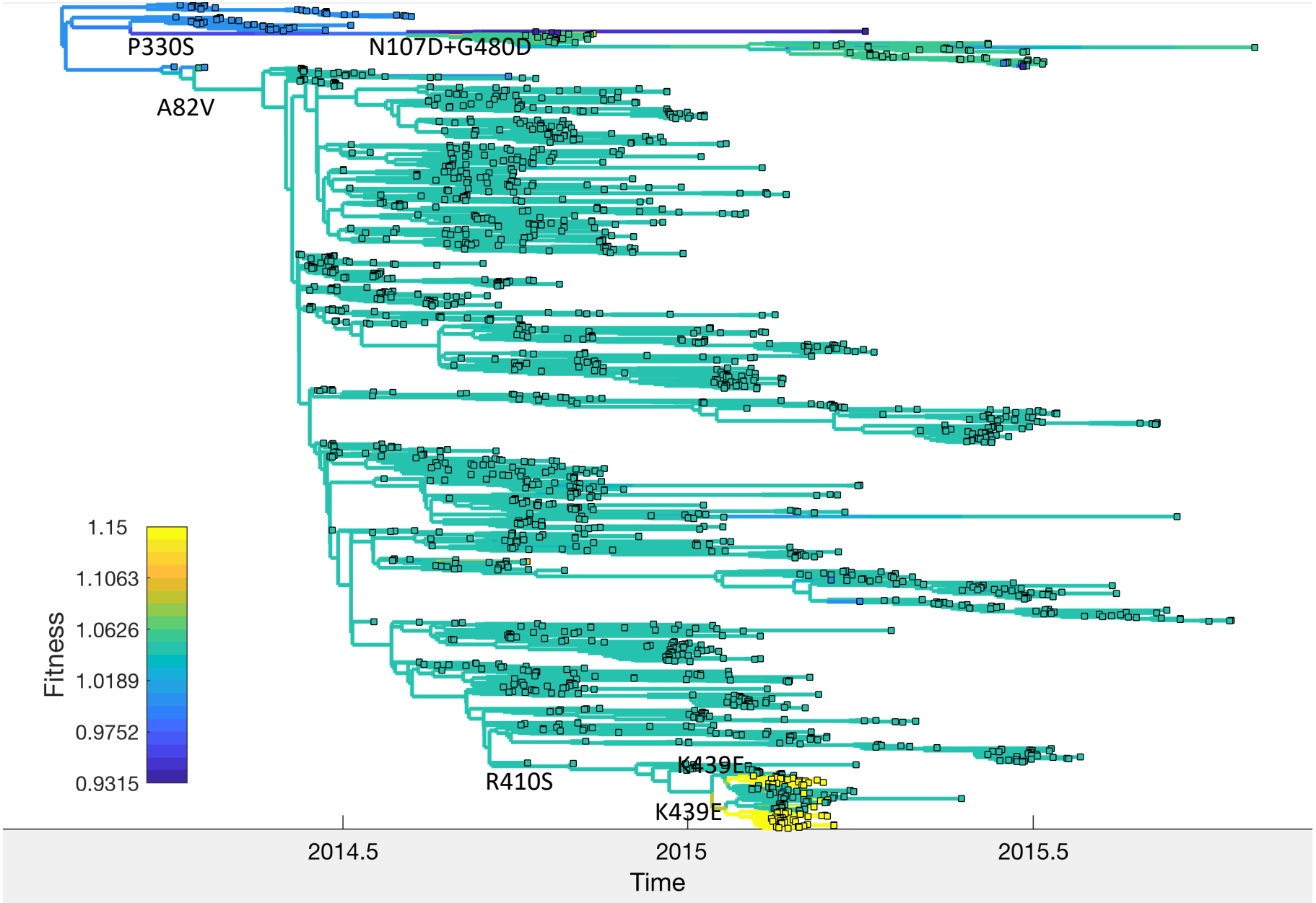


Urbanowicz *et al.* (Cell, 2016)

Adaptation of Ebola virus to humans



Adaptation of Ebola virus to humans



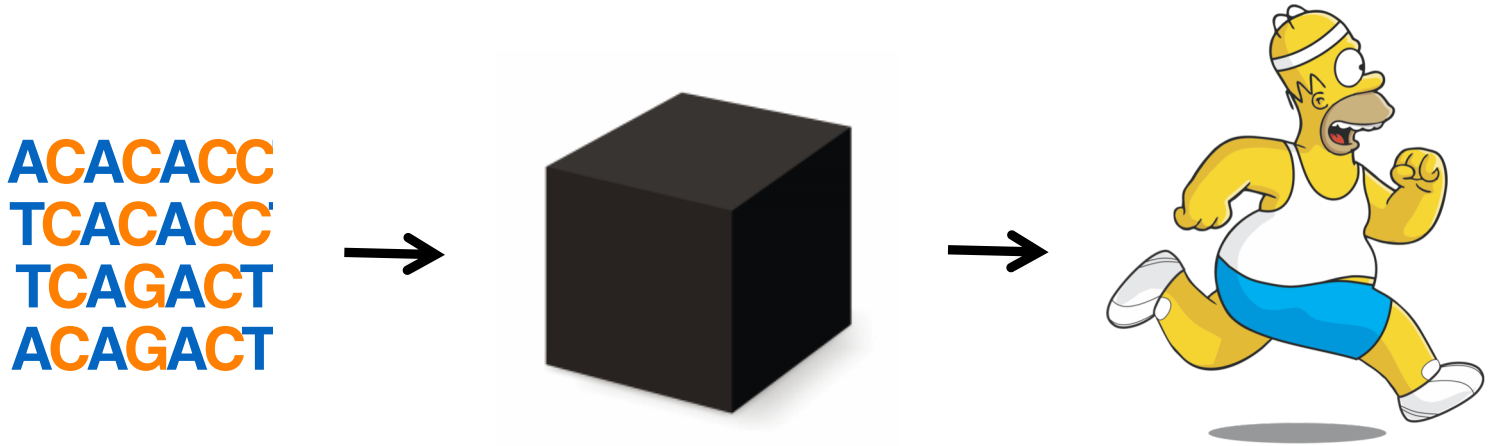
Current and ongoing work

- Marginal Fitness Birth Death model is implemented in *LUMIERE*, a package for BEAST2.

- High performance phylodynamic inference using Generalized Birth-Death Models

Generalized Birth-Death Models

- We would like to be able to learn how pathogen traits map to birth-death fitness parameters:



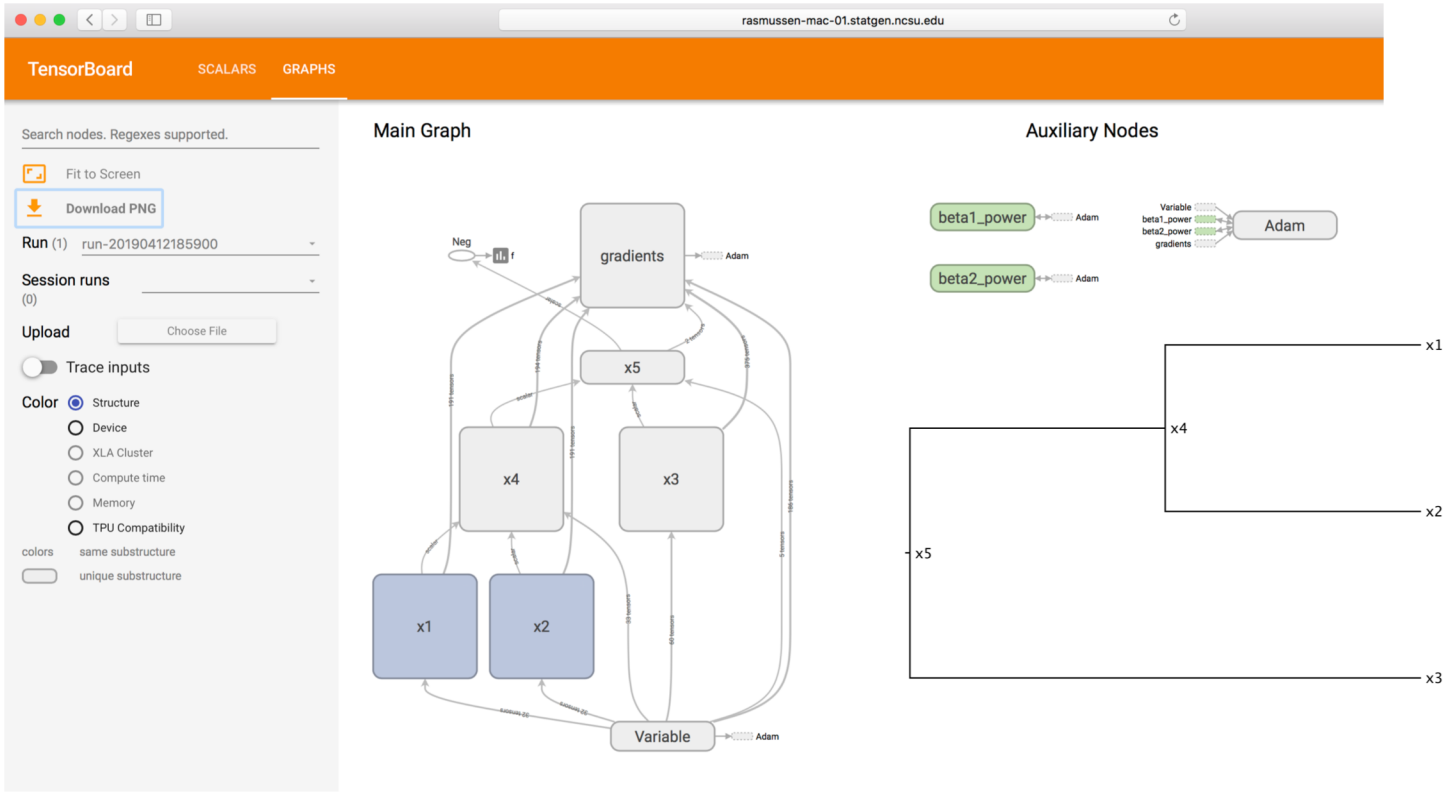
Can we have the best of both worlds?

- Can we perform likelihood-based phylodynamic inference under birth-death models

- While using the tools of machine learning to learn how pathogen features (e.g. genotypes) map to population-level parameters?

Phylodynamics in TensorFlow

- Trees can be represented as computational graphs in TensorFlow with data arrays (i.e. *tensors*) flowing between nodes



Phylodynamics in TensorFlow

```python
import tensorflow as tf
from FitModel import FitModel
from TimeVaryingBirthDeathModel import TimeVaryingBirthDeathModel

def train_model(tree, fasta_file, model, **params):

    "Convert tree to TensorTree object"
    tt = TensorTree(tree,seq_dic,**params)

    "Build birth-death model"
    bdm = TimeVaryingBirthDeathModel()
    bdm.build(tt,**params)

    "Build genotype to fitness model"
    fm = RandomEffectsFitModel()
    fm.build(tt,sites)
    line_fit_vals, birth_fit_vals = fm.update(tt)

    "Compute likelihood of tree under fitness model"
    tree_log_like = bdm.likelihood(line_fit_vals,birth_fit_vals,tt)
    penalty_log_like = fm.penalty(tt)
    tree_loss = -(tree_log_like + penalty_log_like)

    "Set up optimizer"
    learning_rate = 0.01;
    n_epochs = 1000
    optimizer = tf.train.AdamOptimizer(learning_rate = learning_rate)
    training_op = optimizer.minimize(tree_loss, var_list = fm.get_variables())

    with tf.Session() as sess:

        sess.run(tf.global_variables_initializer())

        for epoch in range(n_epochs):
            sess.run(training_op)

    return bdm, fm
```

Convert tree to tensor-like object

Specify birth-death model

Specify fitness model

Specify loss function

Set up optimizer

Run!!!

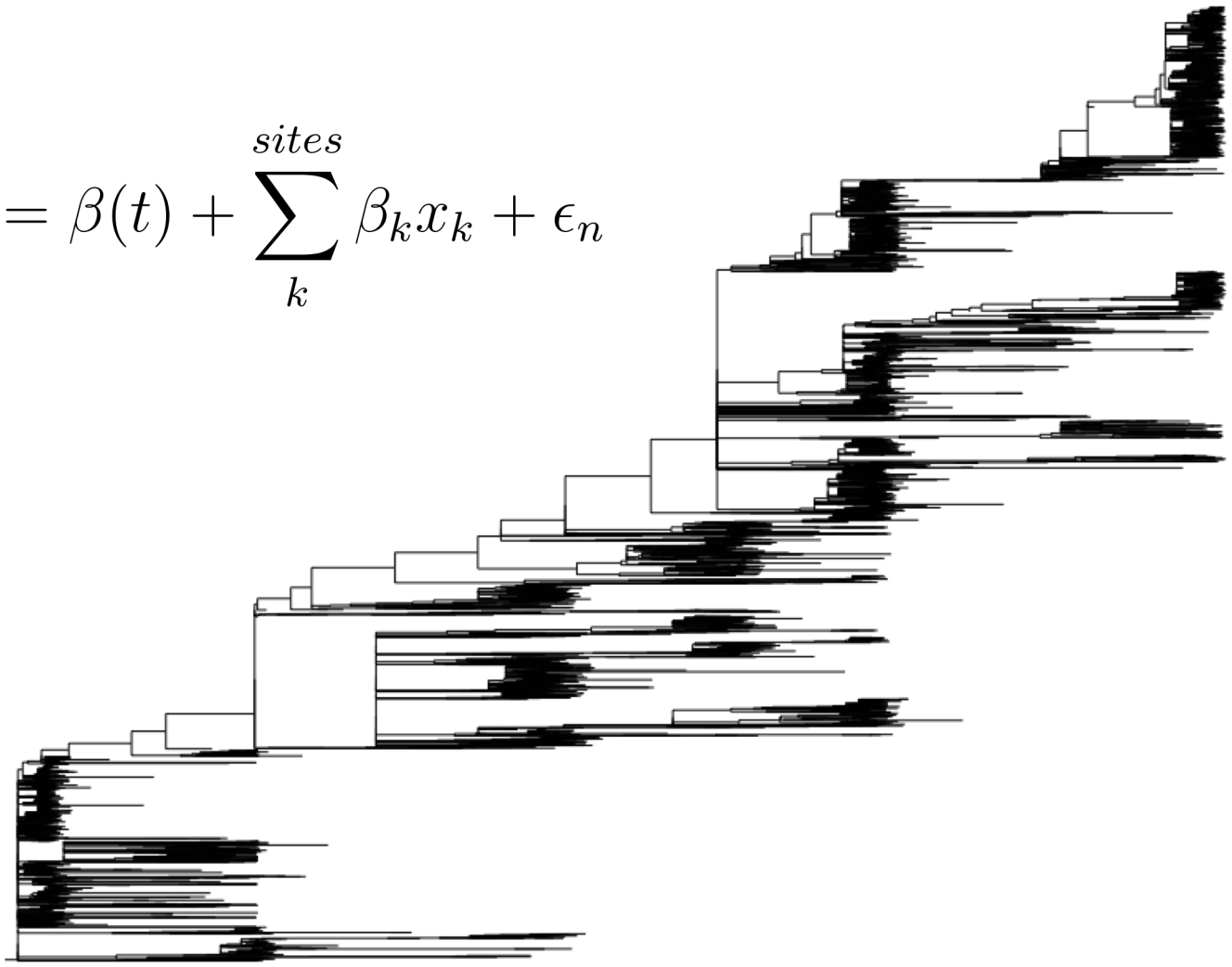
Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants

Juhye M. Lee[a,d,e,1], **John Huddleston**[b,f,1], **Michael B. Doud**[a,d,e], **Kathryn A. Hooper**[a,f], **Nicholas C. Wu**[g], **Trevor Bedford**[b,c,1], **and Jesse D. Bloom**[a,c,d,1]

[a]Basic Sciences Division; [b]Vaccine and Infectious Disease Division; [c]and Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [d]Department of Genome Sciences; [e]Medical Scientist Training Program; [f]and Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195, USA; [g]Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037

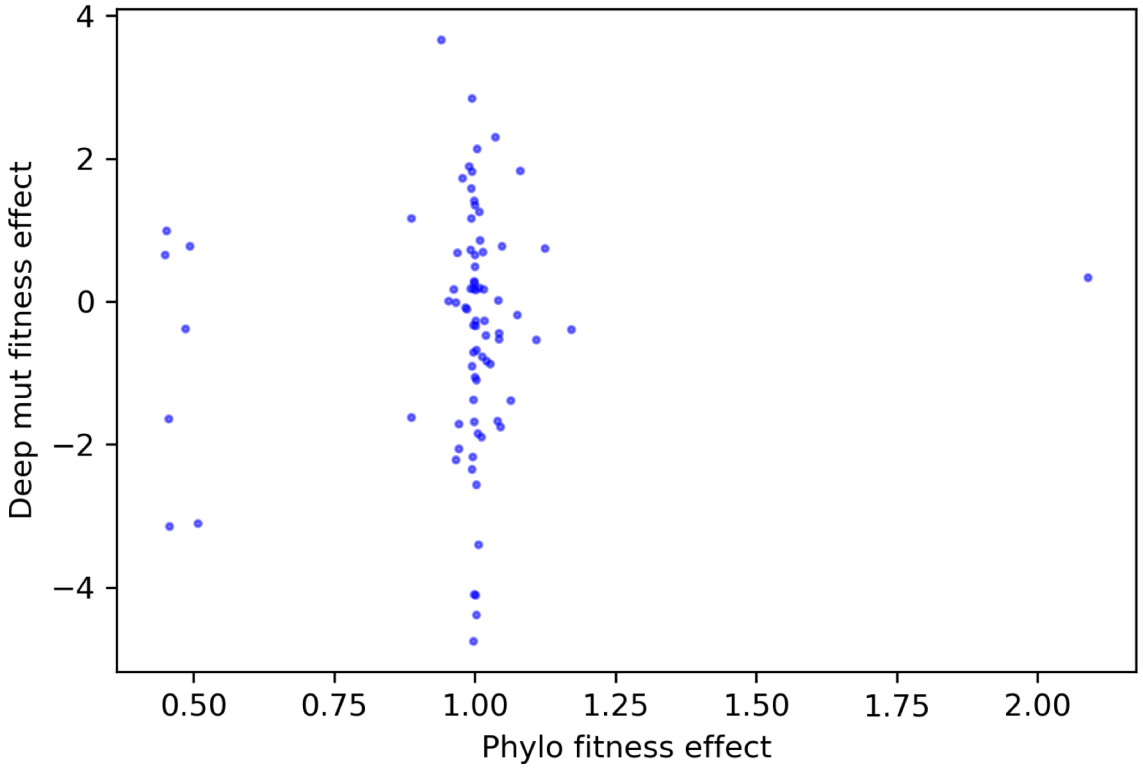
Fitness model for H3N2 influenza

$$f_n = \beta(t) + \sum_{k}^{sites} \beta_k x_k + \epsilon_n$$



Phylodynamics vs. Deep Mutational Scanning

- Very small fitness effects once we account for seasonal fluctuations in flu transmission



A big thank you to:

Prof. Tanja Stadler

The Phylodynamics Research Group at NC State

phylodynamics.wordpress.ncsu.edu

For the South African HIV project:

Tulio de Oliviera
Eduan Wilkinson
Africa Health Research Institute

AHRI AFRICA HEALTH RESEARCH INSTITUTE

For funding:

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

European Commission

erc European Research Council

The marginal fitness birth-death model

- We now have a new system of ODEs for tracking the probability that a lineage evolved exactly as observed at each site *k*:

$$\begin{aligned}\frac{d}{dt}D_{n,k,i}(t) = &-\left(\hat{f}_{n,k,i}\lambda_0 + \sum_{j=1}^{M}\gamma_{i,j} + d\right)D_{n,k,i}(t) \\ &+ 2\hat{f}_{n,k,i}\lambda_0 E_u(t) D_{n,k,i}(t) \\ &+ \sum_{j=1}^{M}\gamma_{i,j}D_{n,k,j}(t).\end{aligned}$$

- Note 1: We can track evolution at each site individually without tracking all genotypes.

The marginal fitness birth-death model

- We now have a new system of ODEs for tracking the probability that a lineage evolved exactly as observed at each site *k*:

$$\frac{d}{dt}D_{n,k,i}(t) = -\left(\hat{f}_{n,k,i}\lambda_0 + \sum_{j=1}^{M}\gamma_{i,j} + d\right)D_{n,k,i}(t) + 2\hat{f}_{n,k,i}\lambda_0 E_u(t) D_{n,k,i}(t) + \sum_{j=1}^{M}\gamma_{i,j}D_{n,k,j}(t).$$

- Note 2: We can simultaneously take into account the coupled fitness effects of mutations at all other sites.

The marginal fitness birth-death model

- We now have a new system of ODEs for tracking the probability that a lineage evolved exactly as observed at each site *k*:

$$\frac{d}{dt}D_{n,k,i}(t) = -\left(\hat{f}_{n,k,i}\lambda_0 + \sum_{j=1}^{M}\gamma_{i,j} + d\right)D_{n,k,i}(t) + 2\hat{f}_{n,k,i}\lambda_0 E_u(t) D_{n,k,i}(t) + \sum_{j=1}^{M}\gamma_{i,j}D_{n,k,j}(t).$$

- Note 3: Tracking $D_{n,k,i}$ all the way back to the root allows us to compute the joint likelihood of the tree and the sequence data at site *k*.

$$L(\mathcal{S}_k, \mathcal{T}|\theta) = \sum_{i=1}^{M} D_{n,k,i}(t_{root})$$

Adaptation of Ebola virus to humans