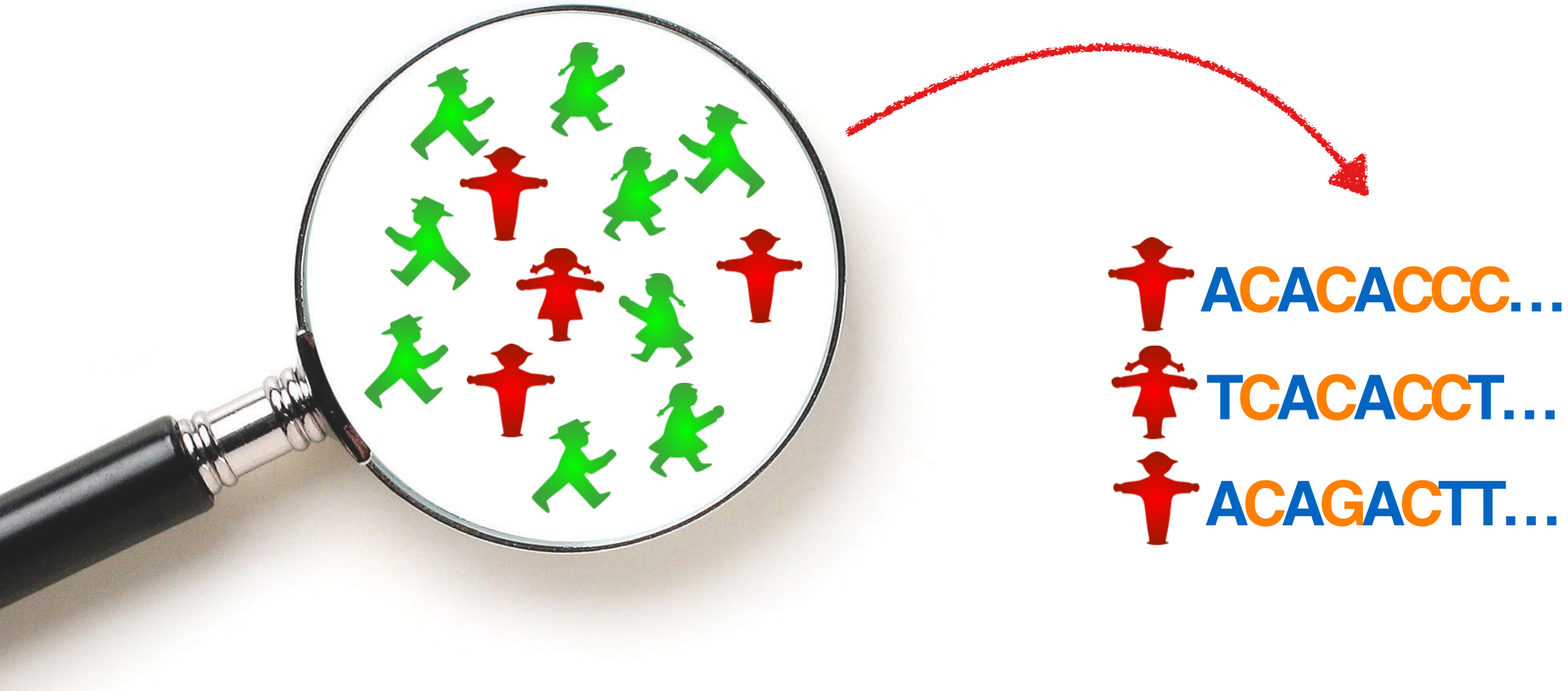
Quantifying the **contribution** of **external covariates** to **pathogen population dynamics** in a birth-death framework

---

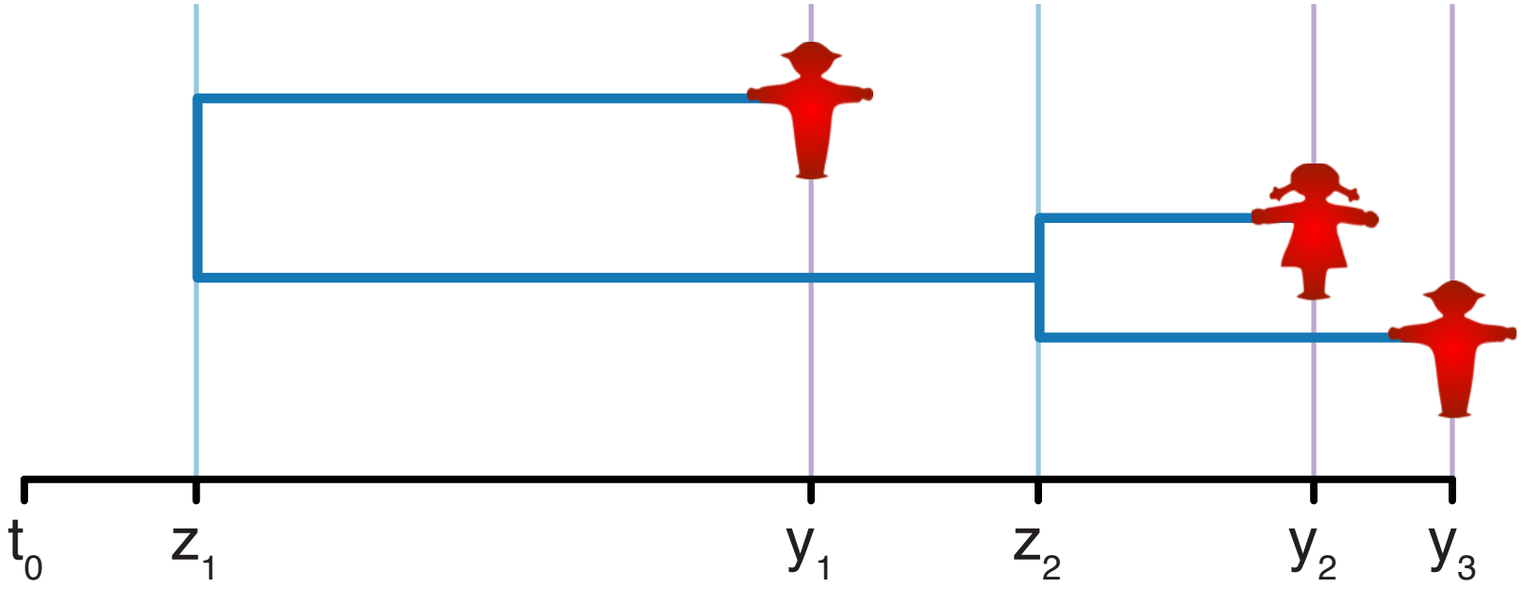
Louis du Plessis

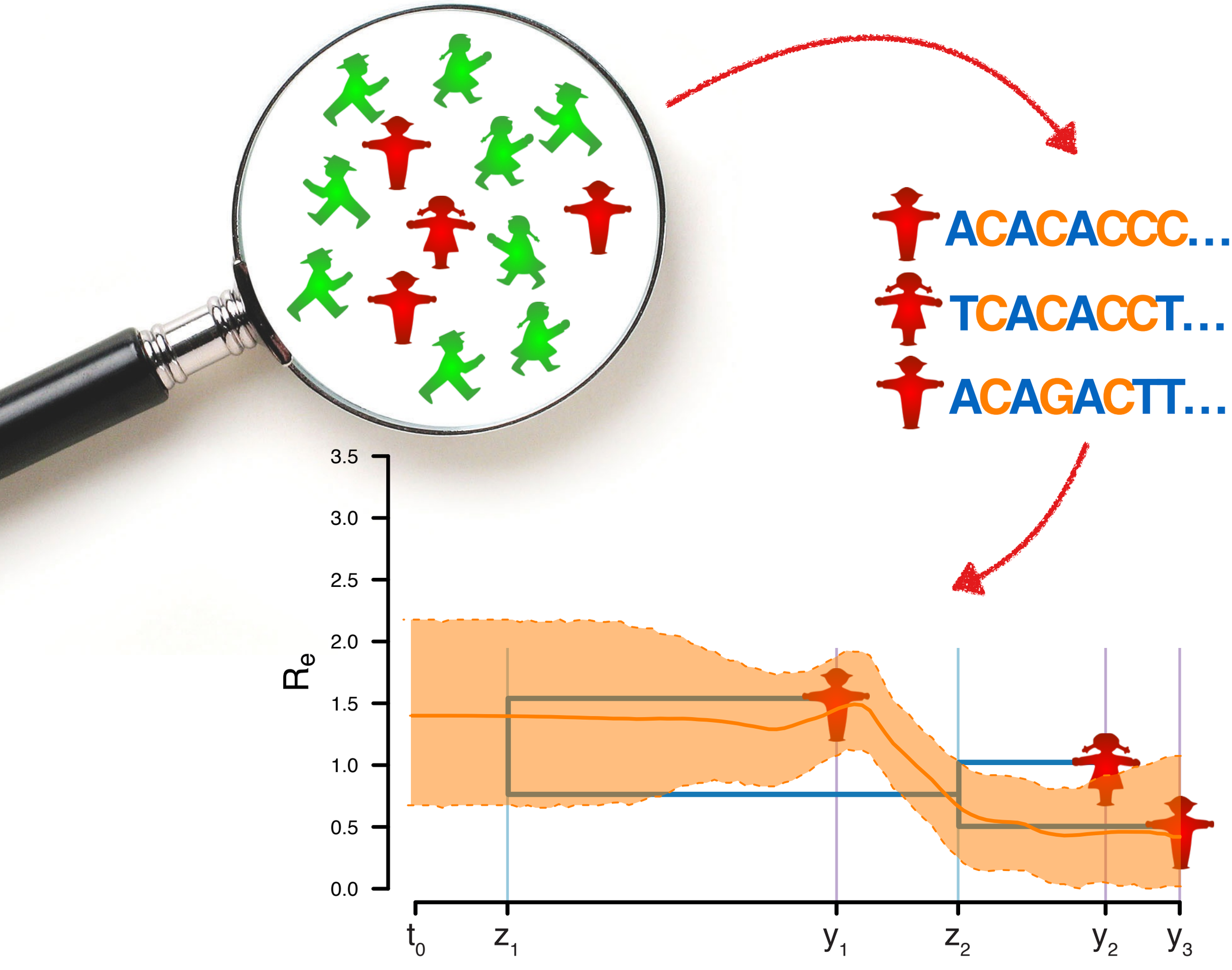

ACACACCC…
TCACACCT…
ACAGACTT…

ACACACCC…

TCACACCT…

ACAGACTT…



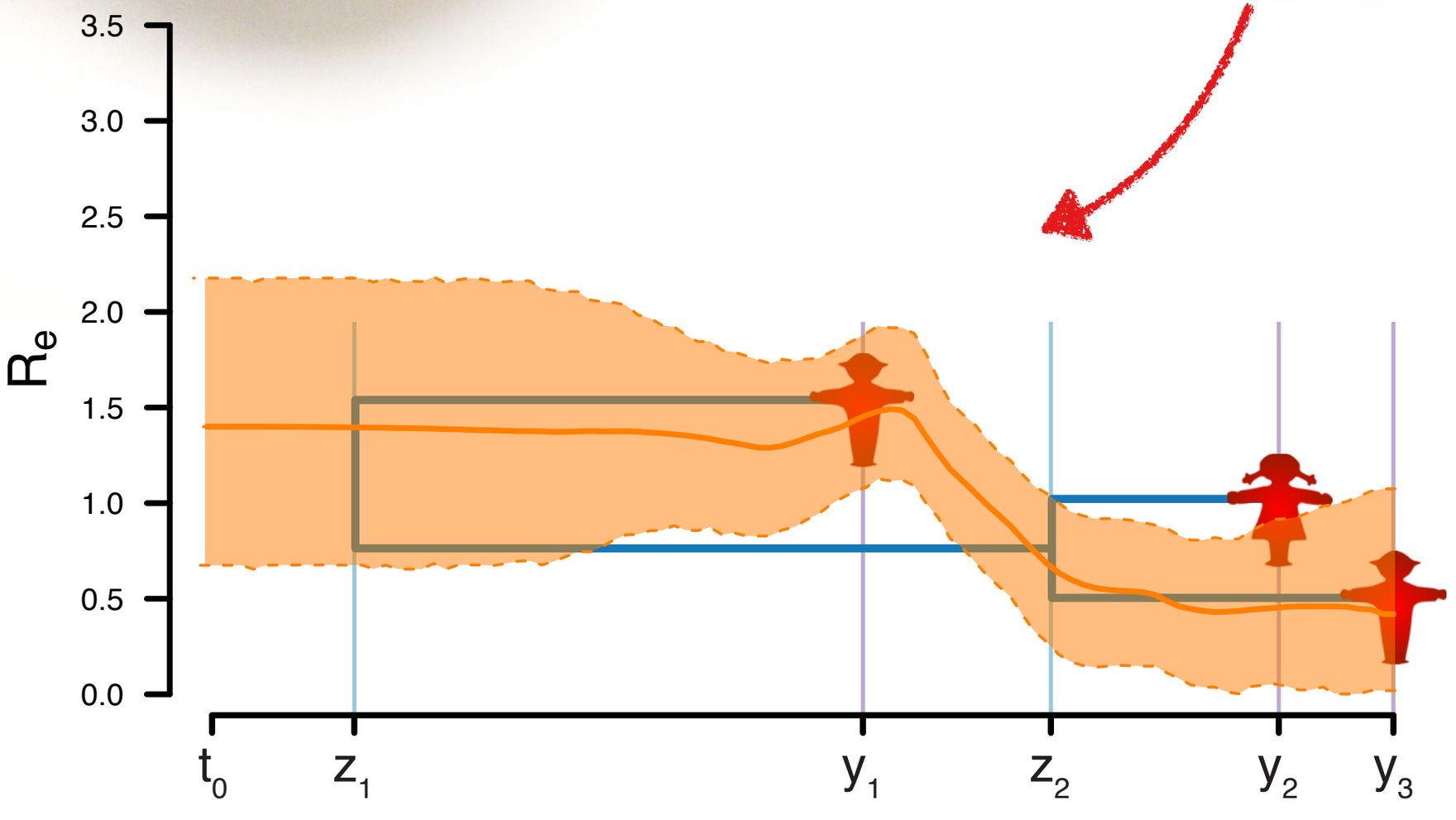

ACACACCC…
TCACACCT…
ACAGACTT…
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
R_e
t_0
z_1
y_1
z_2
y_2
y_3

Tree
Realisation of a stochastic process

$P(\mathcal{E} \mid \theta)$

Demographic model
Describes the growth of the tree (spread of epidemic)

ACAGACTT...



Tree
Realisation of a stochastic process

Demographic model
Describes the growth of the tree (spread of epidemic)

Epidemiological and **evolutionary** dynamics should occur on the same **timescale!**

Measurably evolving populations:

- Large population size
- High mutation rate
- Short generation times

This is the case for **many** (if not most) viral epidemics!

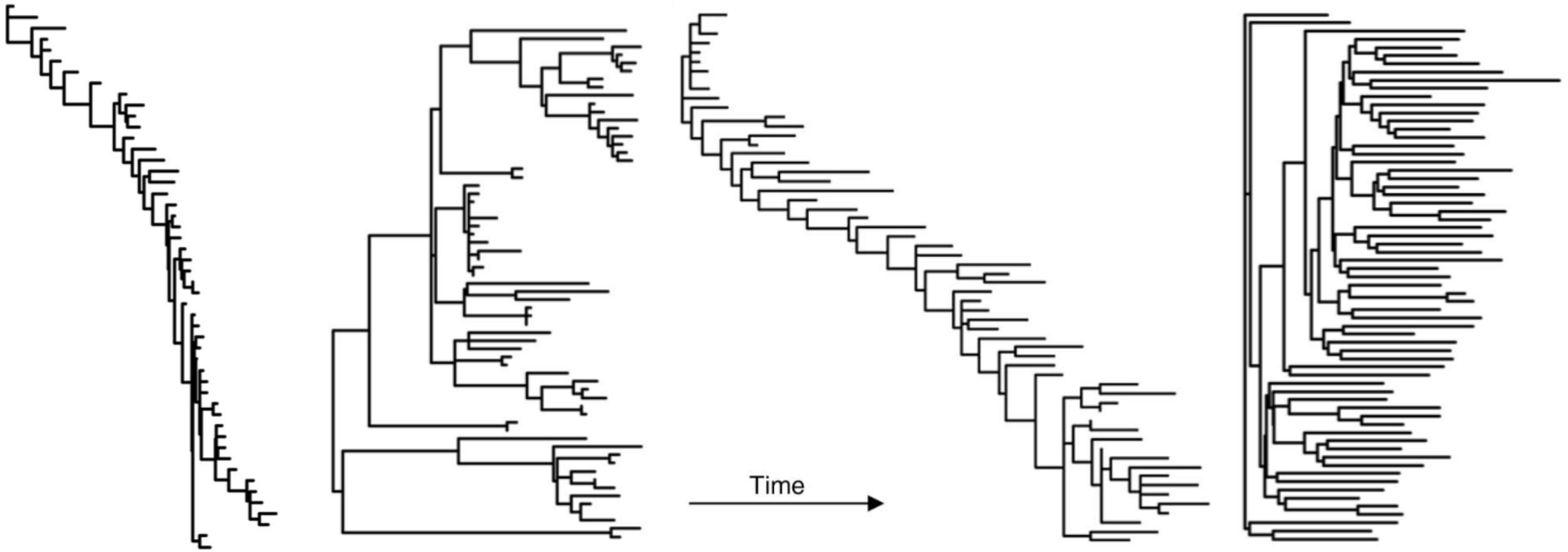
Genomes contain a **signature** of the **epidemiological** dynamics

Influenza A

Measles

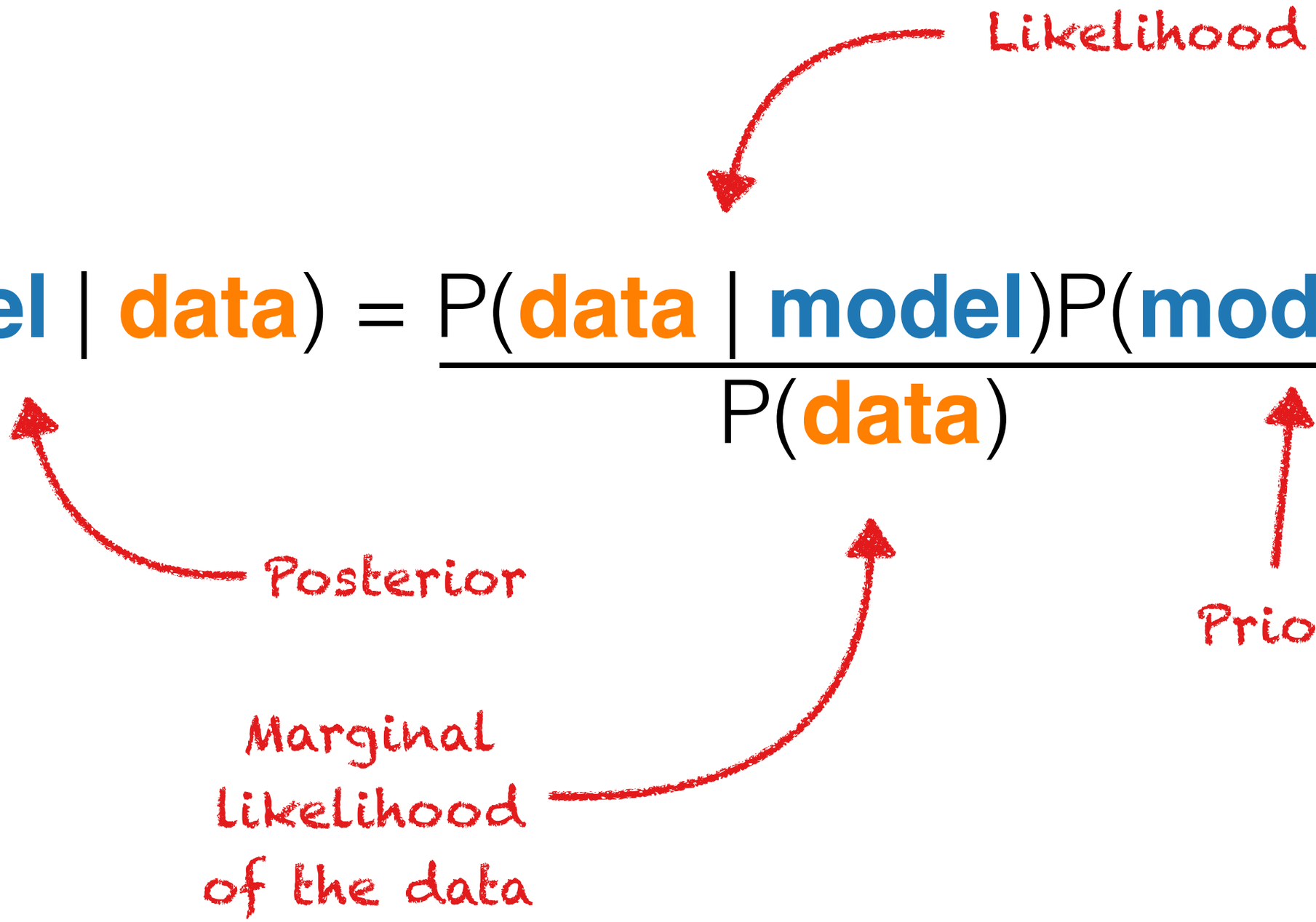
HIV within-host

HCV

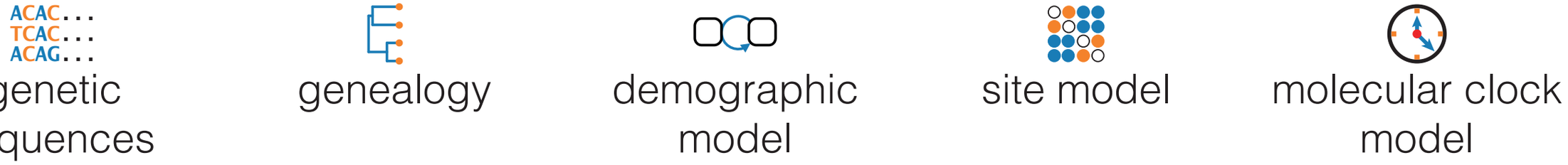


Grenfell *et al.* **Science 2004**

Bayesian **phylogenetic** and **phylodynamic** inference



Hierarchical **Bayesian** model in **Beast2**



Hierarchical **Bayesian** model in **Beast2**

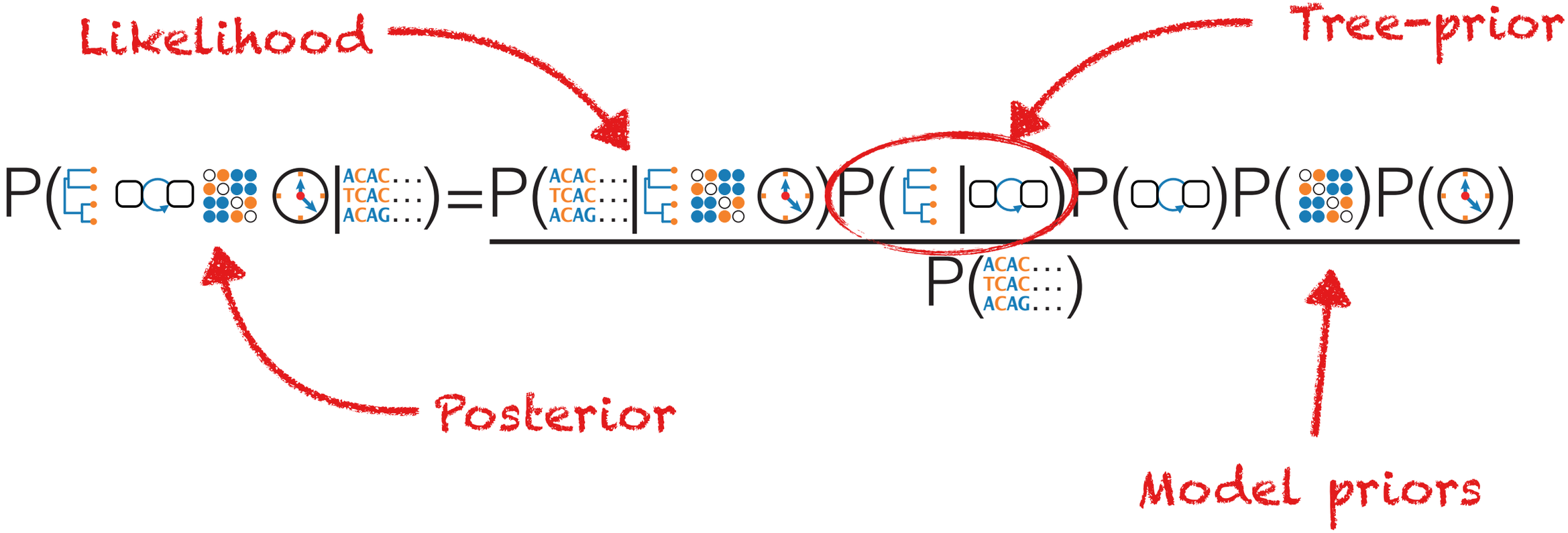
genetic sequences

genealogy

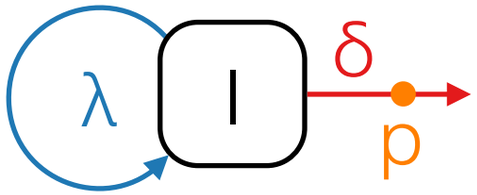
demographic model

site model

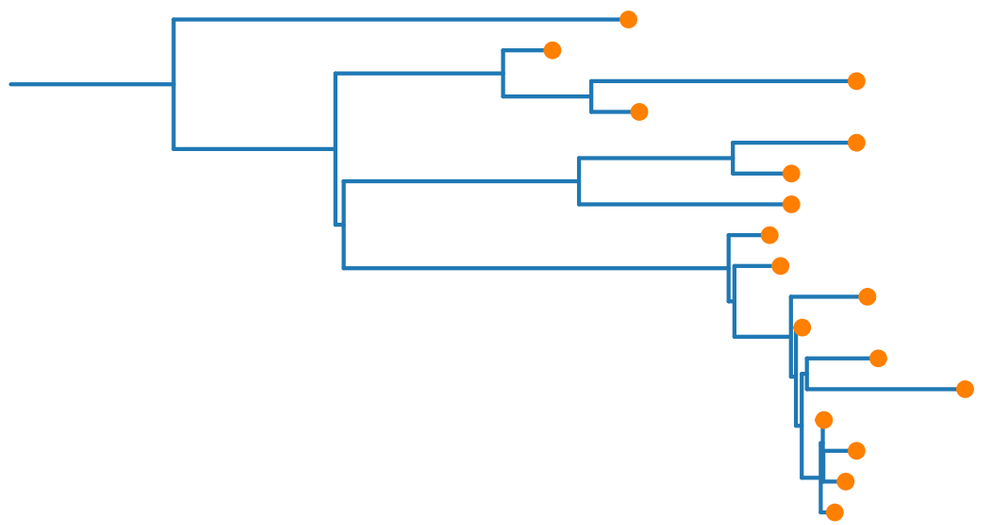
molecular clock model



Birth-death skyline

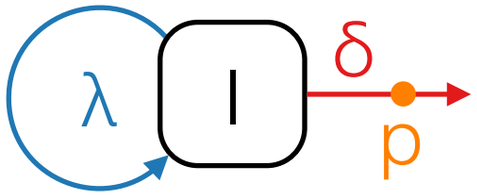


- λ — infection rate
- δ — becoming-noninfectious rate
- p — sampling proportion
- t_0 — Origin of the process

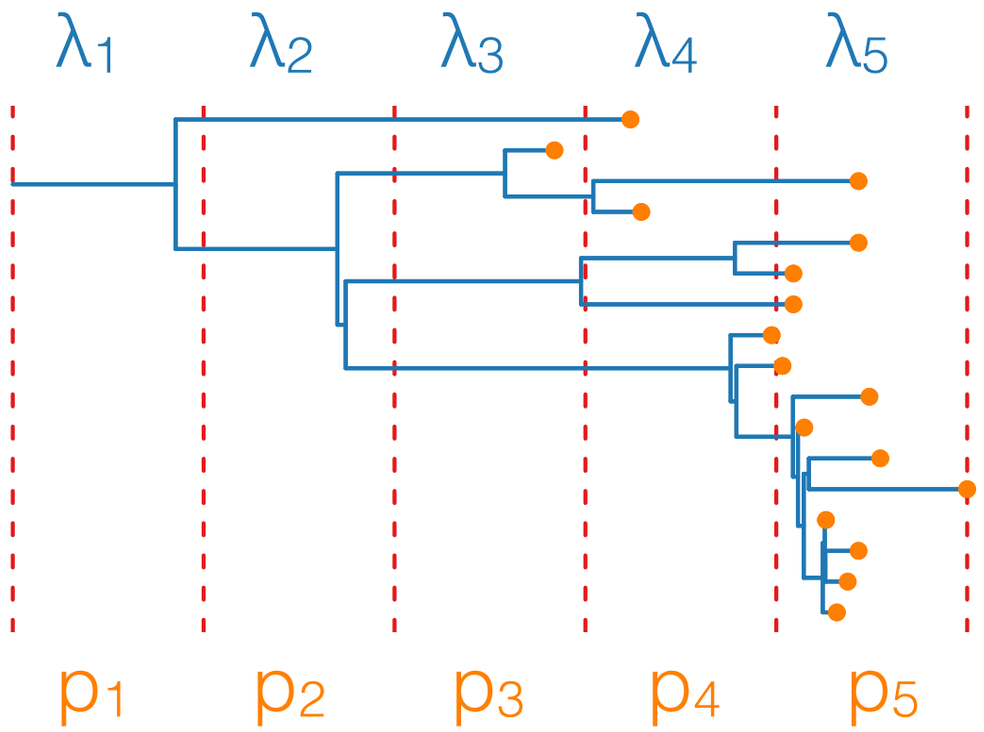


- Model parameters are rates that describe the growth of the tree
- Effective reproduction number = λ/δ
  (is it spreading or not?)

Birth-death skyline

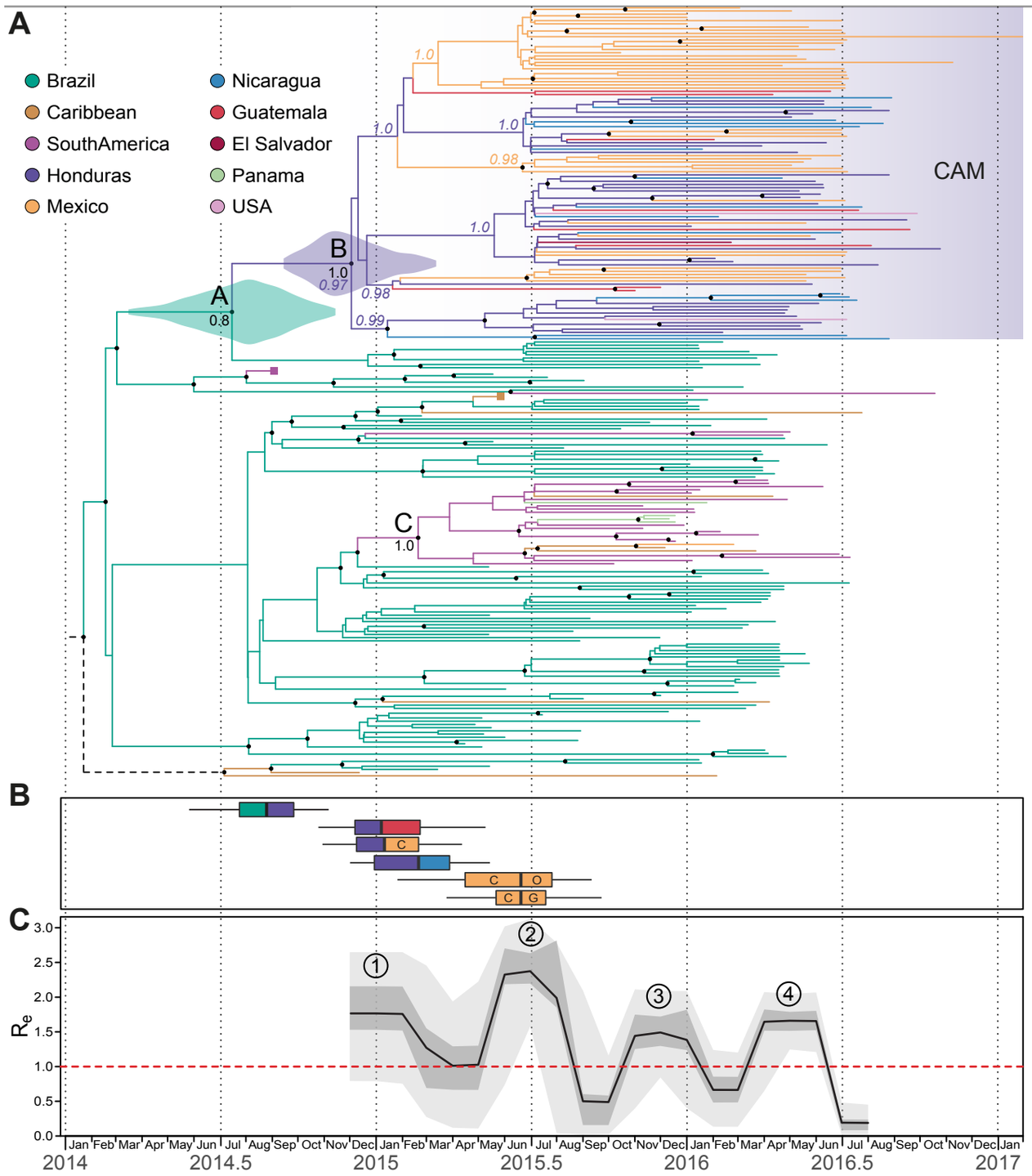


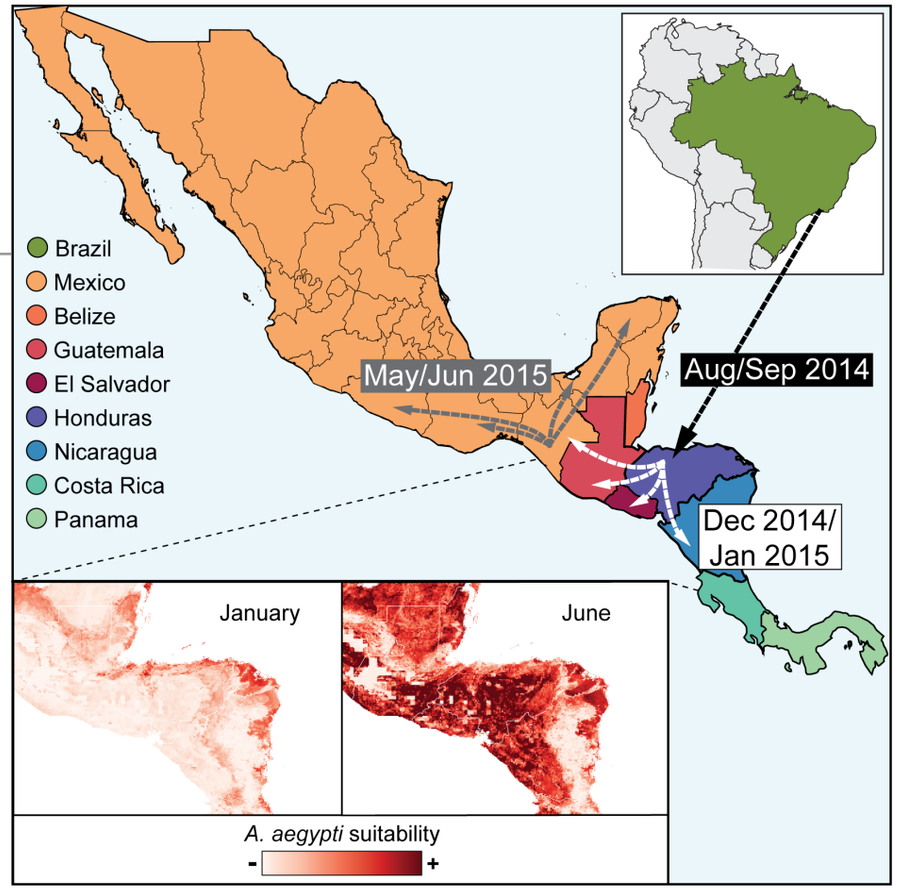
- λ — infection rate
- δ — becoming-noninfectious rate
- p — sampling proportion
- t_0 — Origin of the process



- Model parameters are rates that describe the growth of the tree
- Effective reproduction number = λ/δ
  (is it spreading or not?)
- Allow parameters to change through time
- Shifts in rates can be anywhere
  (usually evenly distributed)

Zika virus in Central America

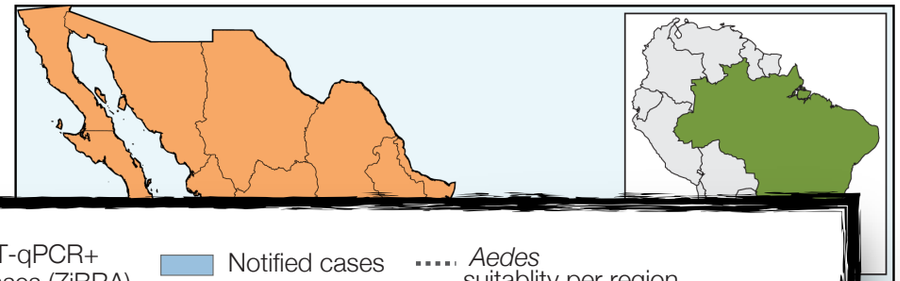


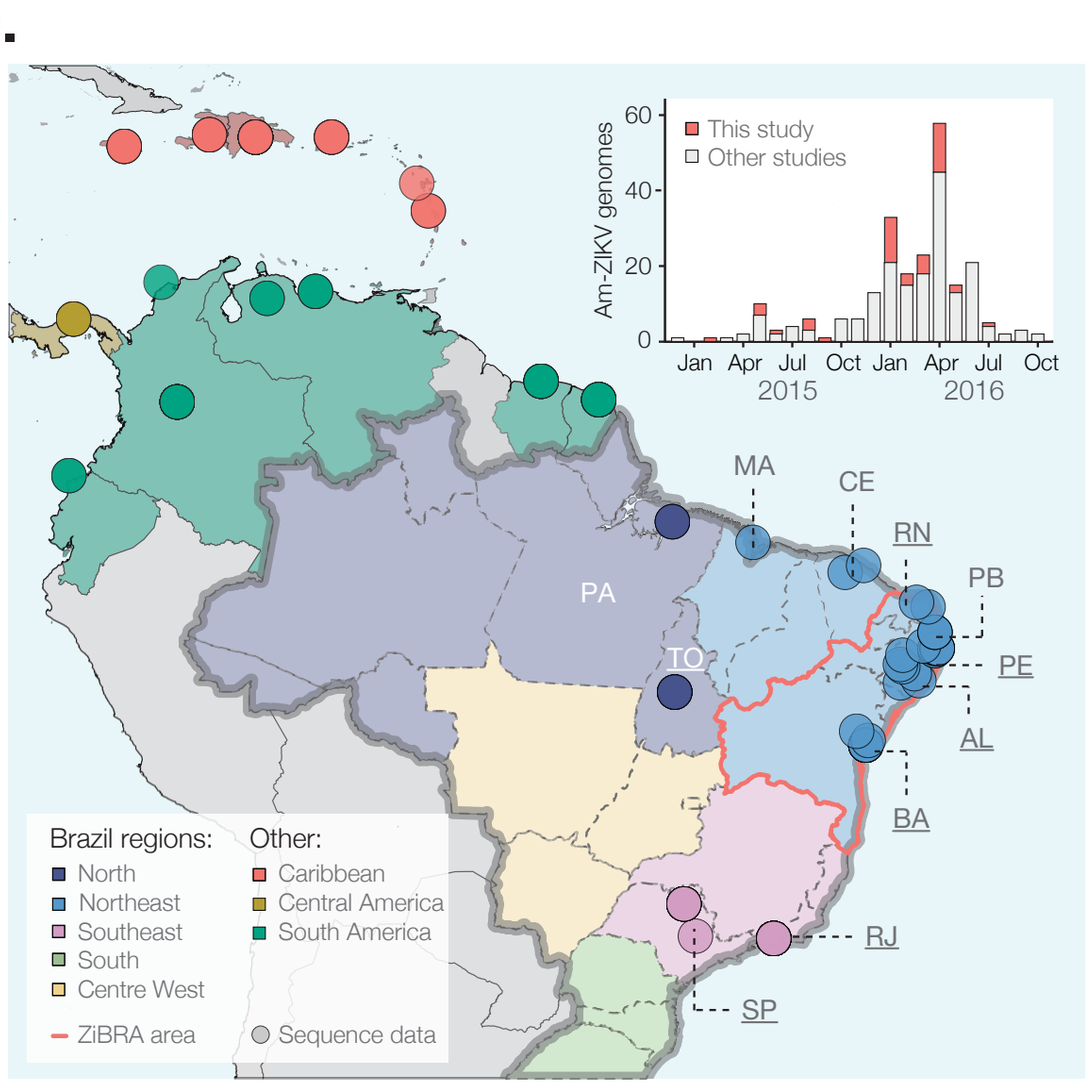


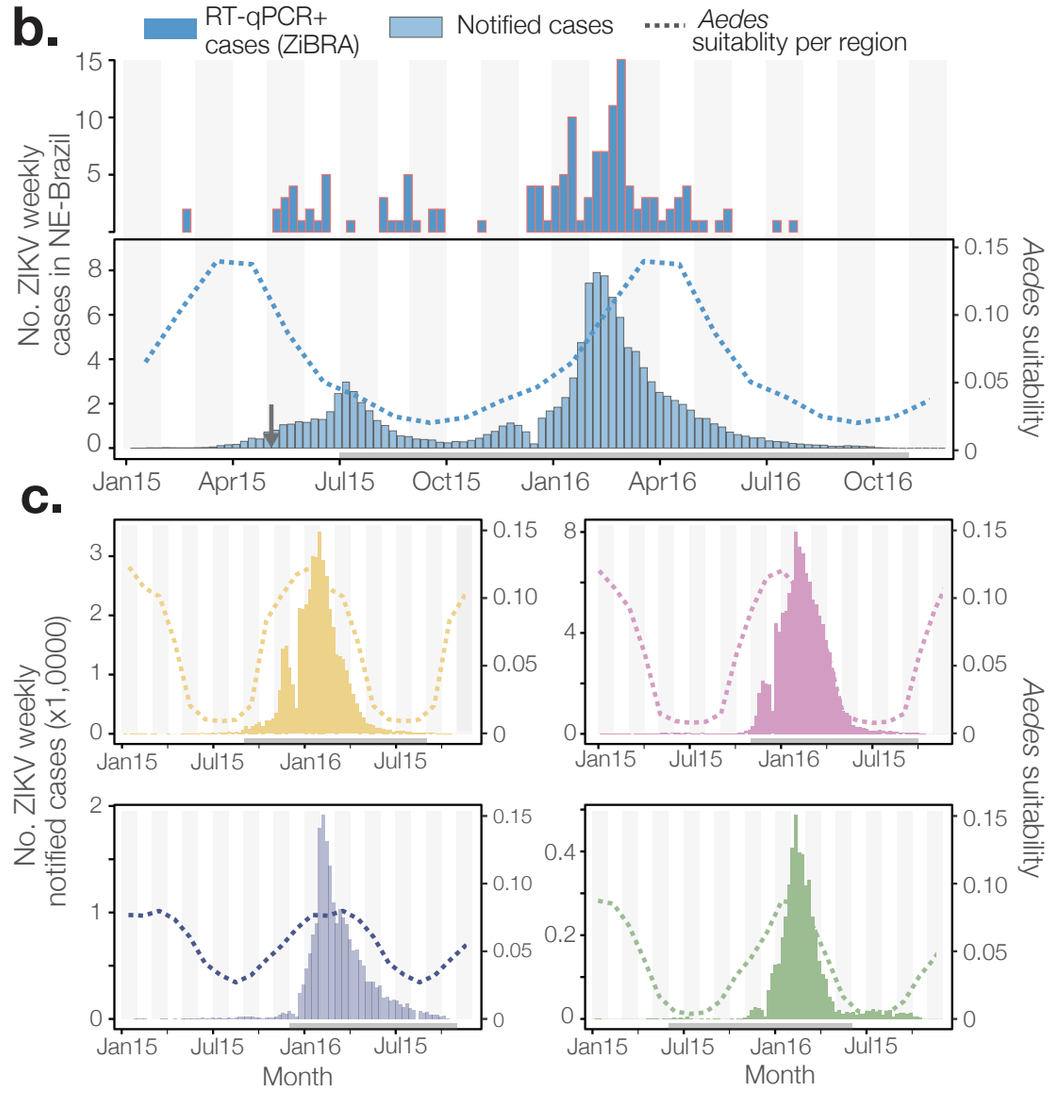
- 104 genomes
- Single introduction in Honduras in late 2014 / early 2015
- Birth-death skyline estimates of R_e show some evidence for seasonality

Thézé, Li, du Plessis *et al.* **Cell H&M 2018**

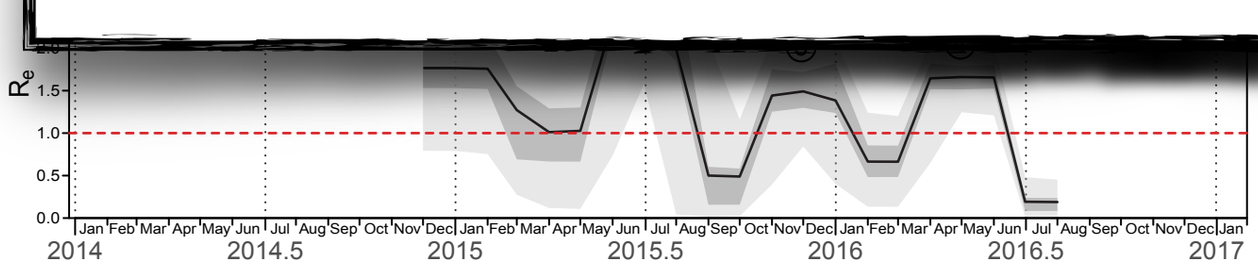
Zika virus in Central America





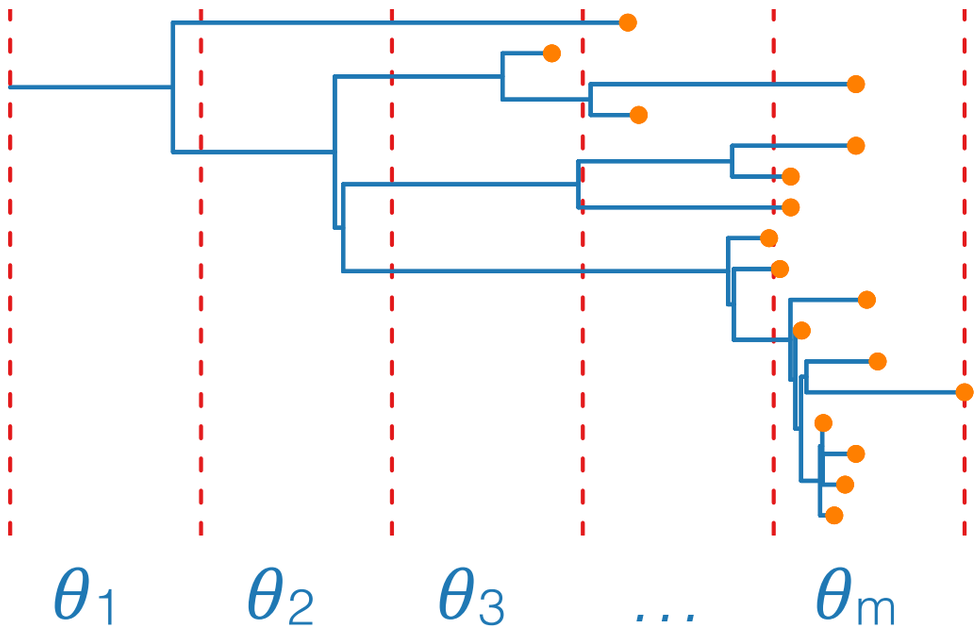


Faria *et al.* **Nature 2017**



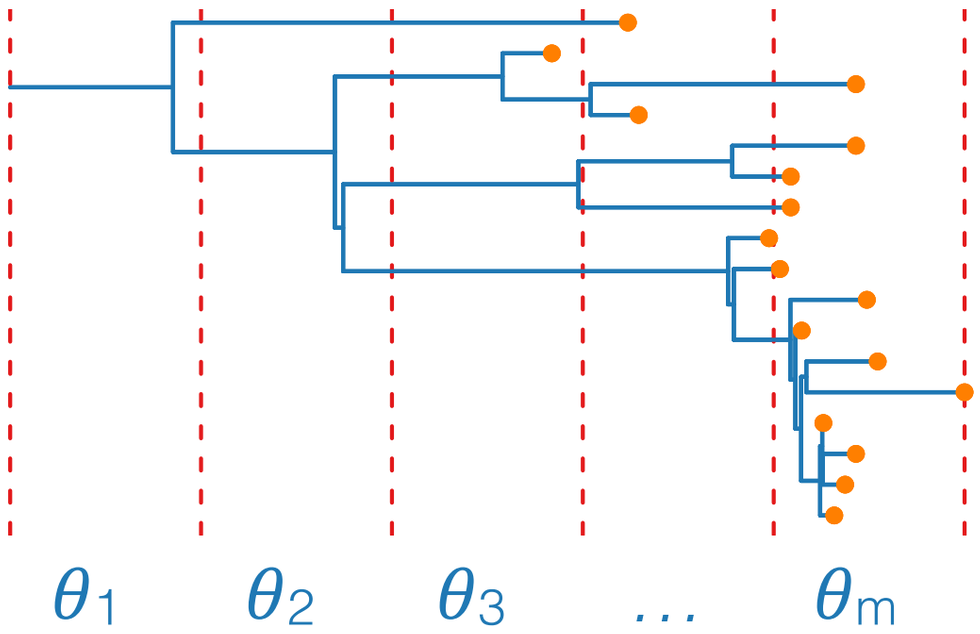
Thézé, Li, du Plessis *et al.* **Cell H&M 2018**

Phylodynamic linear model formulation



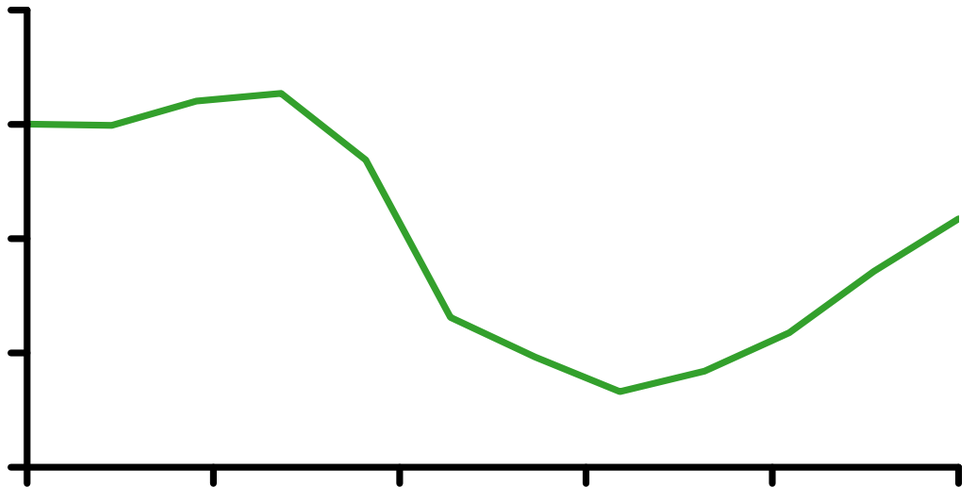
$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Phylodynamic linear model formulation

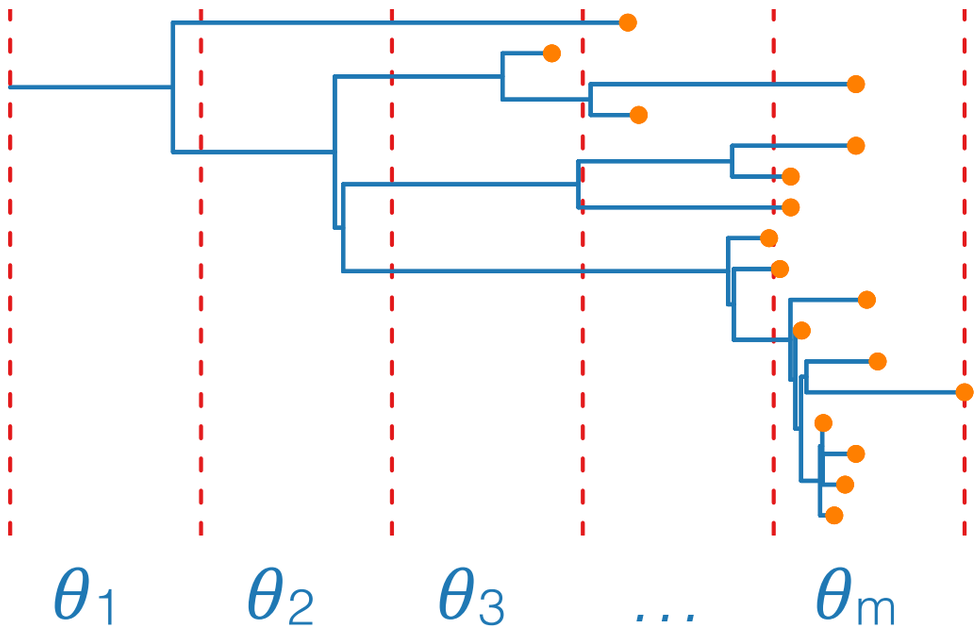


$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Covariates

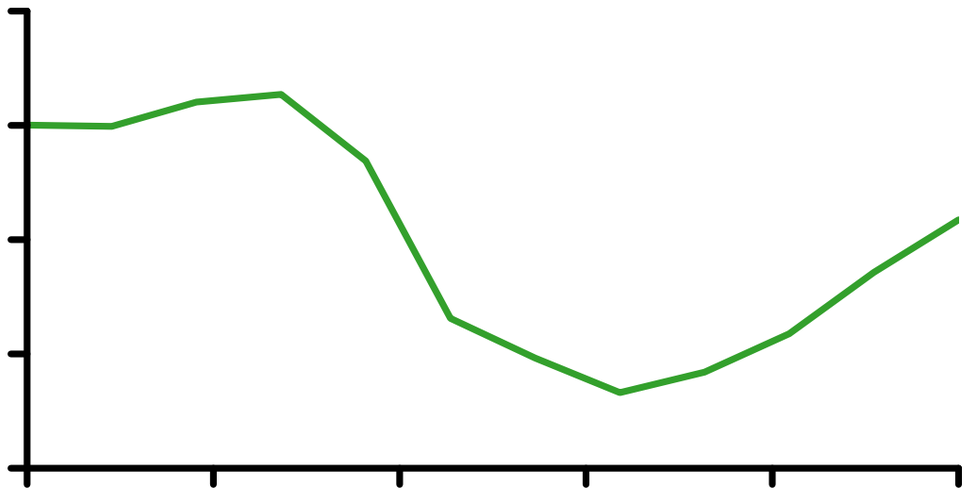


Phylodynamic linear model formulation



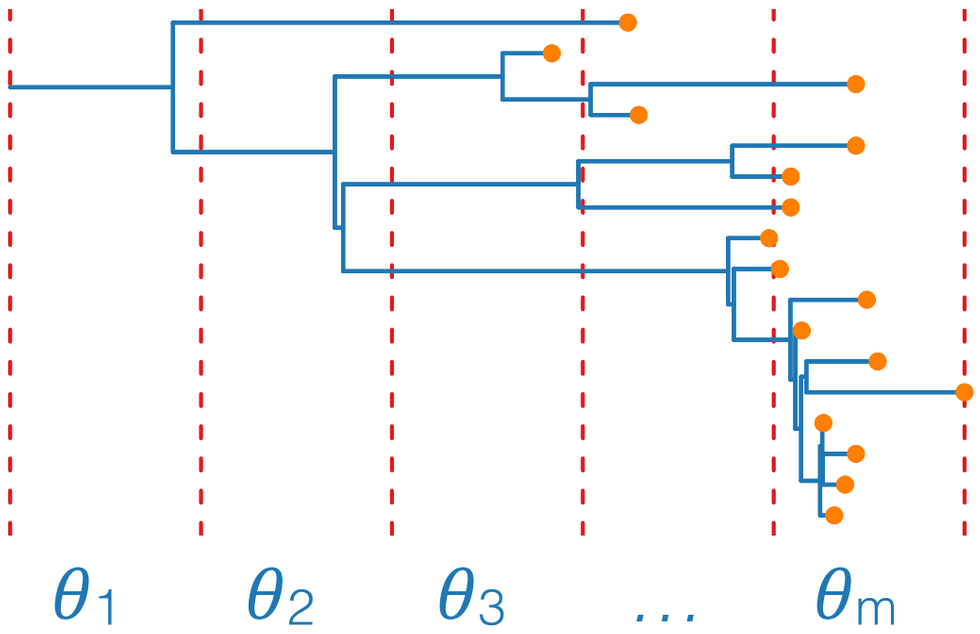
$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Covariates



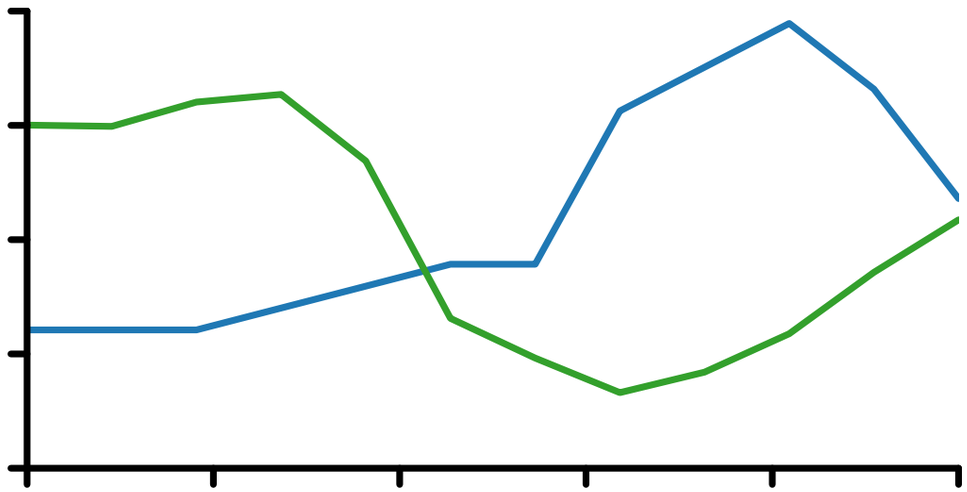
$$g(\theta_i) = x_{i,1}\beta_1$$

Phylodynamic linear model formulation



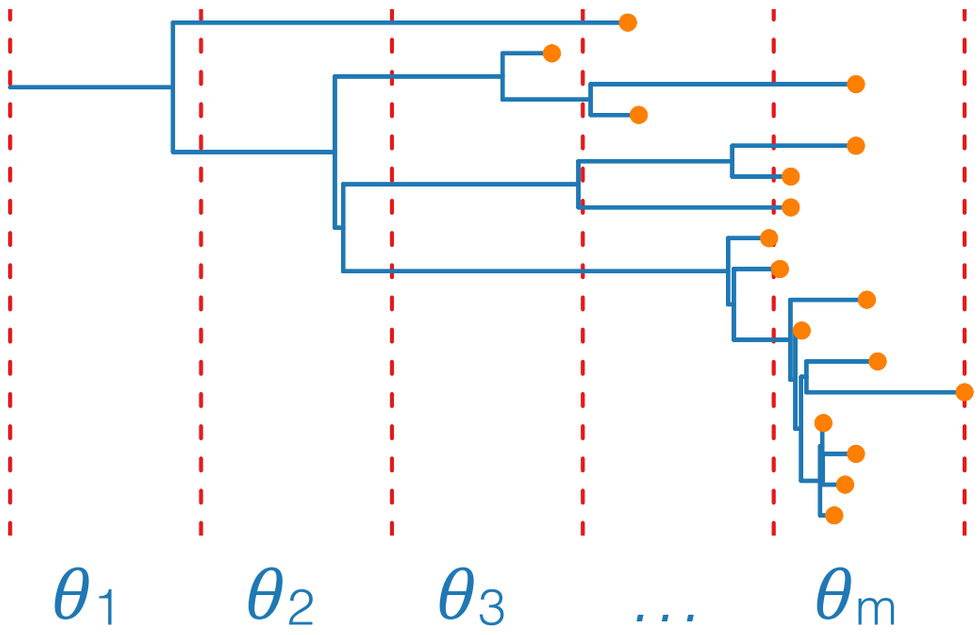
$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Covariates



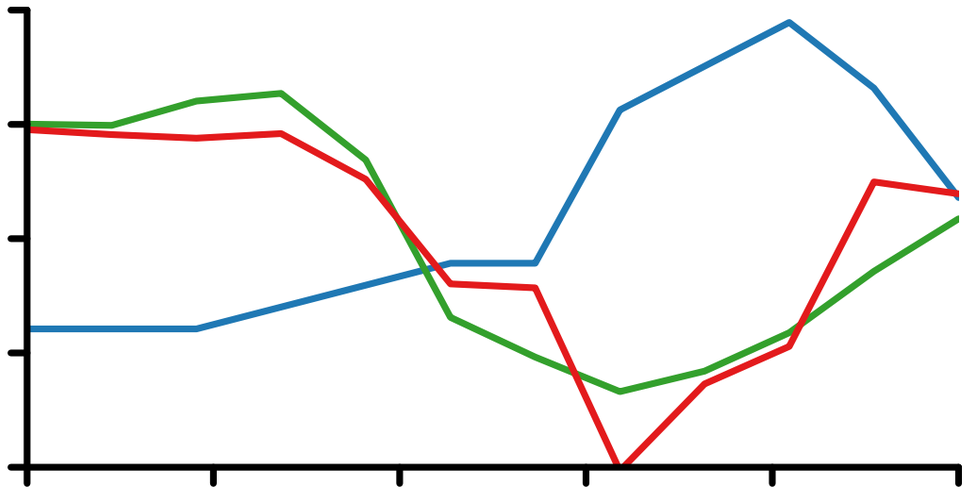
$$g(\theta_i) = x_{i,1}\beta_1 + x_{i,2}\beta_2$$

Phylodynamic linear model formulation



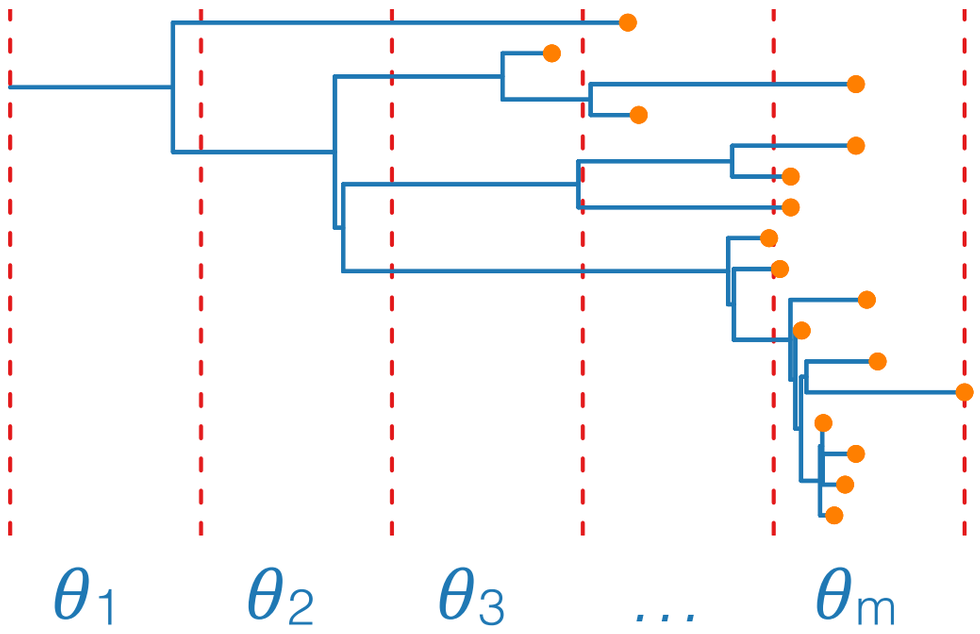
$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Covariates



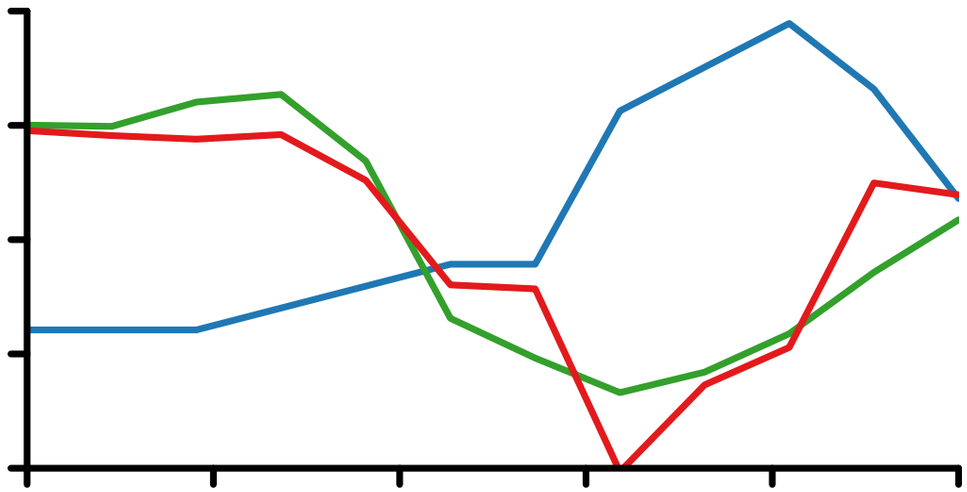
$$g(\theta_i) = x_{i,1}\beta_1 + x_{i,2}\beta_2 + x_{i,3}\beta_3$$

Phylodynamic linear model formulation



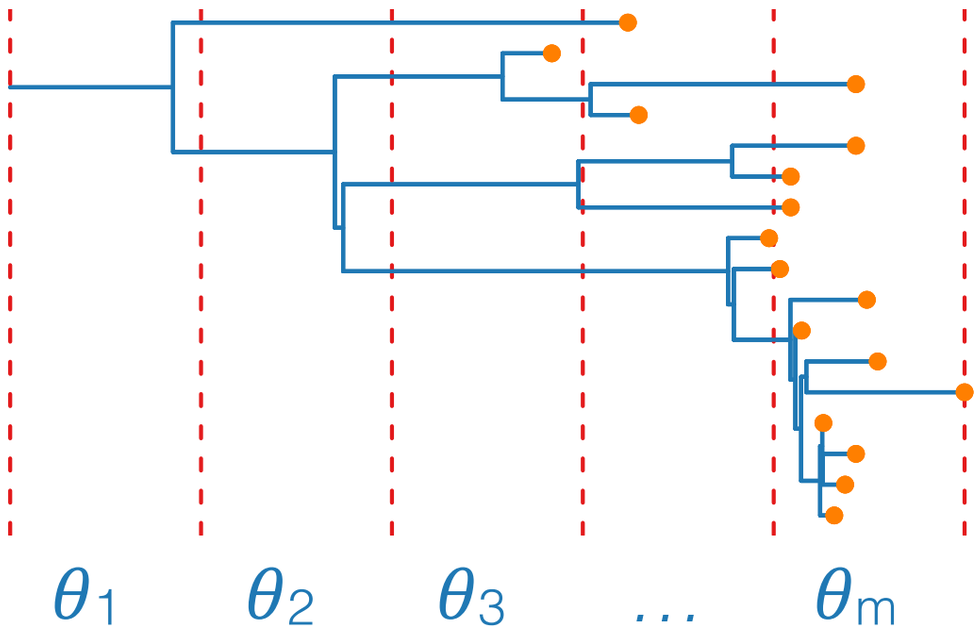
$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Covariates



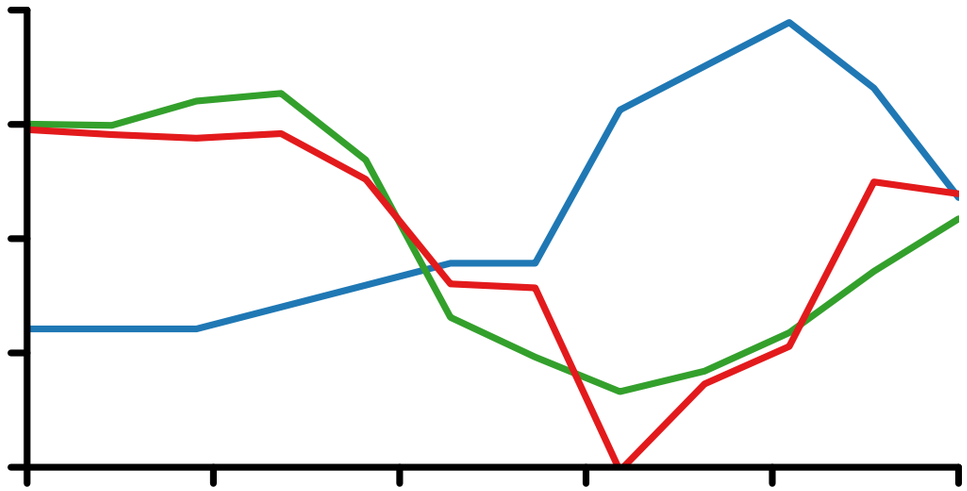
$$g(\theta_i) = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,k}\beta_k$$

Phylodynamic linear model formulation



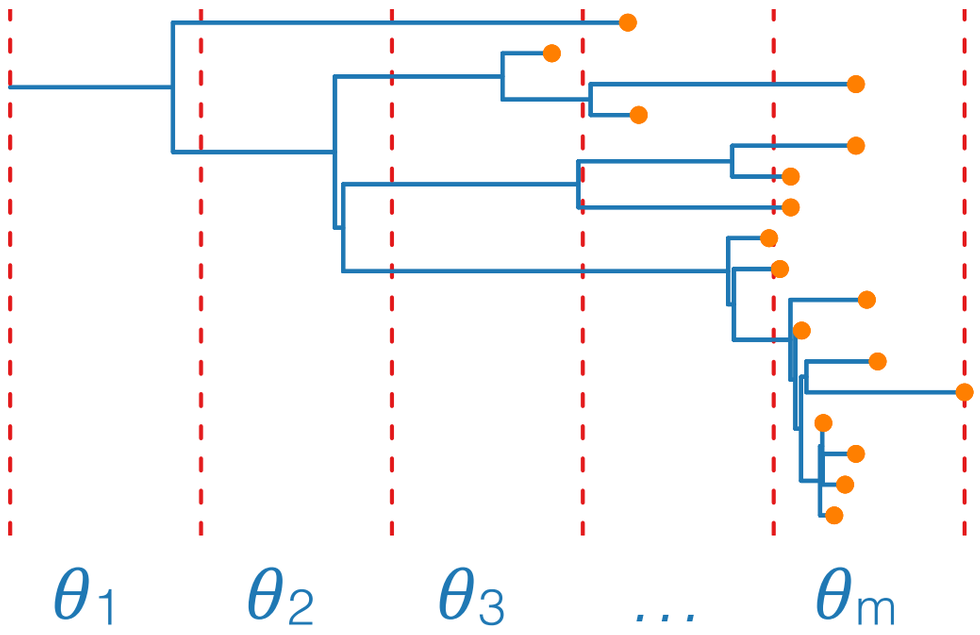
$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Covariates



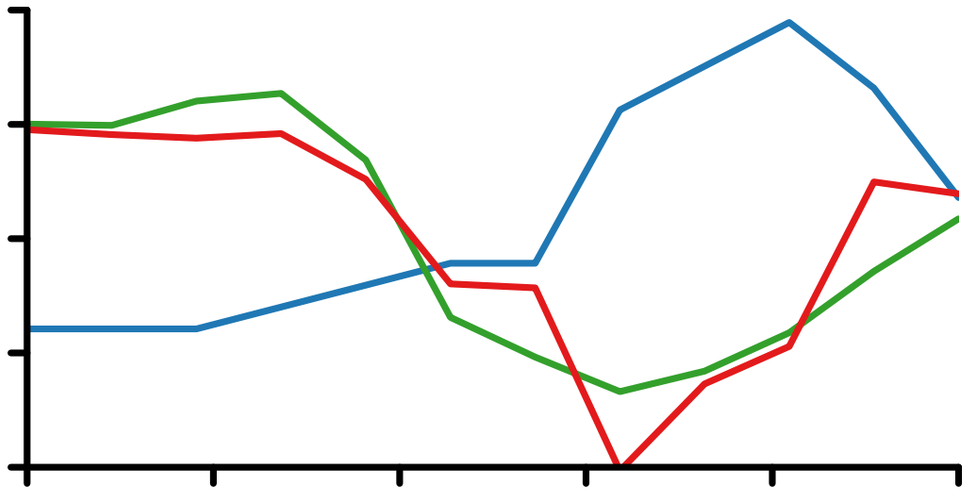
$$g(\theta_i) = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,k}\beta_k + \varepsilon_i$$

Phylodynamic linear model formulation



$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

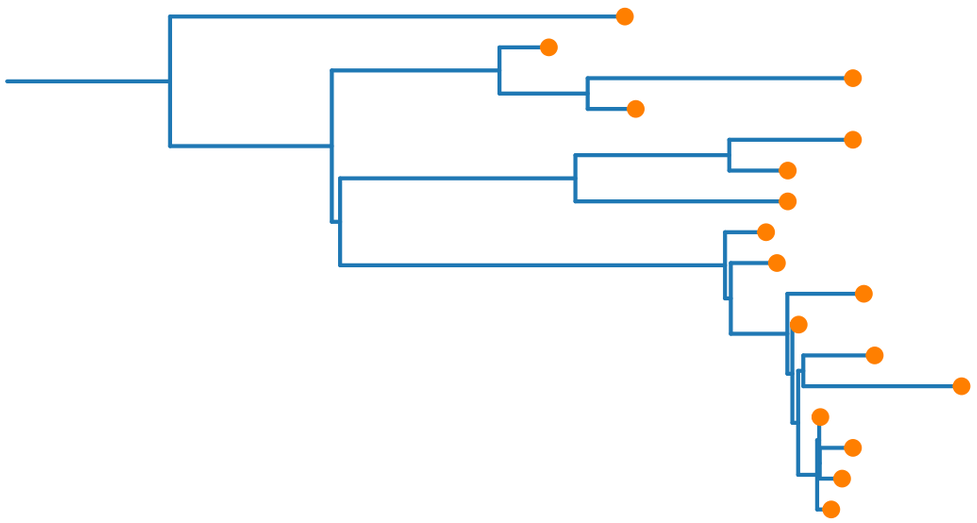
Covariates



$$g(\theta_i) = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,k}\beta_k + \varepsilon_i$$

$$g(\boldsymbol{\theta}) = \underset{[mx1]}{} \underset{[mxk]}{\boldsymbol{X}} \underset{[kx1]}{\boldsymbol{\beta}} + \underset{[mx1]}{\boldsymbol{\varepsilon}}$$

Linear model posteriors

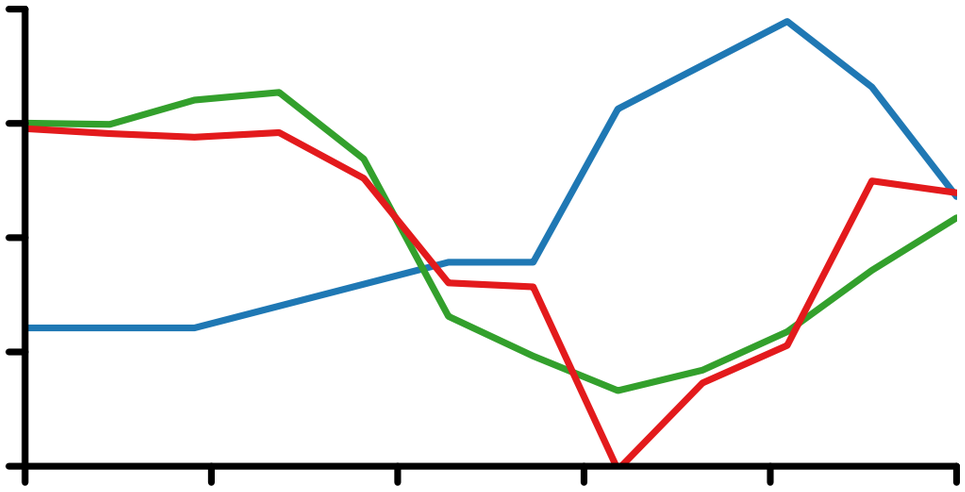


θ_1 θ_2 θ_3 ... θ_m

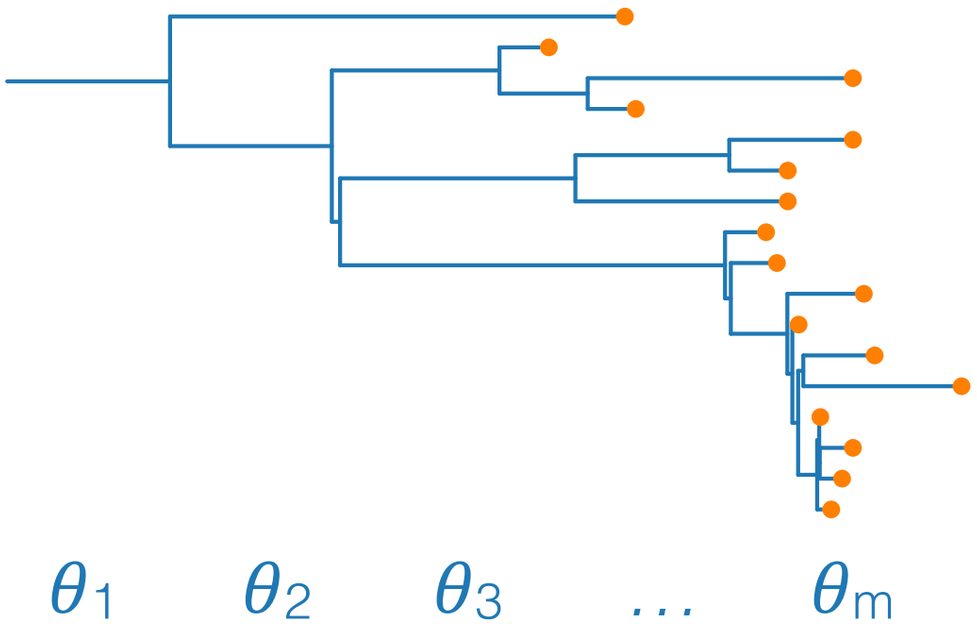
No linear model

$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Covariates



Linear model posteriors



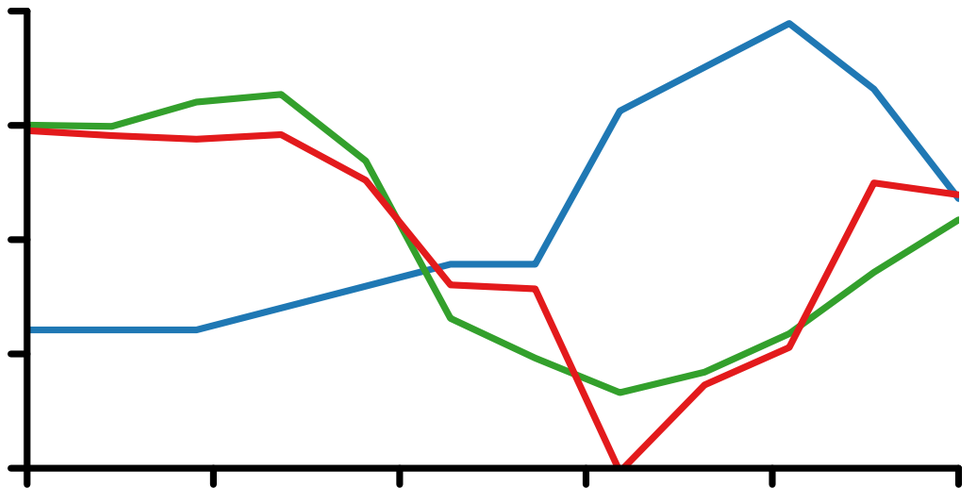
No linear model

$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

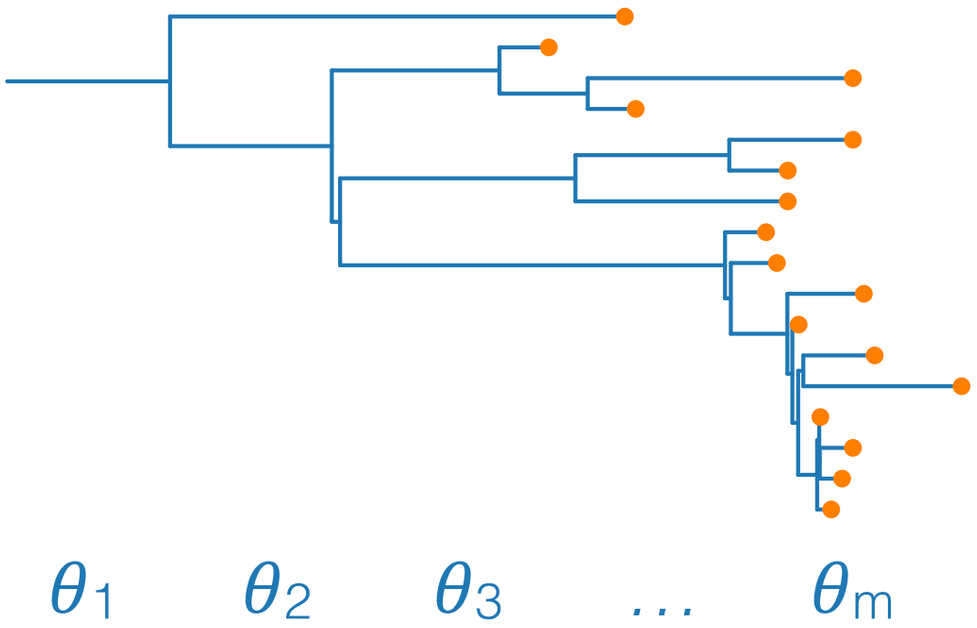
Full model: $g(\boldsymbol{\theta}) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(g(\boldsymbol{\theta}) \mid \boldsymbol{X}, \boldsymbol{\beta})\, P(\boldsymbol{\beta})$$

Covariates



Linear model posteriors



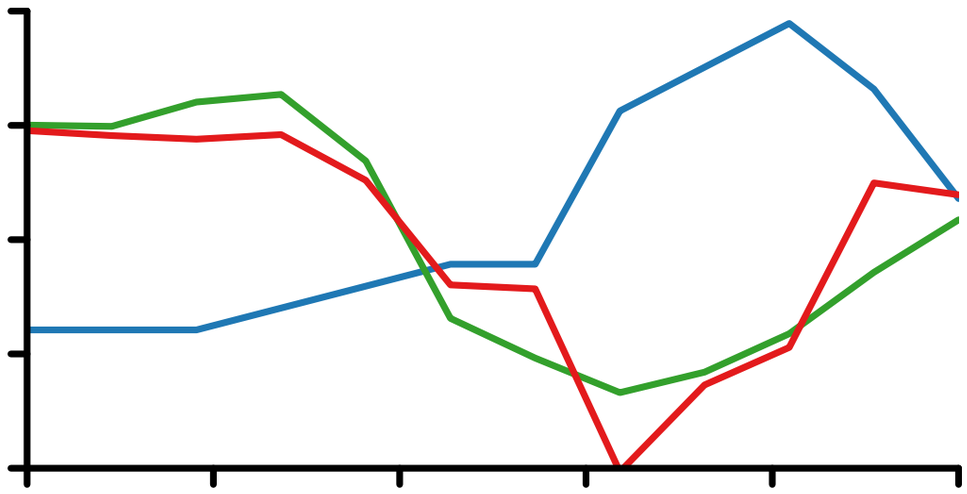
No linear model

$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Full model: $g(\boldsymbol{\theta}) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid \boldsymbol{\theta})\, P(g(\boldsymbol{\theta}) \mid \boldsymbol{X}, \boldsymbol{\beta})\, P(\boldsymbol{\beta})$$

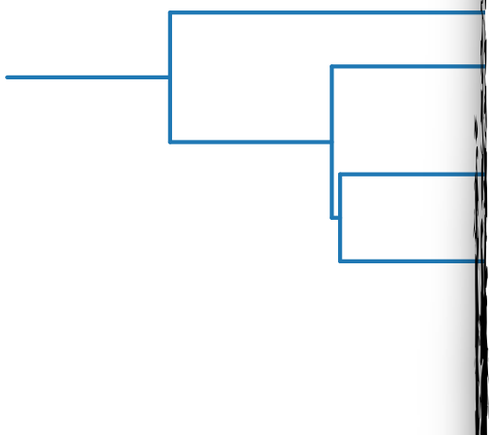
Covariates



Exact model: $g(\boldsymbol{\theta}) = \boldsymbol{X\beta}$ $\boldsymbol{\varepsilon} = \boldsymbol{0}$

$$P(\boldsymbol{\theta} \mid T) \propto P(T \mid g^{-1}(\boldsymbol{X\beta}))\, P(\boldsymbol{\beta})$$

Linear model posteriors



Parameter distributions

$$g(\boldsymbol{\theta}) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

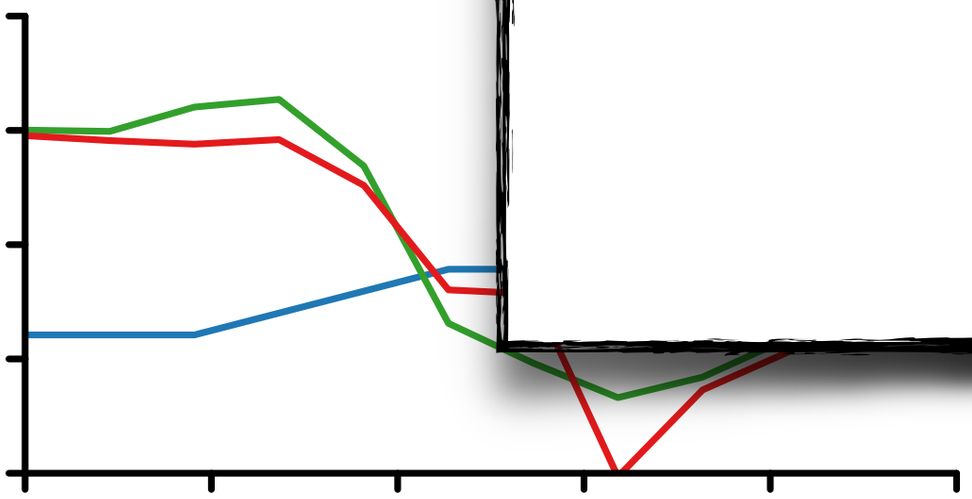
[mx1] [mxk] [kx1] [mx1]

***m* time points x *k* covariates**

θ_1 θ_2 θ

$\boldsymbol{X},\boldsymbol{\beta})\ P(\boldsymbol{\beta})$

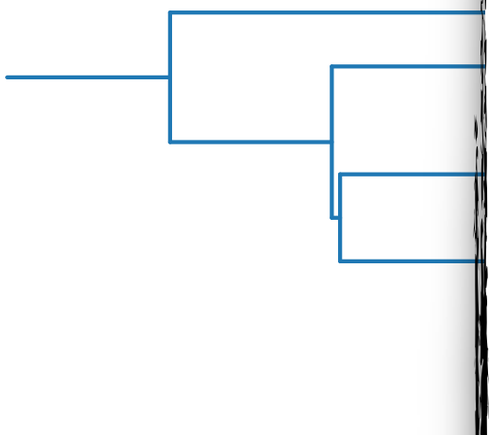
Covariates



$\boldsymbol{\varepsilon} = \boldsymbol{0}$

$(\boldsymbol{\beta})$

Linear model posteriors



Parameter distributions

$$g(\boldsymbol{\theta}) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

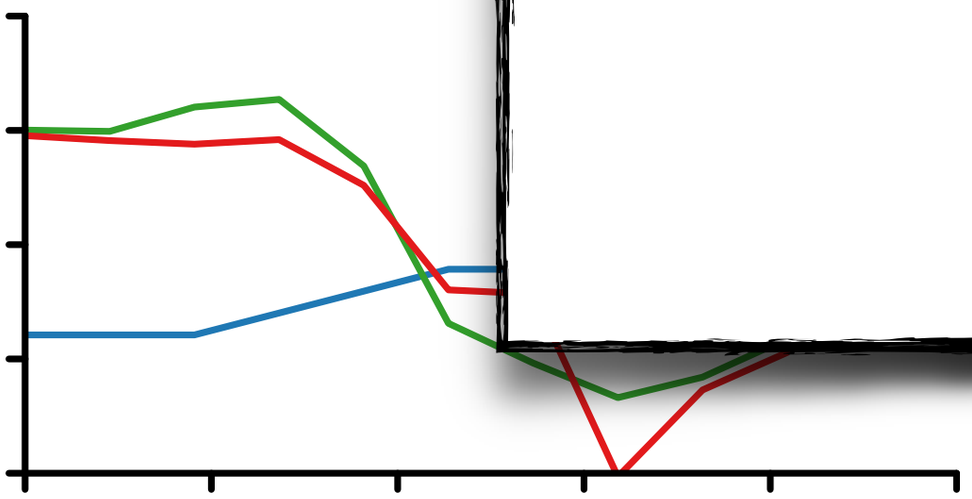
[mx1] [mxk] [kx1] [mx1]

***m* time points** **x** ***k* covariates**

$$\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$$

θ_1 θ_2 θ

Covariates

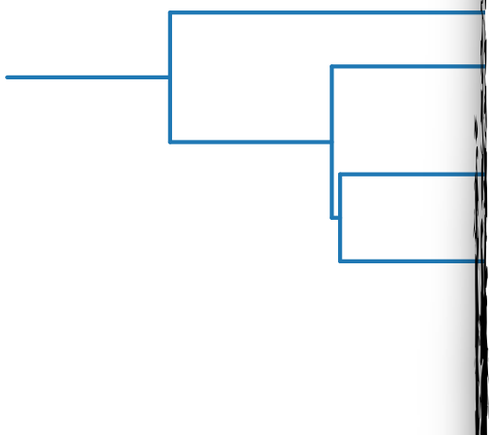


$\boldsymbol{X},\boldsymbol{\beta})\ P(\boldsymbol{\beta})$

$\boldsymbol{\varepsilon = 0}$

$(\boldsymbol{\beta})$

Linear model posteriors



Parameter distributions

$$g(\boldsymbol{\theta}) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

[mx1] [mxk] [kx1] [mx1]

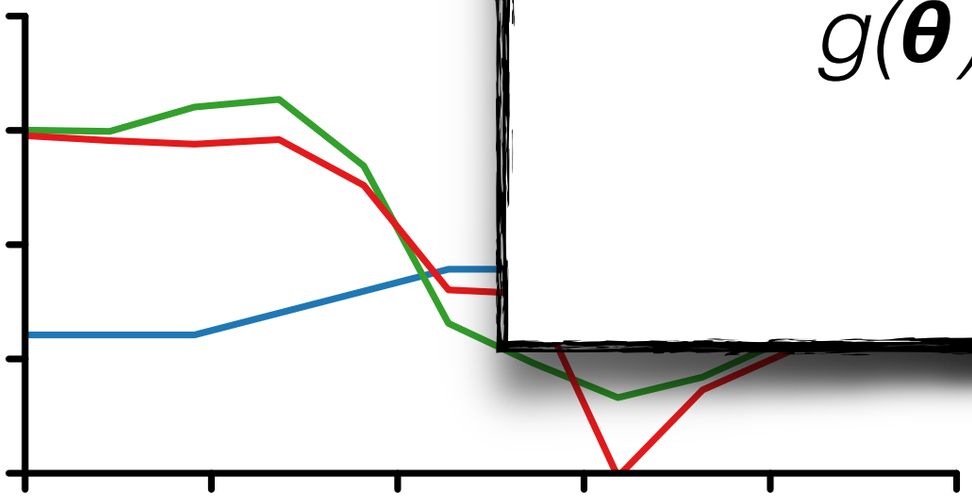
***m* time points** **x** ***k* covariates**

$$\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$$

$$g(\boldsymbol{\theta}) \sim \mathcal{N}_m(\boldsymbol{X\beta}, \sigma^2 I)$$

θ_1 θ_2 θ

Covariates

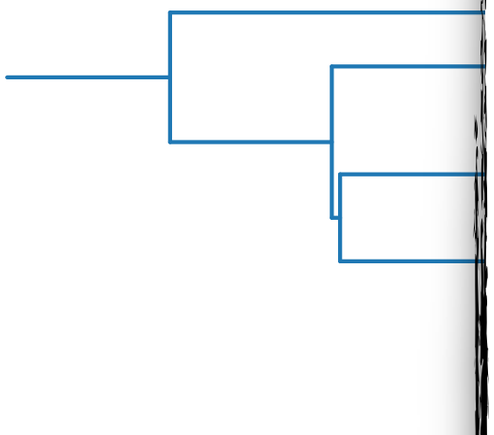


$\boldsymbol{X},\boldsymbol{\beta})\ P(\boldsymbol{\beta})$

$\boldsymbol{\varepsilon = 0}$

$(\boldsymbol{\beta})$

Linear model posteriors



Parameter distributions

$$g(\boldsymbol{\theta}) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

[mx1] [mxk] [kx1] [mx1]

***m* time points** **x** ***k* covariates**

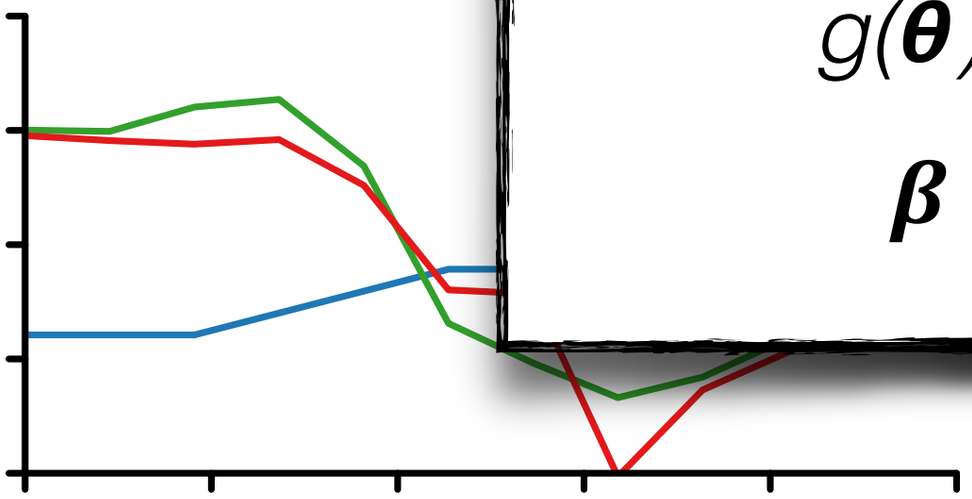
$$\varepsilon_i \sim \mathcal{N}(0,\sigma^2)$$

$$g(\boldsymbol{\theta}) \sim \mathcal{N}_m(\boldsymbol{X\beta}, \sigma^2 I)$$

$$\boldsymbol{\beta} \sim \mathcal{N}_k(\hat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$$

θ_1 θ_2 θ

Covariates



$\boldsymbol{X},\boldsymbol{\beta})\ P(\boldsymbol{\beta})$

$\boldsymbol{\varepsilon = 0}$

$(\boldsymbol{\beta})$

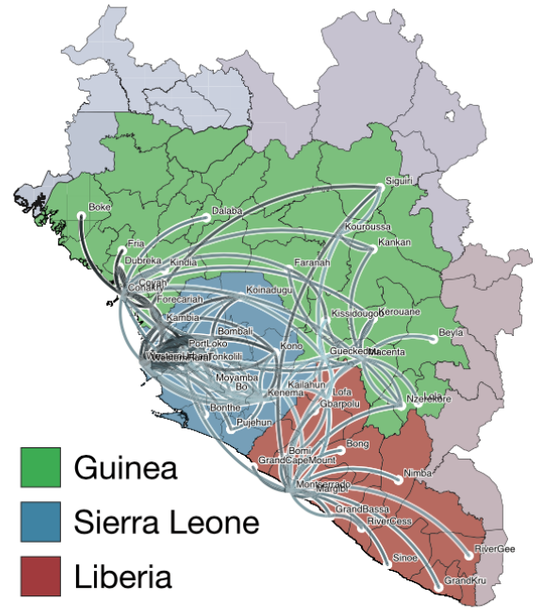
Spatial GLM model

OPEN ACCESS Freely available online

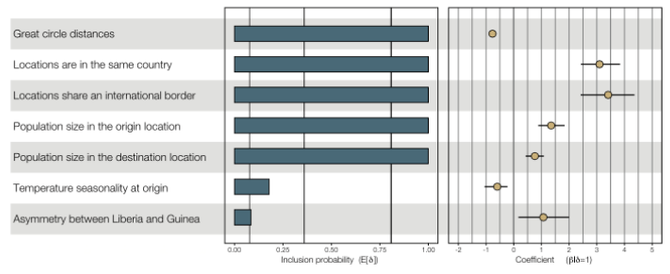
PLOS | PATHOGENS

Unifying Viral Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of Human Influenza H3N2

Philippe Lemey[1]*, Andrew Rambaut[2,3], Trevor Bedford[2], Nuno Faria[1], Filip Bielejec[1], Guy Baele[1], Colin A. Russell[4,5], Derek J. Smith[4,5,6], Oliver G. Pybus[7], Dirk Brockmann[8,9,10], Marc A. Suchard[11,12]



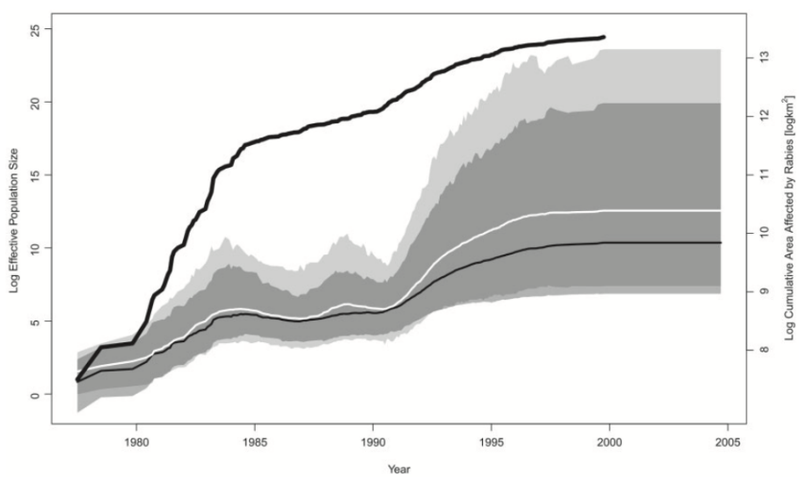
- Linear model for migration rates
- Assume $\boldsymbol{\varepsilon = 0}$ (exact model)



Lemey *et al.* **PLOS Pathogens 2014**

Suchard *et al.* **Virus Evolution 2018**

Skygrid GLM



Gill *et al.* **Systematic Biology 2016**

Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates

MANDEV S. GILL[1], PHILIPPE LEMEY[2], SHANNON N. BENNETT[3], ROMAN BIEK[4], AND MARC A. SUCHARD[5,6,7,*]

- Linear model for changes in N_e over time
- Assume $\boldsymbol{\varepsilon} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma})$

Penalised regression

The Bayesian Lasso

Trevor PARK and George CASELLA

The Lasso estimate for linear regression parameters can be interpreted as a Bayesian posterior mode estimate when the regression parameters have independent Laplace (i.e., double-exponential) priors. Gibbs sampling from this posterior is possible using an expanded hierarchy with conjugate normal priors for the regression parameters and independent exponential priors on their variances. A connection with the inverse-Gaussian distribution provides tractable full conditional distributions. The Bayesian Lasso provides interval estimates (Bayesian credible intervals) that can guide variable selection. Moreover, the structure of the hierarchical model provides both Bayesian and likelihood methods for selecting the Lasso parameter. Slight modifications lead to Bayesian versions of other Lasso-related estimation methods, including bridge regression and a robust variant.

KEY WORDS: Empirical Bayes; Gibbs sampler; Hierarchical model; Inverse Gaussian; Linear regression; Penalized regression; Scale mixture of normals.

Park and Casella **Journal of the American Statistical Association 2008**

Penalised regression

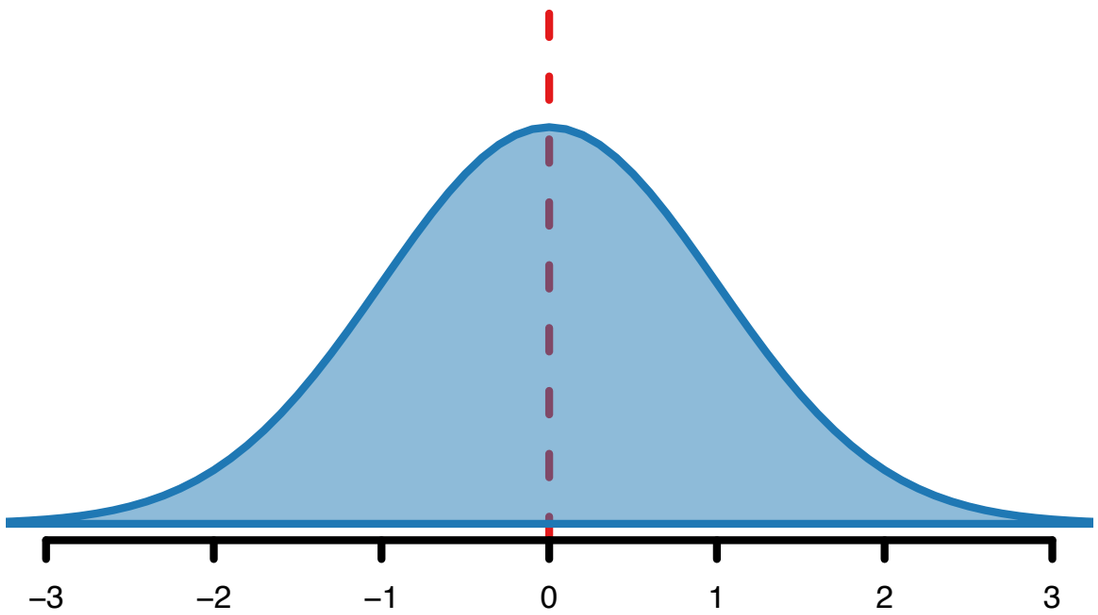
The Bayesian Lasso

Trevor PARK and George CASELLA

The Lasso estimate for linear regression parameters can be interpreted as a Bayesian posterior mode estimate when the regression parameters have independent Laplace (i.e., double-exponential) priors. Gibbs sampling from this posterior is possible using an expanded hierarchy with conjugate normal priors for the regression parameters and independent exponential priors on their variances. A connection with the inverse-Gaussian distribution provides tractable full conditional distributions. The Bayesian Lasso provides interval estimates (Bayesian credible intervals) that can guide variable selection. Moreover, the structure of the hierarchical model provides both Bayesian and likelihood methods for selecting the Lasso parameter. Slight modifications lead to Bayesian versions of other Lasso-related estimation methods, including bridge regression and a robust variant.

KEY WORDS: Empirical Bayes; Gibbs sampler; Hierarchical model; Inverse Gaussian; Linear regression; Penalized regression; Scale mixture of normals.

Park and Casella **Journal of the American Statistical Association 2008**



Penalised regression

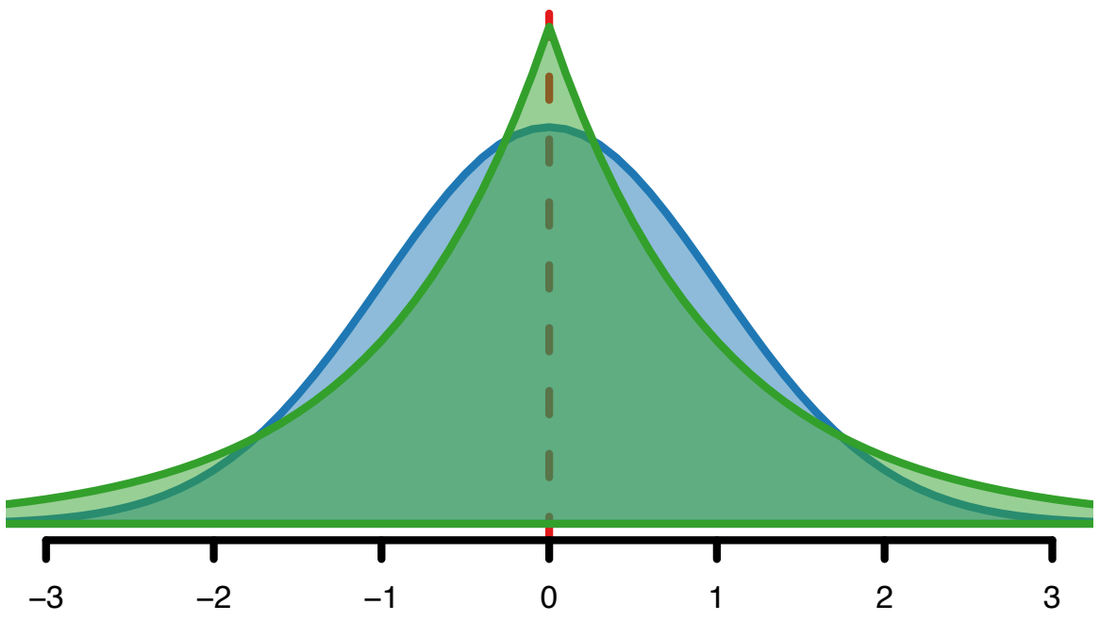
The Bayesian Lasso

Trevor PARK and George CASELLA

The Lasso estimate for linear regression parameters can be interpreted as a Bayesian posterior mode estimate when the regression parameters have independent Laplace (i.e., double-exponential) priors. Gibbs sampling from this posterior is possible using an expanded hierarchy with conjugate normal priors for the regression parameters and independent exponential priors on their variances. A connection with the inverse-Gaussian distribution provides tractable full conditional distributions. The Bayesian Lasso provides interval estimates (Bayesian credible intervals) that can guide variable selection. Moreover, the structure of the hierarchical model provides both Bayesian and likelihood methods for selecting the Lasso parameter. Slight modifications lead to Bayesian versions of other Lasso-related estimation methods, including bridge regression and a robust variant.

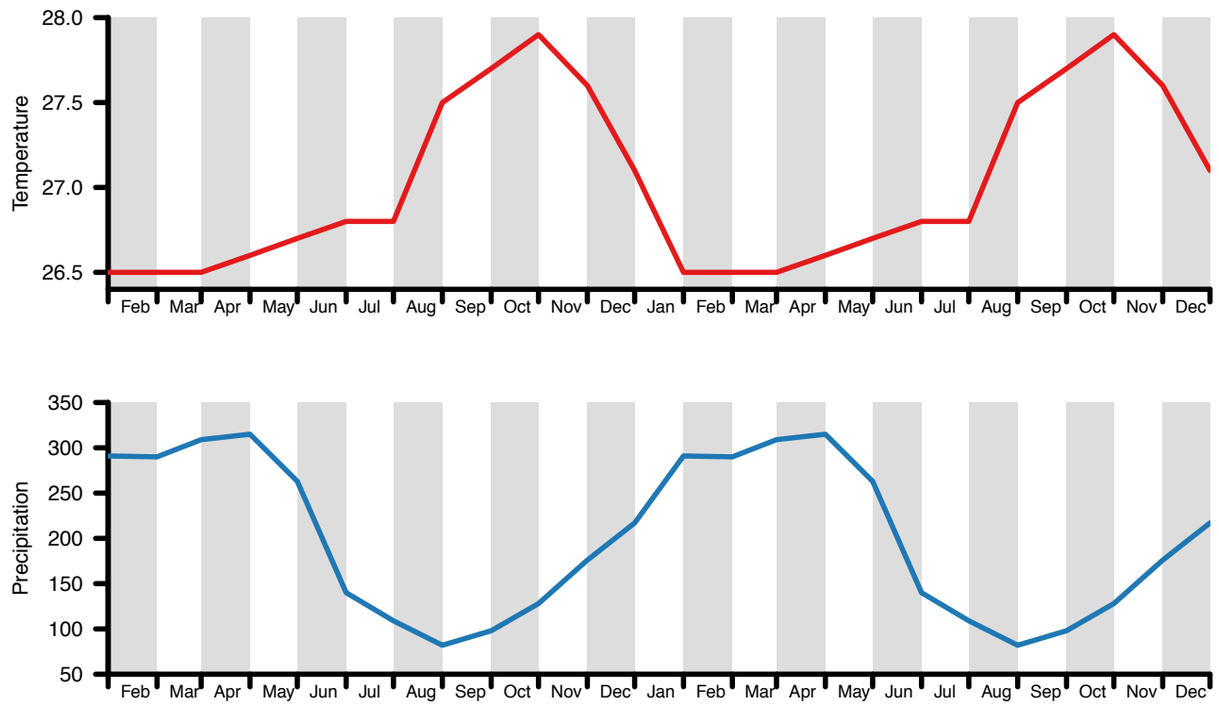
KEY WORDS: Empirical Bayes; Gibbs sampler; Hierarchical model; Inverse Gaussian; Linear regression; Penalized regression; Scale mixture of normals.

Park and Casella **Journal of the American Statistical Association 2008**



Lagged covariates

- Delays between covariate values and effects on θ
- Covariates may be measured at more time points than sequence data (especially climatic covariates)

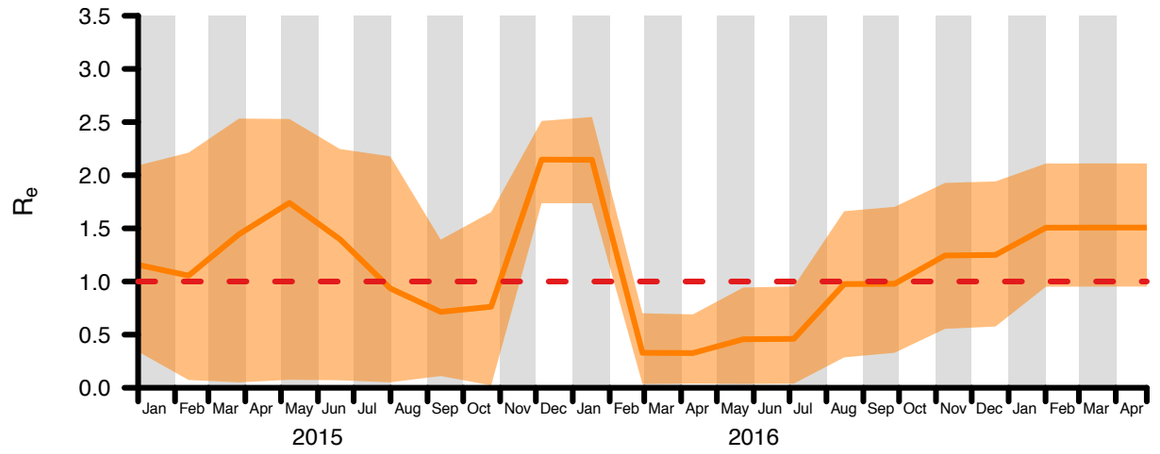


Introduce a timelag parameter:

$$g(\theta_i) = x_{i-s_1,1}\beta_1 + x_{i-s_2,2}\beta_2 + \varepsilon_i$$

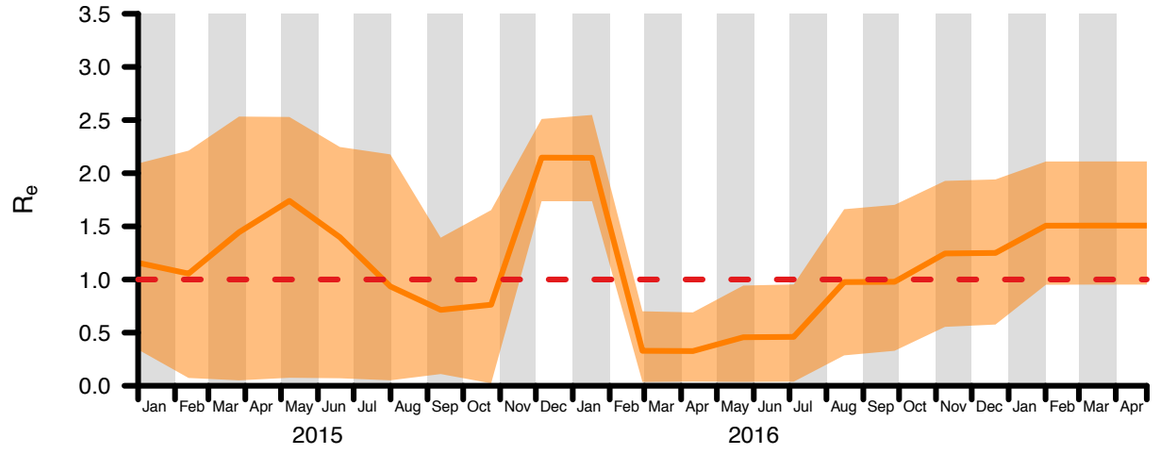
Zika virus in Manaus (57 genomes over 2 years)

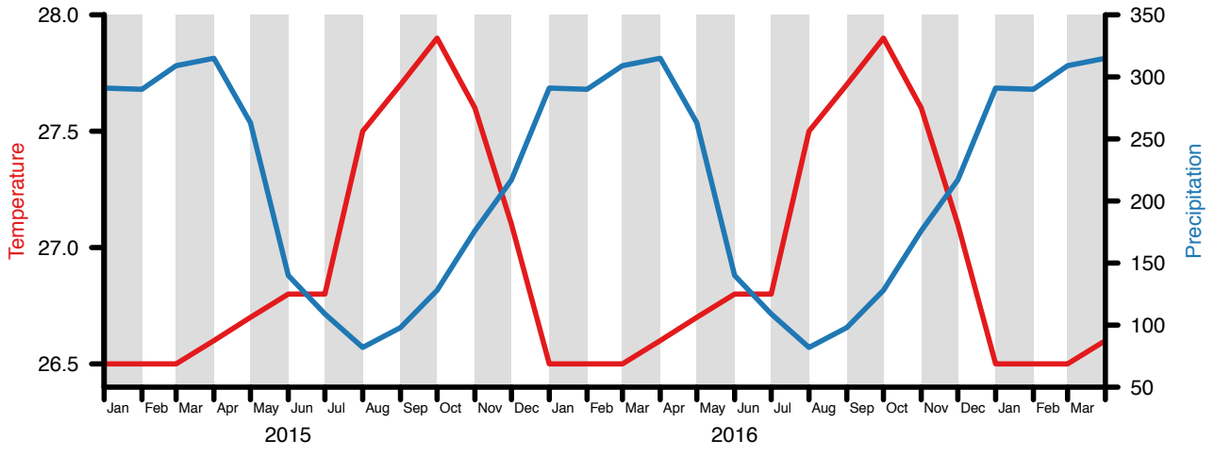
No linear model ▶



Zika virus in Manaus (57 genomes over 2 years)

No linear model ▶

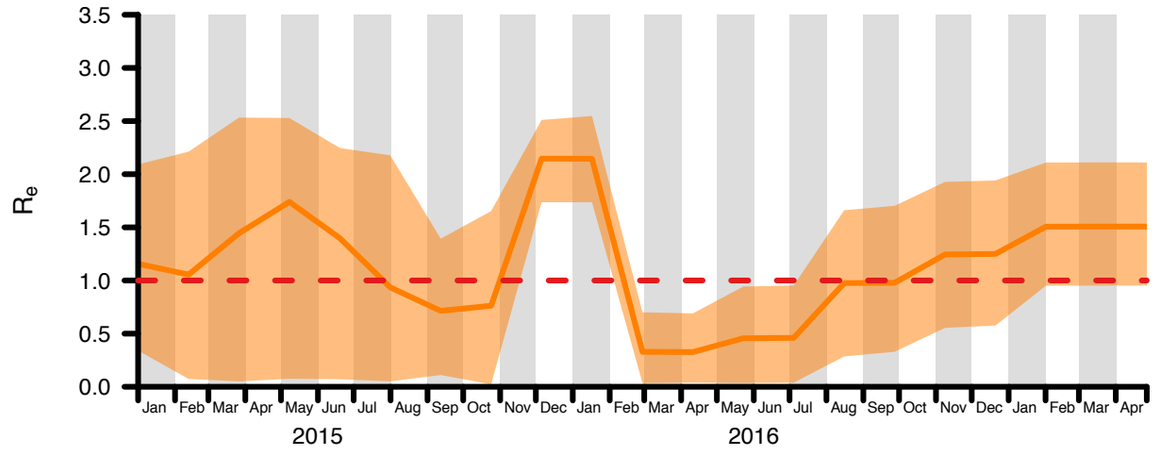


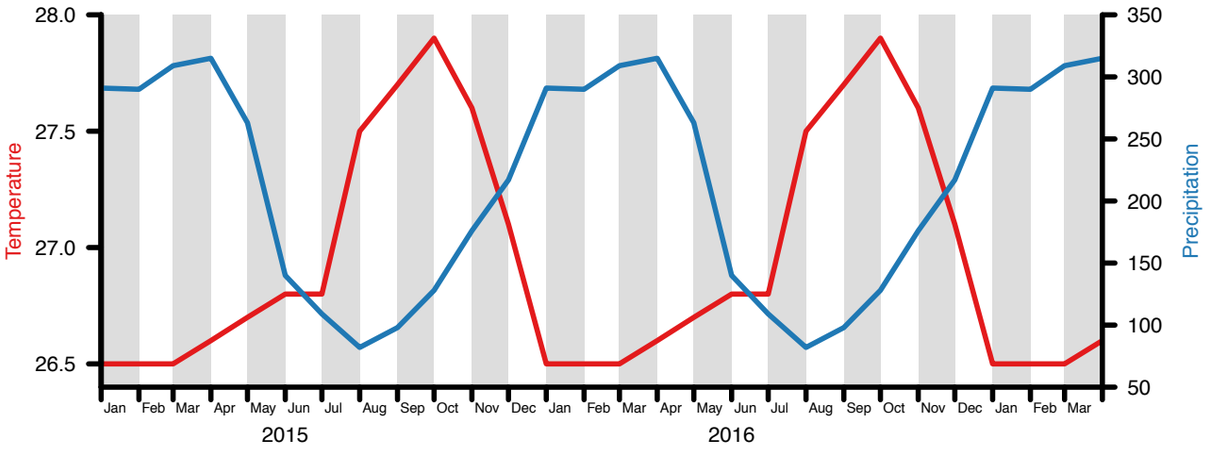


◀ Use rainfall and temperature as covariates

Zika virus in Manaus (57 genomes over 2 years)

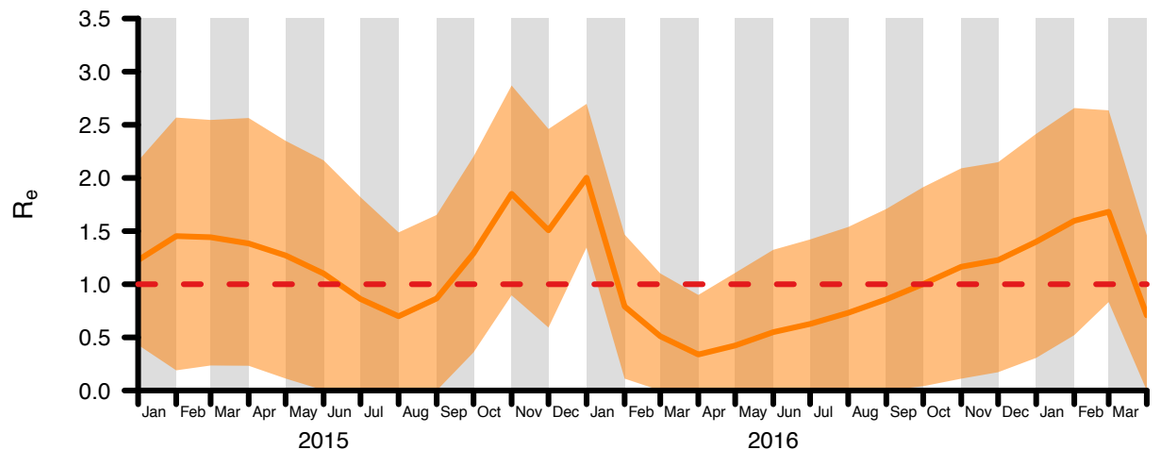
No linear model ▶





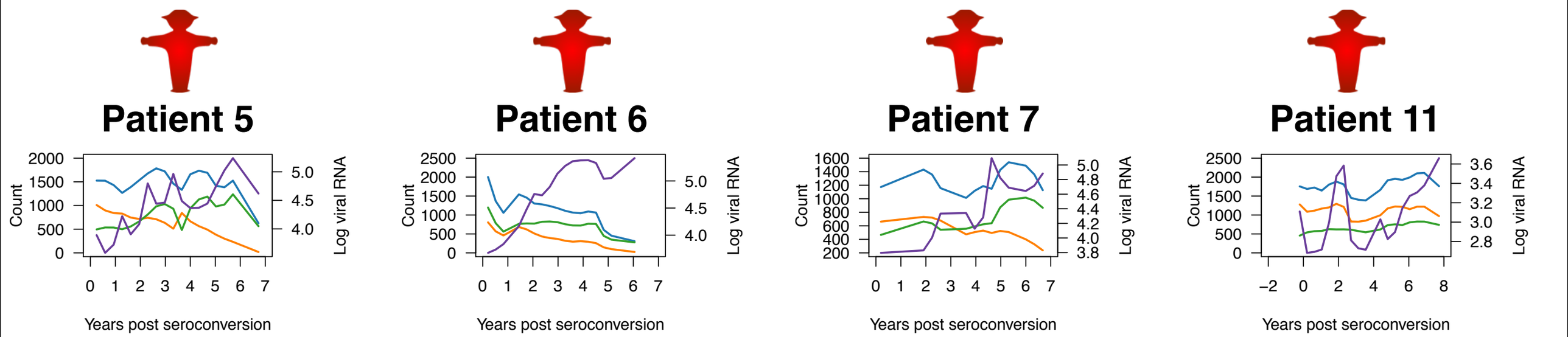
◀ Use rainfall and temperature as covariates

With linear model ▶



Hierarchical within-host model

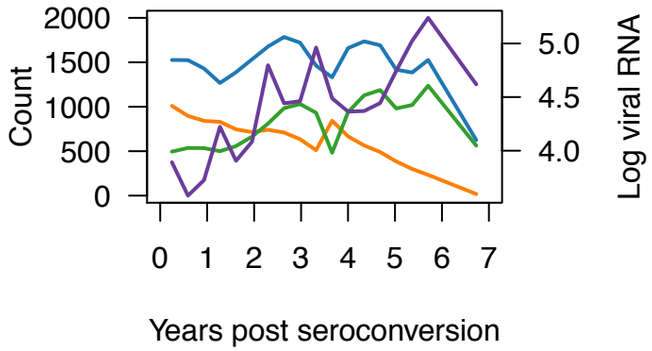
- Within-host HIV sequences and covariates from different patients sampled over time (Shankarappa et al. **Journal of Virology 1999**)
- Hierarchical linear model for R_e across multiple patients

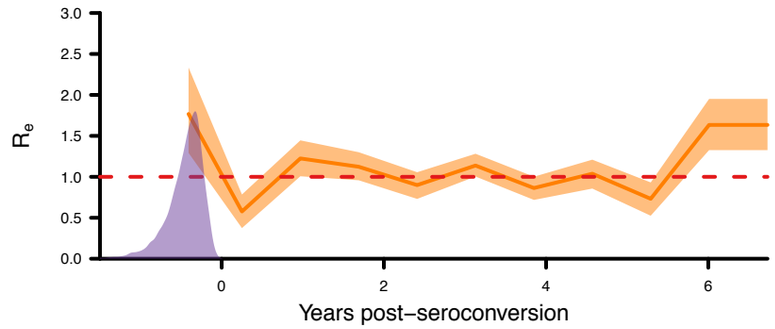


$$log(R_e) = log(CD3_{i,p})\beta_{CD3} + log(CD4_{i,p})\beta_{CD4} + log(CD8_{i,p})\beta_{CD8} + log(RNA_{i,p})\beta_{RNA} + \varepsilon_i$$

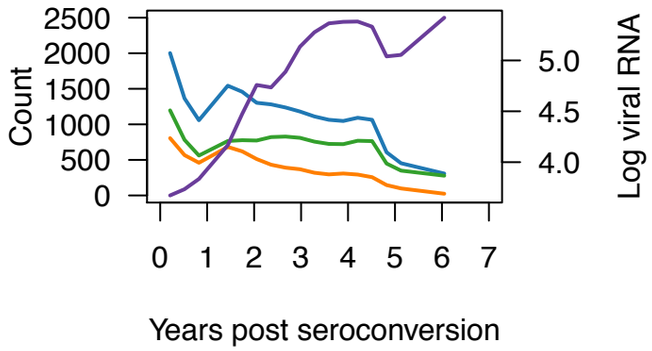
Hierarchical within-host model results

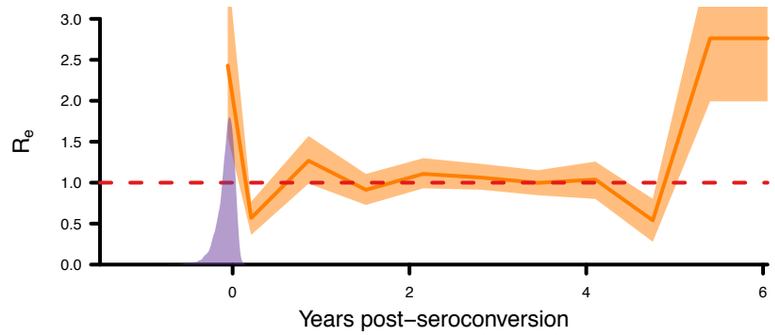
Patient 5



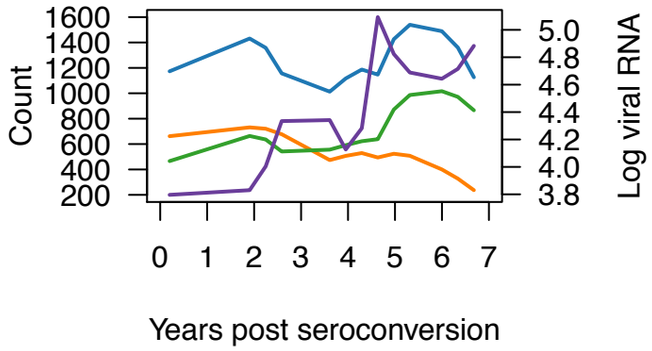


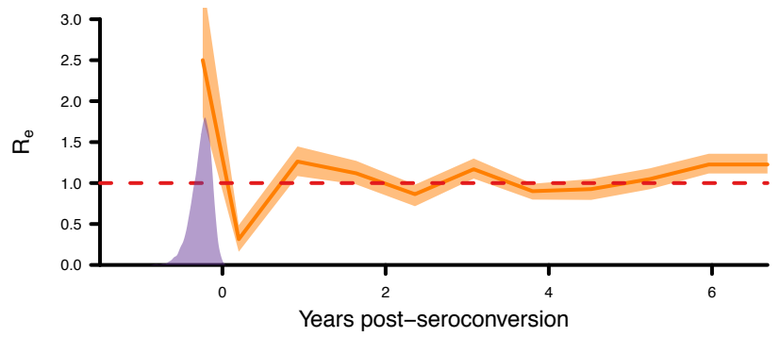
Patient 6



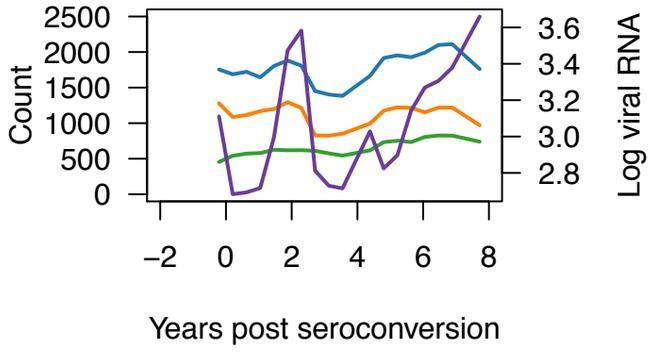


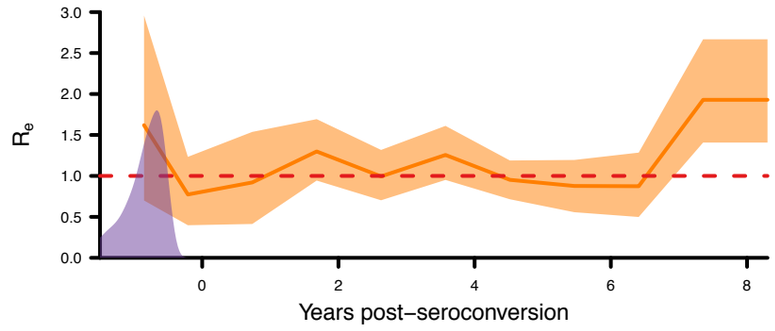
Patient 7





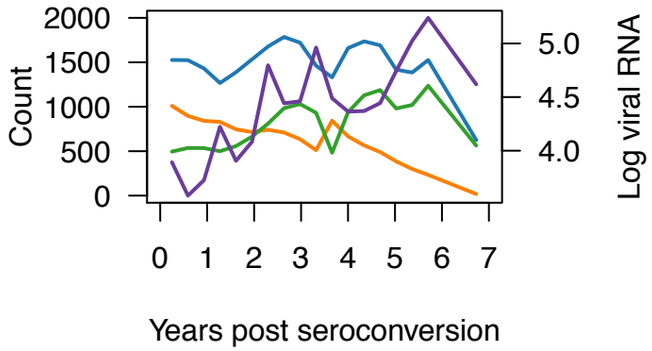
Patient 11

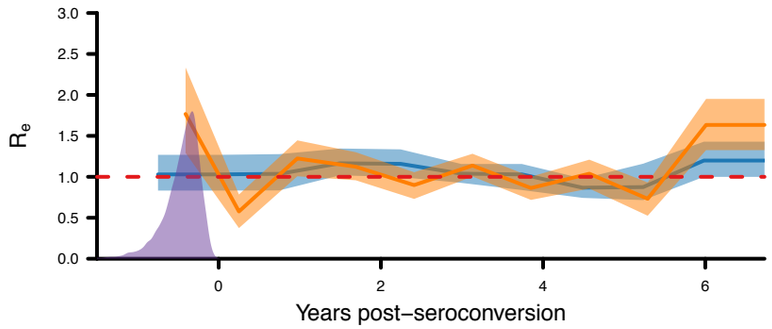




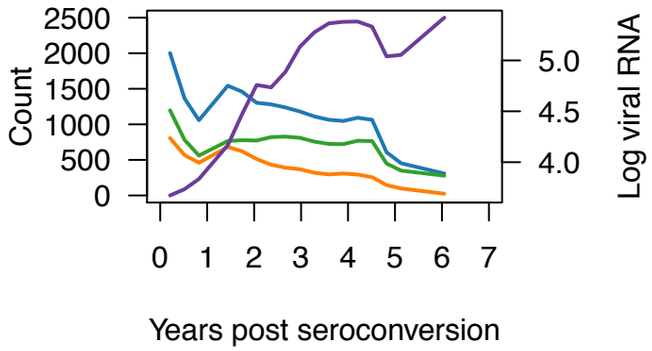
Hierarchical within-host model results

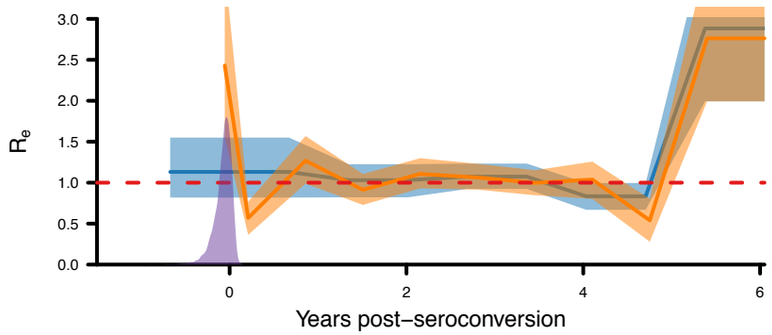
Patient 5



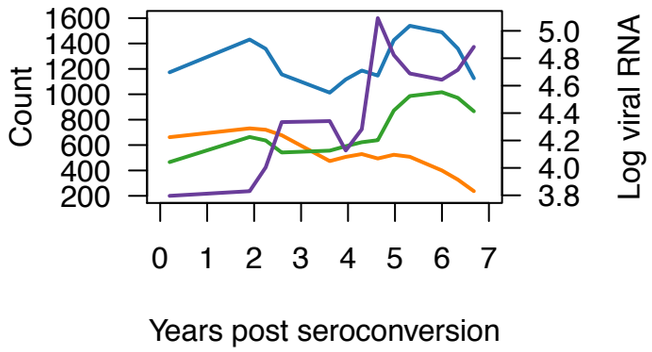


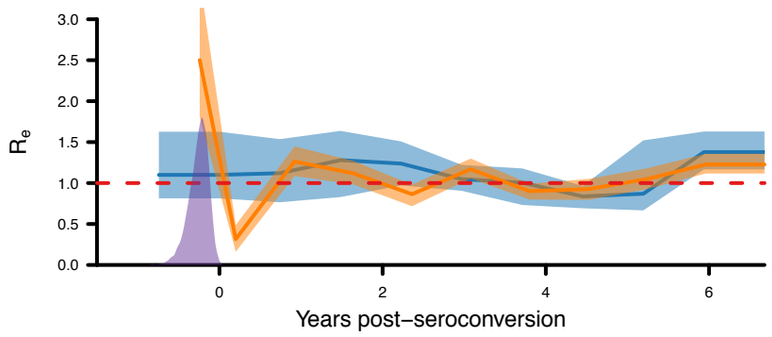
Patient 6



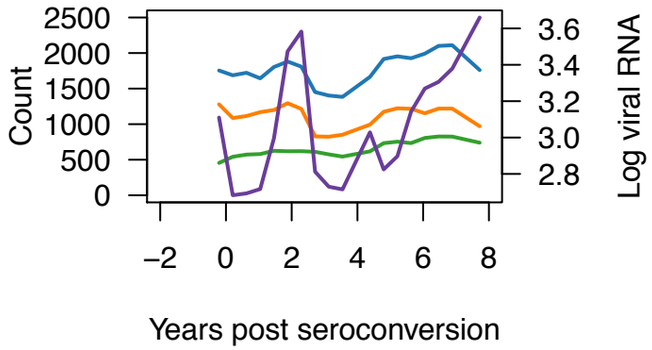


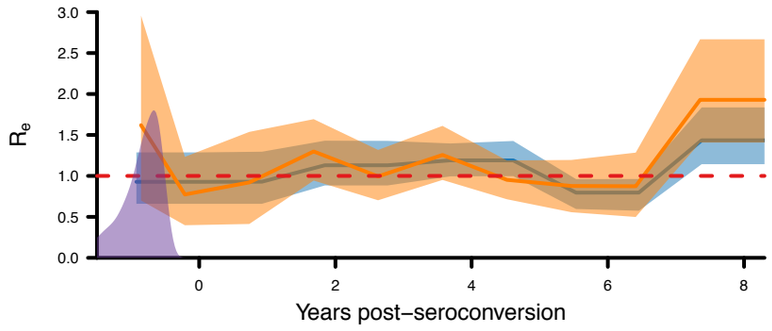
Patient 7





Patient 11

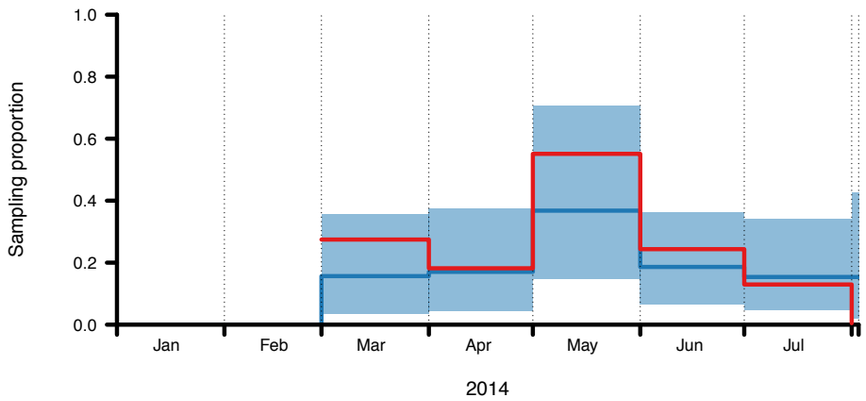


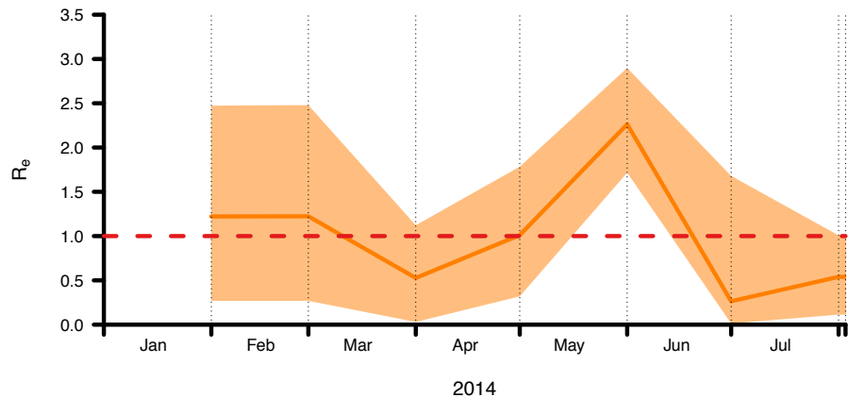


West African Ebola early epidemic

- 244 genomes sampled and 1424 cases reported until 4 August 2014 (WHO declares international public health emergency)
- Sampling proportion varies a lot over time
- Use linear model to put a multivariate prior on sampling proportion: $p_i = (sequences_i/cases_i)\beta + \varepsilon_i$

Monthly

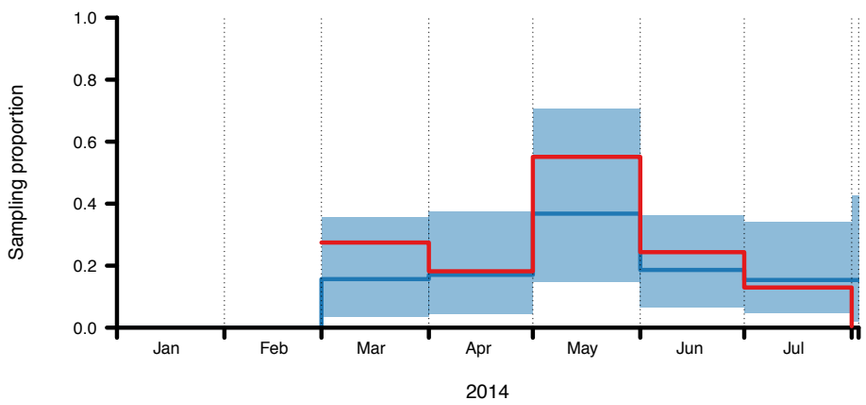


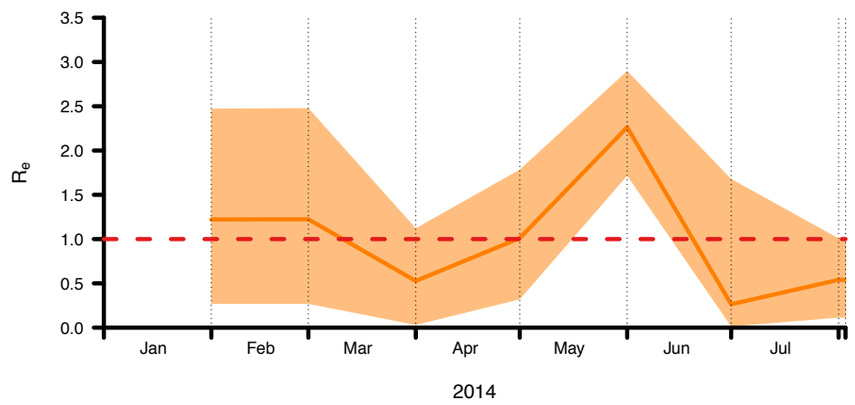


West African Ebola early epidemic

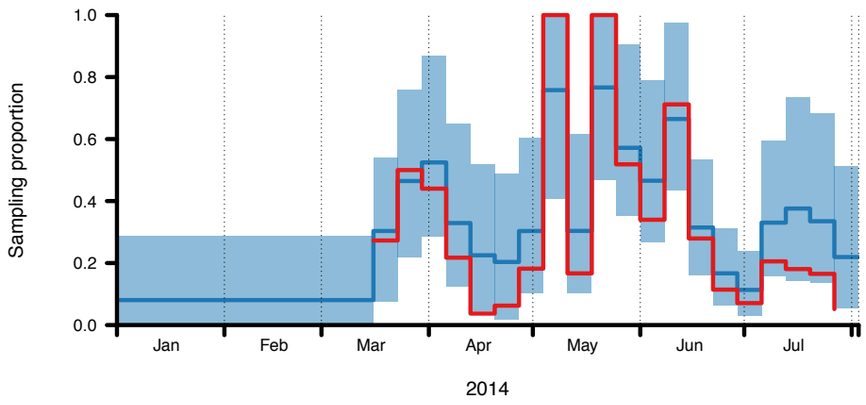
- 244 genomes sampled and 1424 cases reported until 4 August 2014 (WHO declares international public health emergency)
- Sampling proportion varies a lot over time
- Use linear model to put a multivariate prior on sampling proportion: $p_i = (sequences_i/cases_i)\beta + \varepsilon_i$

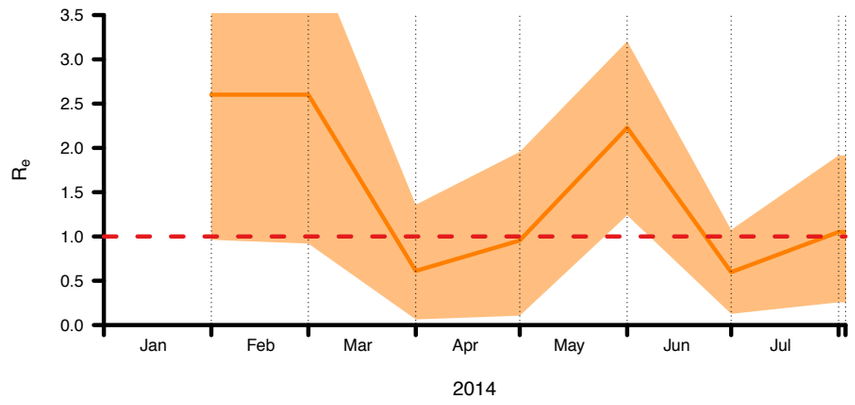
Monthly





Weekly





Conclusions and future directions

- As data from covariates keeps growing we can use it to **restrict** the **model space** and obtain more **accurate** answers
- Need to have some prior knowledge to **pick appropriate covariates**

Conclusions and future directions

- As data from covariates keeps growing we can use it to **restrict** the **model space** and obtain more **accurate** answers
- Need to have some prior knowledge to **pick appropriate covariates**

- **Extension I:**
  Use indicator variables to switch off covariates with poor predictive capacity (**remove** unnecessary **covariates**)
- **Extension II:**
  Add autocorrelation between residuals
  (e.g. model ε as a zero-mean **Gaussian process**)
- **Extension III:**
  Use maximum-likelihood estimate of β for better proposals

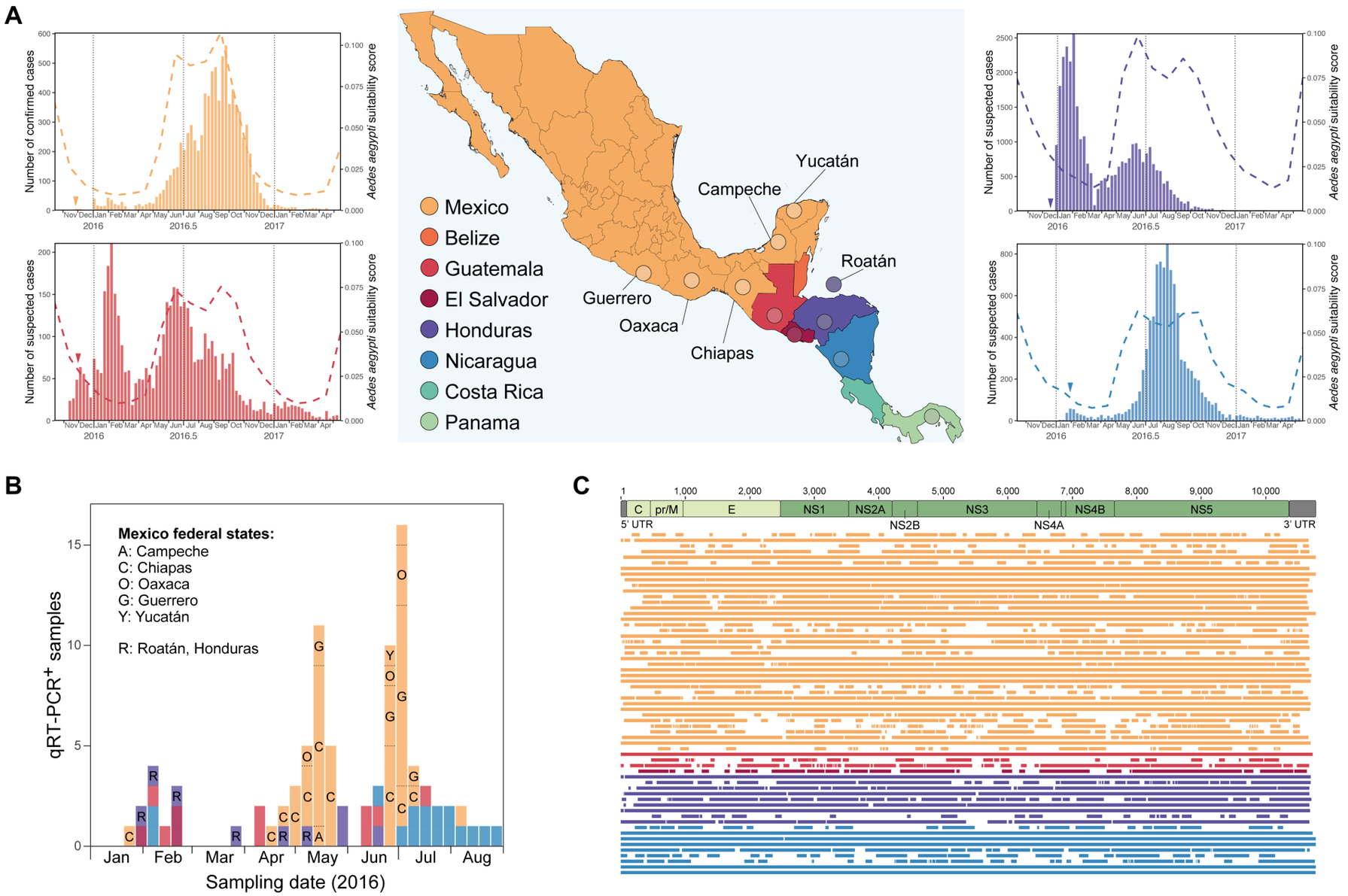
THANK YOU FOR LISTENING!

THANKS TO:

Oliver Pybus, Julien Thézé, Nuno Faria, Kris Parag

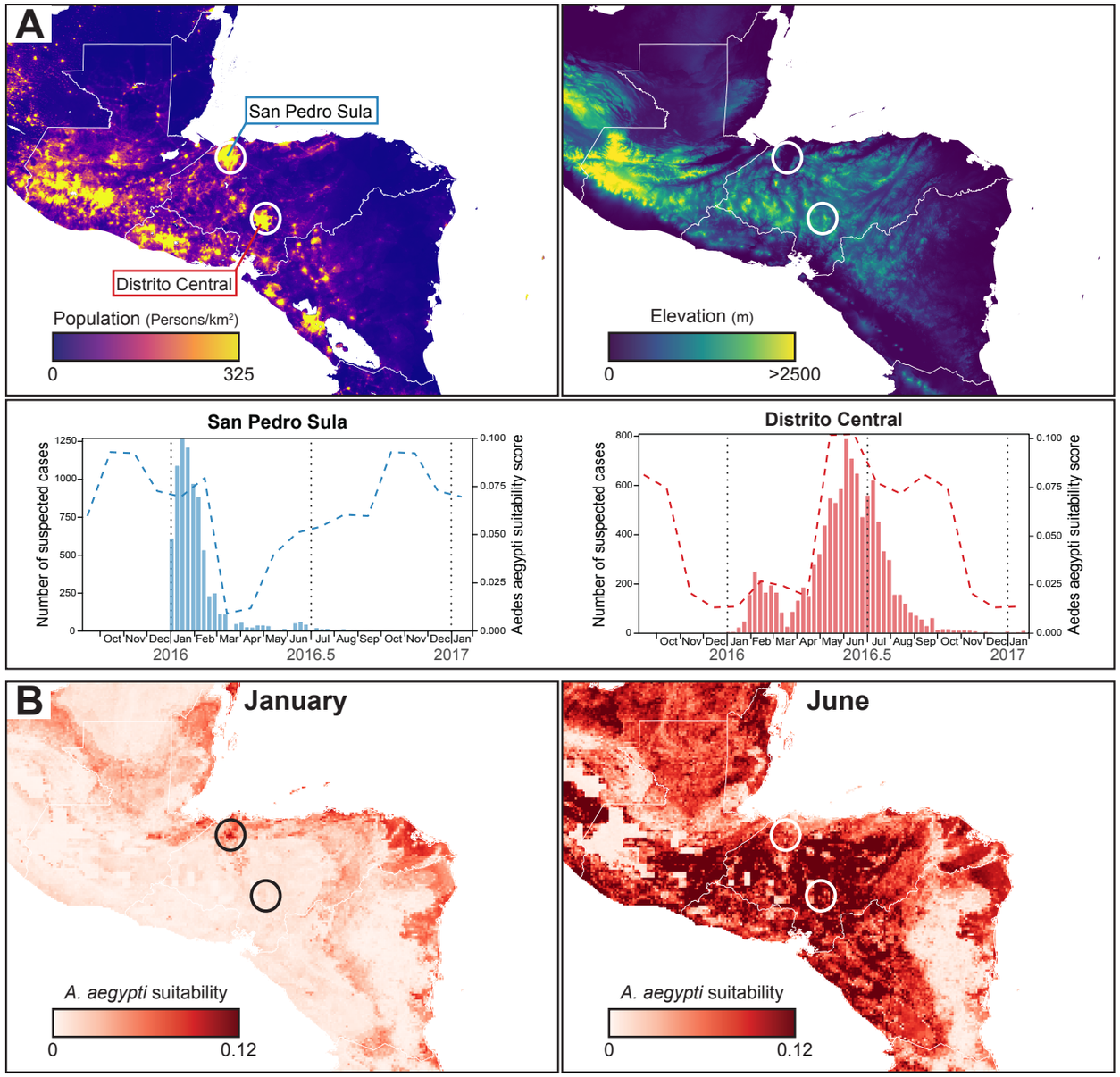
QUESTIONS?

Zika virus in Central America



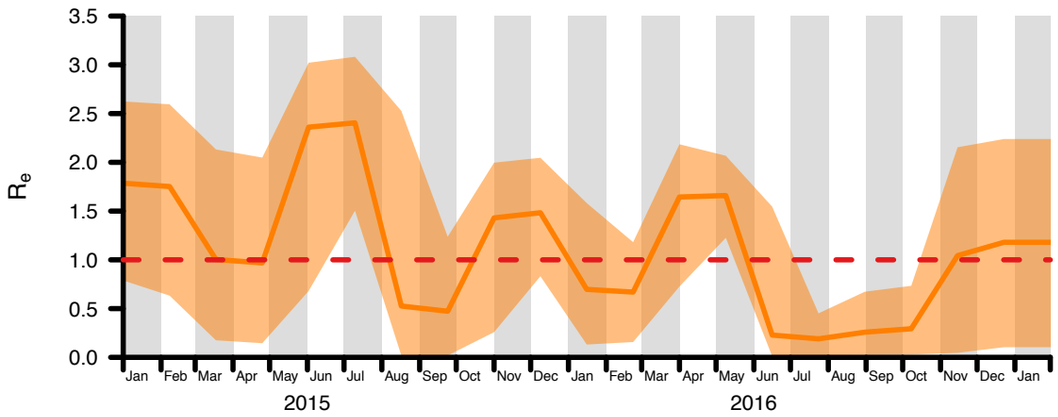
Thézé, Li, du Plessis *et al.* **Cell H&M 2018**

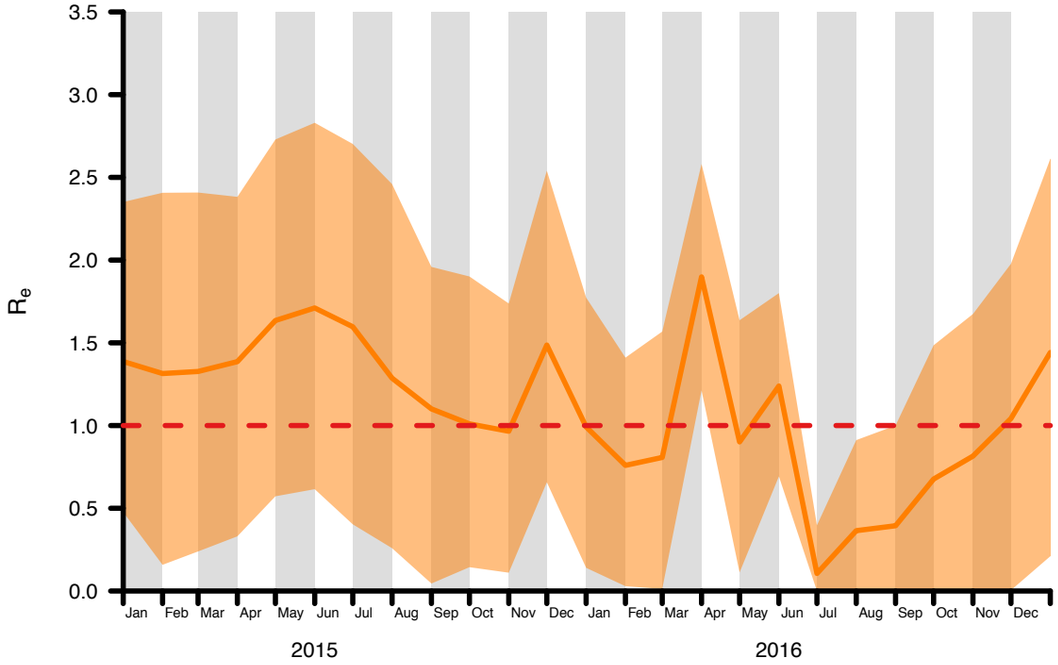
Zika virus in Central America



Thézé, Li, du Plessis *et al.* **Cell H&M 2018**

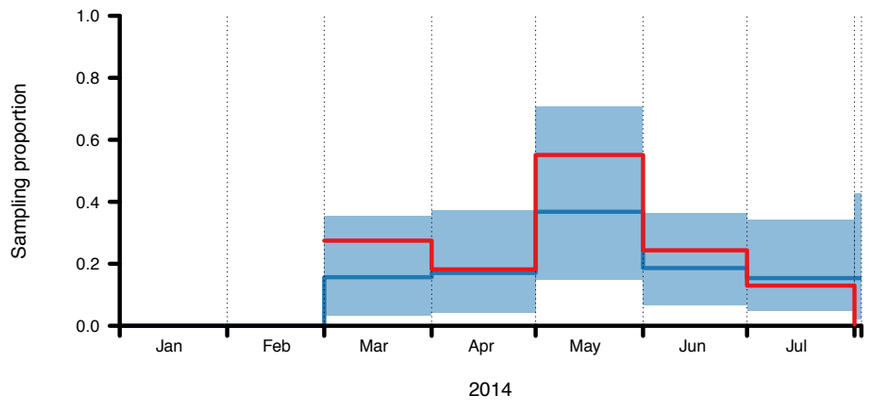
Zika virus in Central America

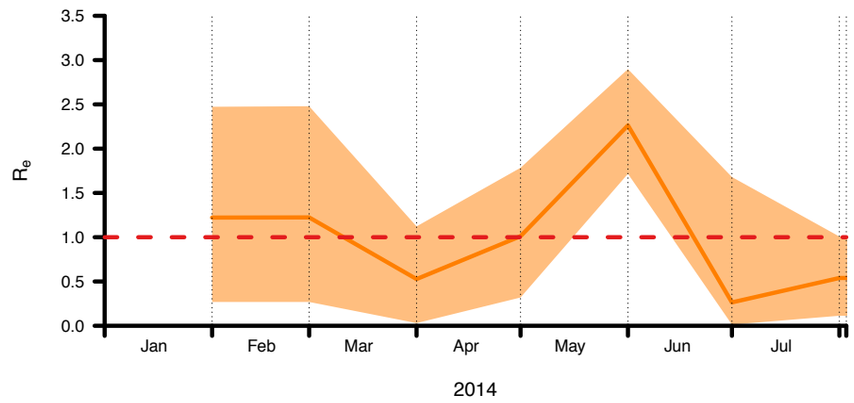




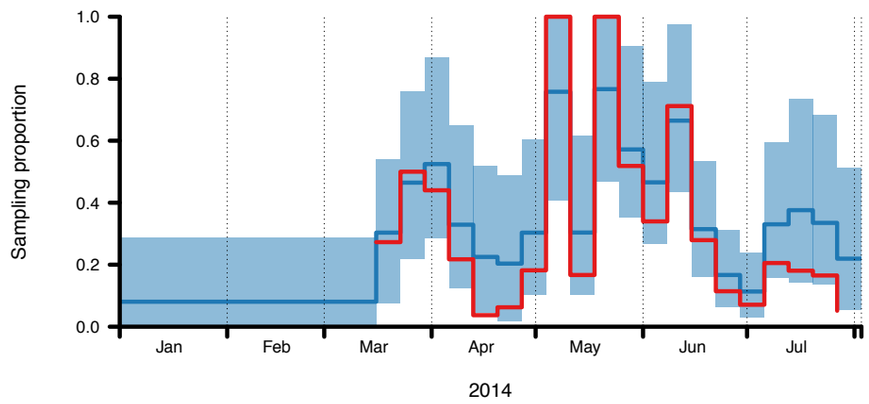
West African Ebola early outbreak

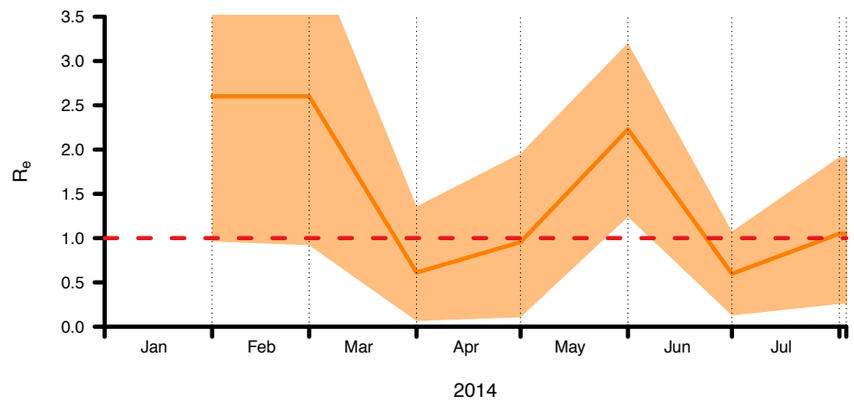
Monthly





Weekly





No linear model