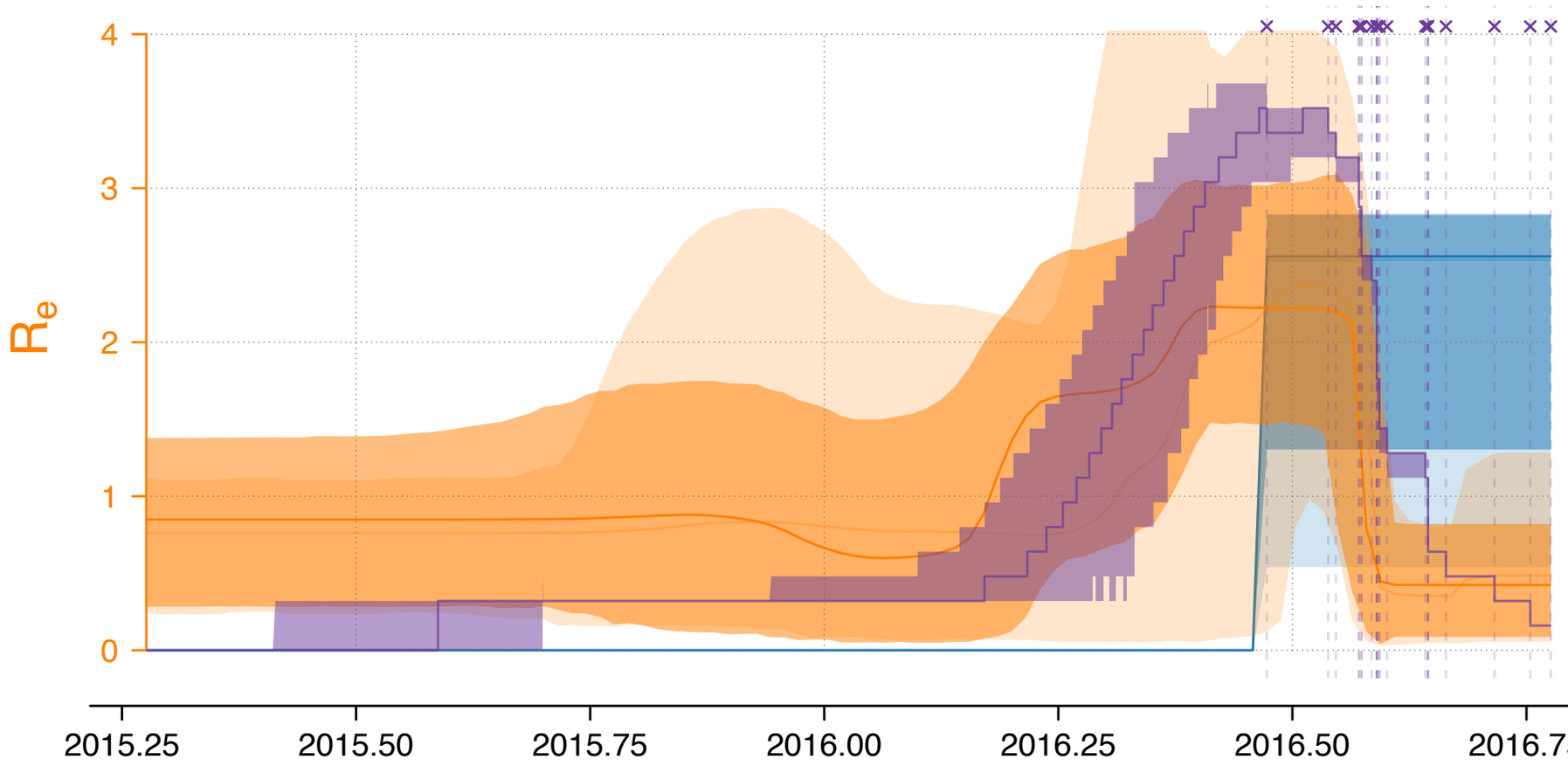
Next-generation phylodynamics

Whole Genome Sequencing for Clinical Microbiology

Translation into routine applications

02.09.2017



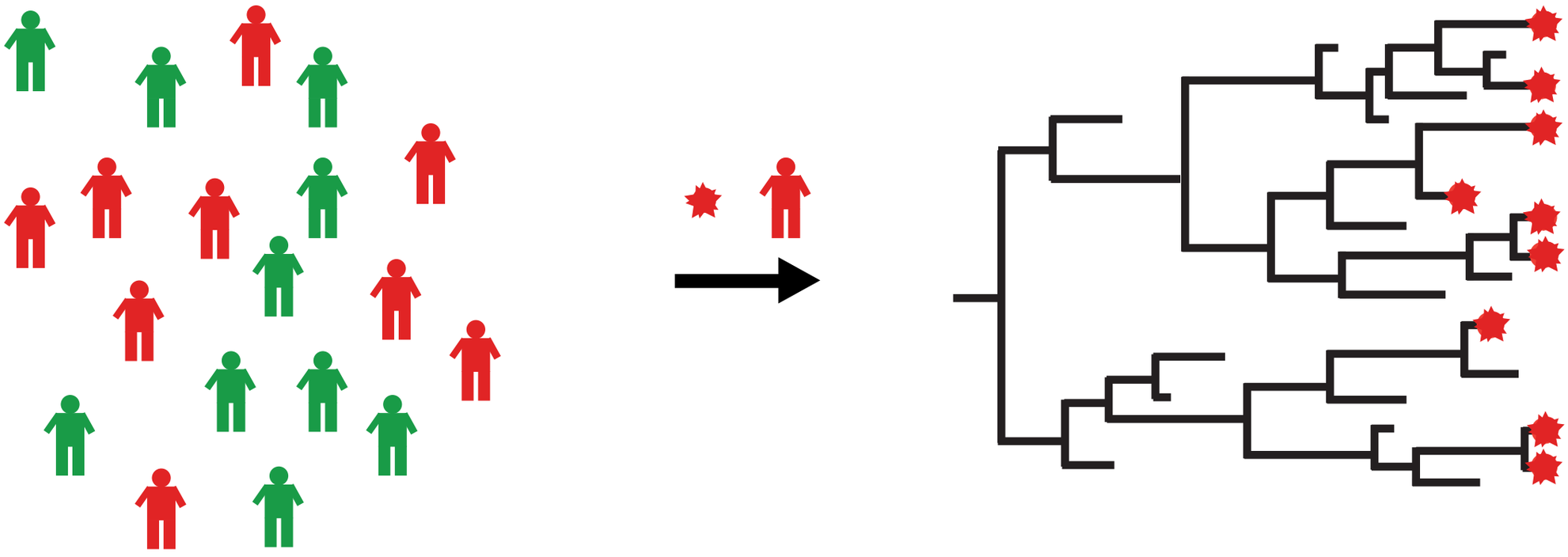
Veronika Boskova
veronika.boskova@bsse.ethz.ch
Computational Evolution group — Department of Biosystems Science and Engineering

ETH zürich

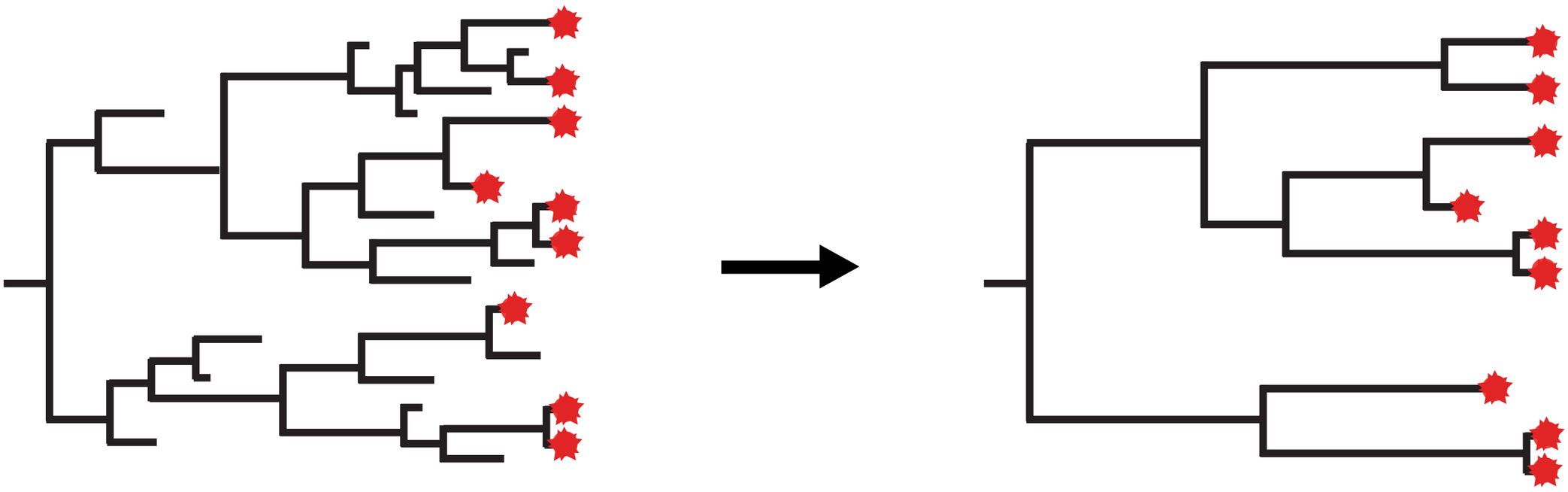
Sequences can be used for …

- association of variants with phenotype
  - humans: disease status
  - pathogens: drug resistance
- BUT also tracking pathogen spread

Pathogen phylogenetics



Pathogen phylogenetics

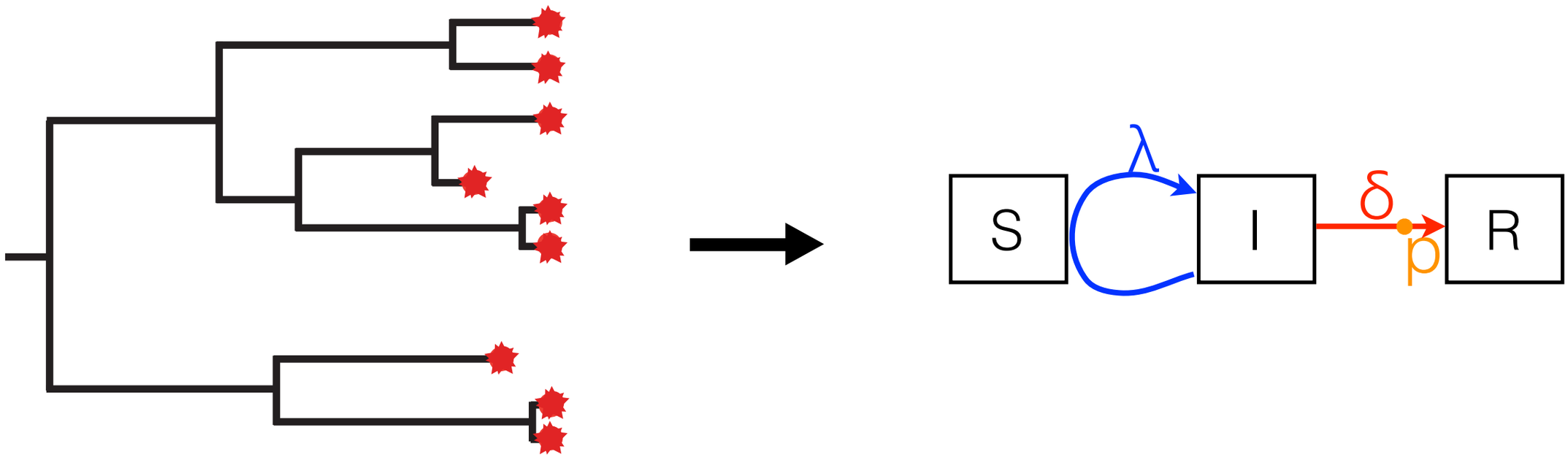


How does the drug resistance spread in the population?

How does the pathogen evolve through time?

When did the epidemic start?

Pathogen phylodynamics



How fast does the epidemic spread?

Where did it originate?

How many individuals in total have been infected?

Phylogenetic and phylodynamic inference

Data

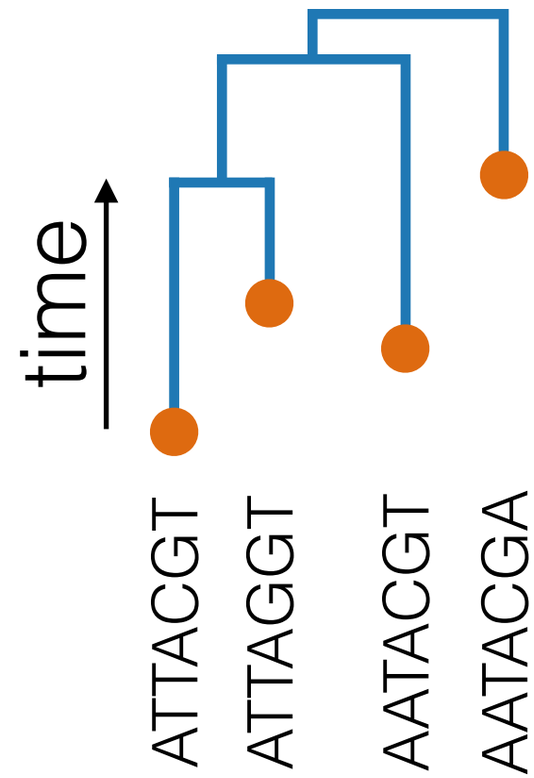
Alignment of sequences of the pathogen from different infected individuals
D

Sequence evolution model

Evolution of sequences along the tree
θ parameters of substitution model
can calculate the likelihood P(D | θ,τ)

Population dynamics model

Evolution of population
η parameters of population dynamics
can calculate the tree prior P(τ | η)



Slide courtesy of T. Stadler (modified)

Bayesian phylodynamics: population dynamics models

Birth-death model

Coalescent model

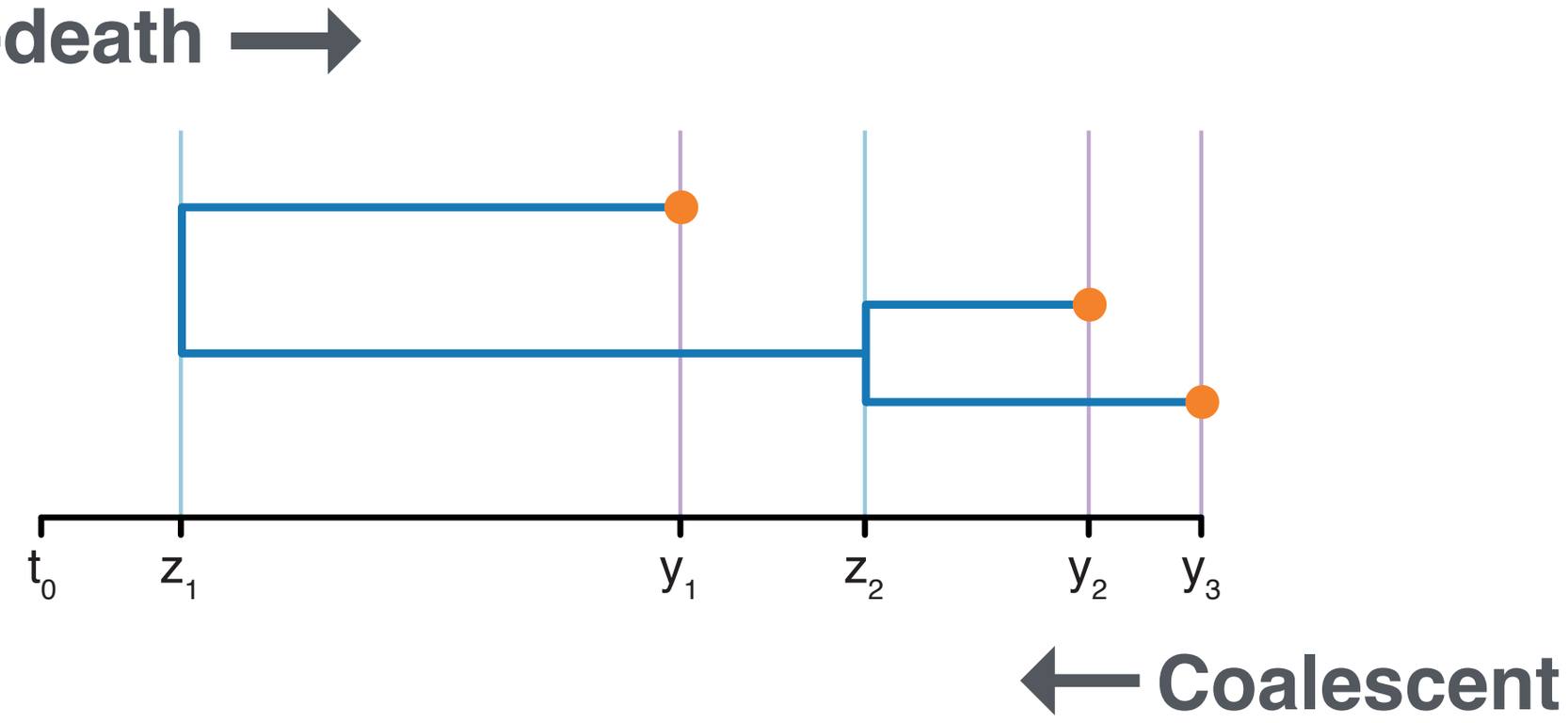
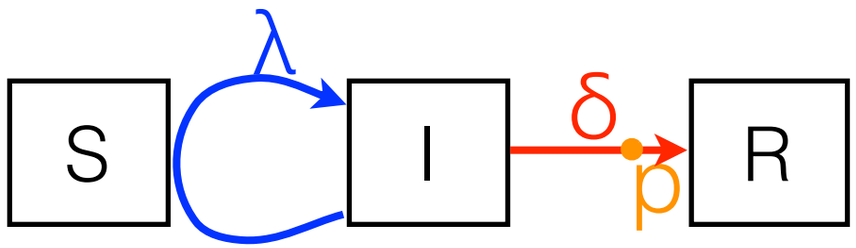


Figure adapted from du Plessis et al. *Trends in microbiology* 23.7 (2015): 383-386.

Bayesian phylodynamics: population dynamics models

Birth-death model



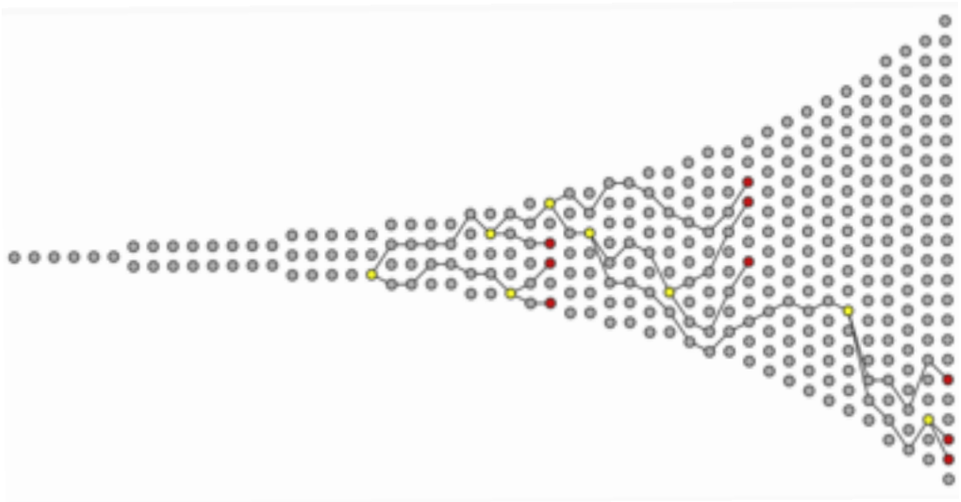
$R = \lambda/\delta$ — Reproductive number (is it spreading or not?)

δ — Becoming noninfectious rate ($1/\delta$ = time from infection to recovery)

p — Sampling probability

$R \in [0,\infty)$ $\delta \in [0,\infty)$ $p \in [0,1]$

Coalescent model



I_e — Population size of infected individuals

$\in [0,\infty)$

Bottom figure adapted from Kühnert et al. *Infection, Genetics and Evolution* 11.8 (2011): 1825-1841., Slide courtesy of T. Stadler (modified)

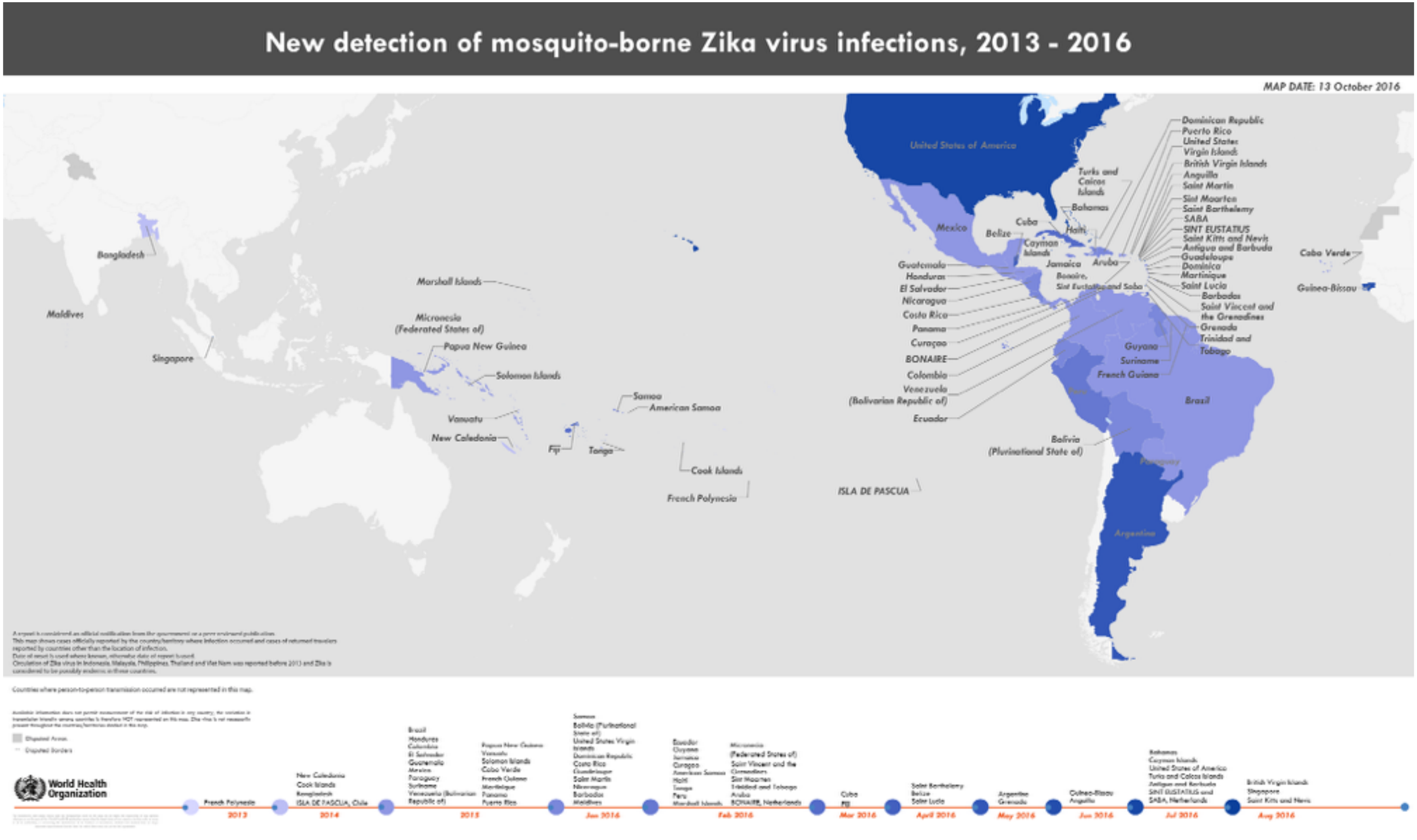
NGS in action

- Recent epidemic outbreaks
    - Ebola in West Africa
        - spreading in 2014
        - mortality >50%
        - portable sequencer MinION[1]

- **Zika in the Americas**
    - detected in 2015 in the Americas
    - MinION for ZiBRA in Brazil[2]

[1] Quick et al. Nature 530.7589 (2016): 228., [2] Faria et al. Genome medicine 8.1 (2016): 97.

Zika virus epidemic: spread



Zika virus epidemic: background

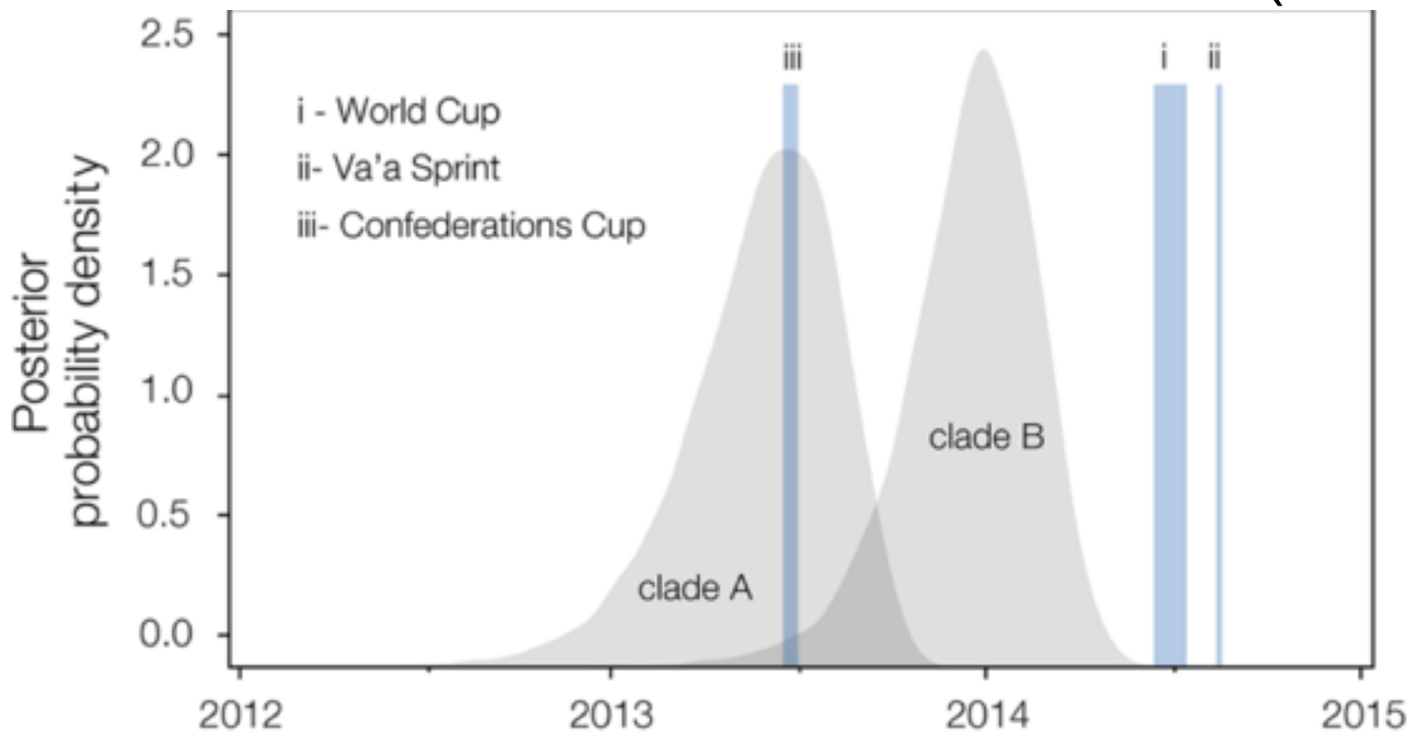
- 2015/2016 ZIKV infection monitoring and sequencing
- 07.2015 association with Guillain-Barré syndrome
- 10.2015 association with microcephaly
- 02.2016 ZIKV declared to be public health emergency of international concern[3]
- 11.2016 end of ZIKV public health emergency[3]
  but ZIKV remains “a highly significant and a long-term problem”

[3] WHO (Geneva, 2016): http://www.who.int

Zika virus epidemic: what is known

Faria et al, Science, March 2016

- tMRCA of American strains 12.2013 (08.2013-04.2014)



- substitution rate 0.98-1.06 $x10^{-3}$ subst/site per year

Zika virus epidemic: data

- Retrieved on 03.11.2016
  139 sequences from GenBank and github pages:

  https://github.com/zibraproject
  https://github.com/andersen-lab
  https://github.com/jtladner/ZIKA_Florida

- tMRCA and clock rate estimates for the Americas
- epidemiology for Florida
- use BEAST2 software

[4] Boskova et al. unpublished (2017)

Zika virus epidemic: conclusions

- In sub-epidemics
  - data not informative on the clock rate
  - cannot disentangle clock rate x tree height
- Florida outbreak
  - Re>1 in mid-2016
  - hard to reconcile with count data

[4] Boskova et al. unpublished (2017)

NGS in action

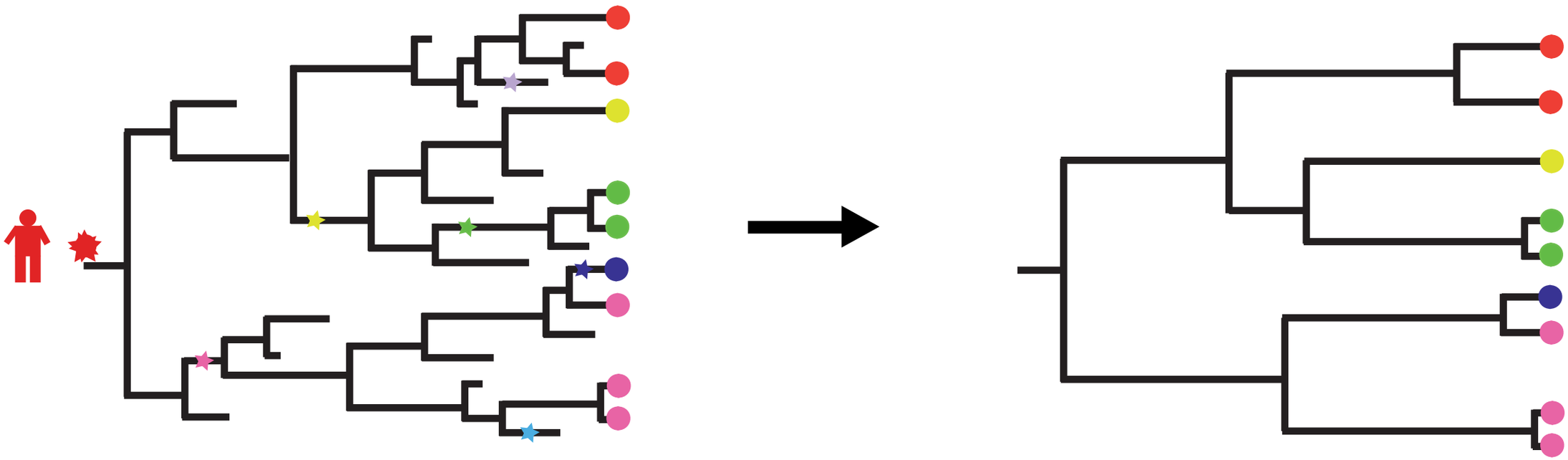
- Recent epidemic outbreaks
    - Ebola in West Africa
        - spreading in 2014
        - mortality >50%
        - portable sequencer MinION[1]

    - Zika in the Americas
        - detected in 2015 in the Americas
        - MinION for ZiBRA in Brazil[2]

- **Within-host pathogen dynamics**

[1] Quick et al. Nature 530.7589 (2016): 228., [2] Faria et al. Genome medicine 8.1 (2016): 97.

Within-host pathogen dynamics



NGS quasispecies datasets

- RNA viruses show high mutation rate
  = high population diversity (time-dependent)

- 10-20 haplotypes = 90% of the population
  = lots of duplicate sequences[7]

- If sequenced by NGS => LOTS of DATA
  - impossible to analyse full dataset with Bayesian phylogenetic/phylodynamic methods
  - unique sequences only
  - random subset of sequences

[7] Töpfer et al. *Journal of Computational Biology* 20.2 (2013): 113-123.

NGS quasispecies datasets - summary

- Analysis of unique datasets - biased
- Analysis of random datasets - biased/imprecise
- New method: PIQMEE
  - accurate and precise
  - fast: handles large datasets well unlike CLASSIC method

[8] Boskova et al. in preparation (2017)

Summary

- Phylodynamics is a powerful tool for (NGS) sequence analysis of early outbreak data
  - Need to be able to determine in which cases phylodynamic analyses are not (yet) possible
- Need (better) tree priors suited for large within-host NGS datasets
  - e.g. PIQMEE model for HIV/HCV

Acknowledgements

Zika project: Tanja Stadler
Carsten Magnus

PIQMEE project: Tanja Stadler
Joëlle Barido-Sottani
Denise Kühnert

D-BSSE
Department of Biosystems
Science and Engineering

ETH zürich

SIB
Swiss Institute of
Bioinformatics