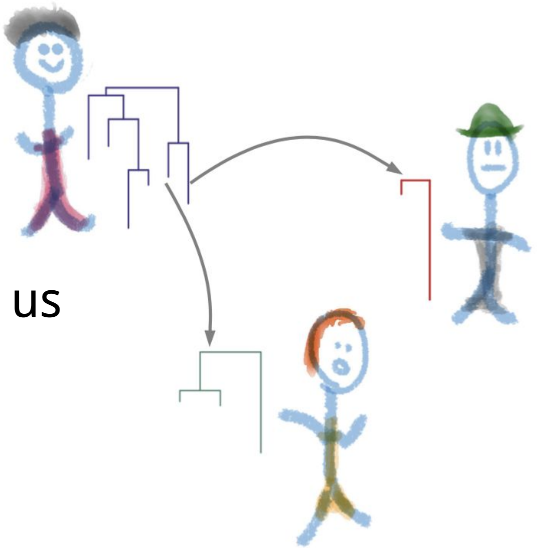
Research Forefronts

Simultaneous phylogenetic and transmission tree reconstruction with BREATH:

Bayesian Reconstruction and Evolutionary Analysis of Transmission Histories

BREATH: A new method that meets a key remaining challenge

- **Diverse infections**: Evolution happens within hosts.
- Multiple samples per host
- **Incomplete sampling**
- **Phylogenetic uncertainty**
- **Limited pathogen variation:** does not fully inform us about a phylogenetic tree
- Bottleneck at transmission
- Environmental pathogens
- Additional data eg timing, contact information, location



The main barrier to simultaneously inferring the phylogeny and transmission tree was the unsampled cases.

Want to allow (unknown numbers of) unsampled cases, while remaining feasible

- Generically, unsampled cases typically come with parameters.
- Changing the number of them changes the dimension during the MCMC, requiring reversible jump MCMC (rjMCMC).
- TransPhylo requires an input phylogenetic tree and rjMCMC.

Here: Create a likelihood for the transmission tree given the epidemiological model, which has unsampled cases without adding new parameters for them, while still allowing within-host diversity.

Main idea: allow chains of unsampled infections on edges

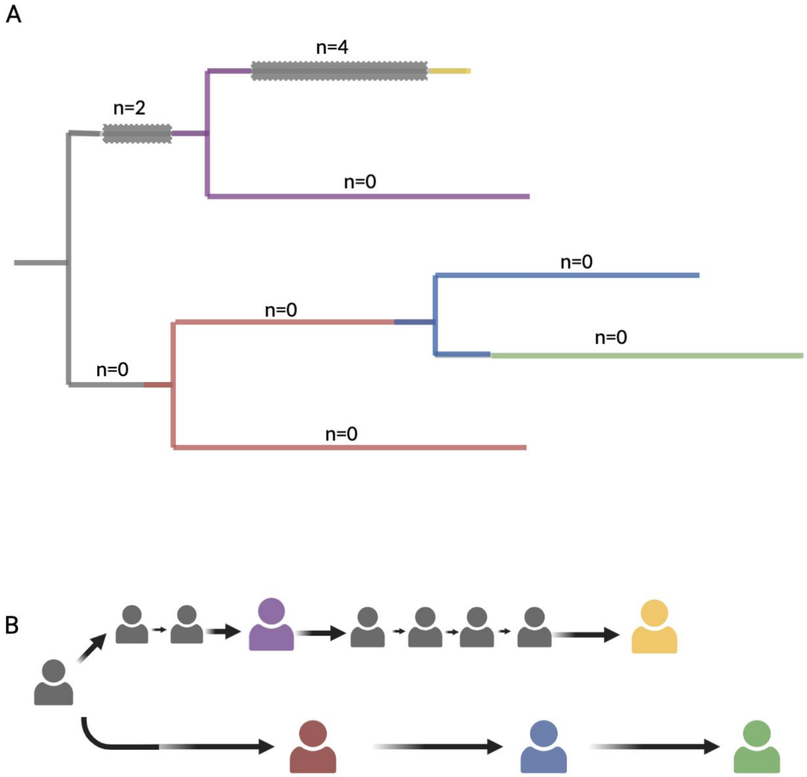
- Each edge of the tree has an integer *n* associated with it.

- *n* represents the number of additional hosts on the edge

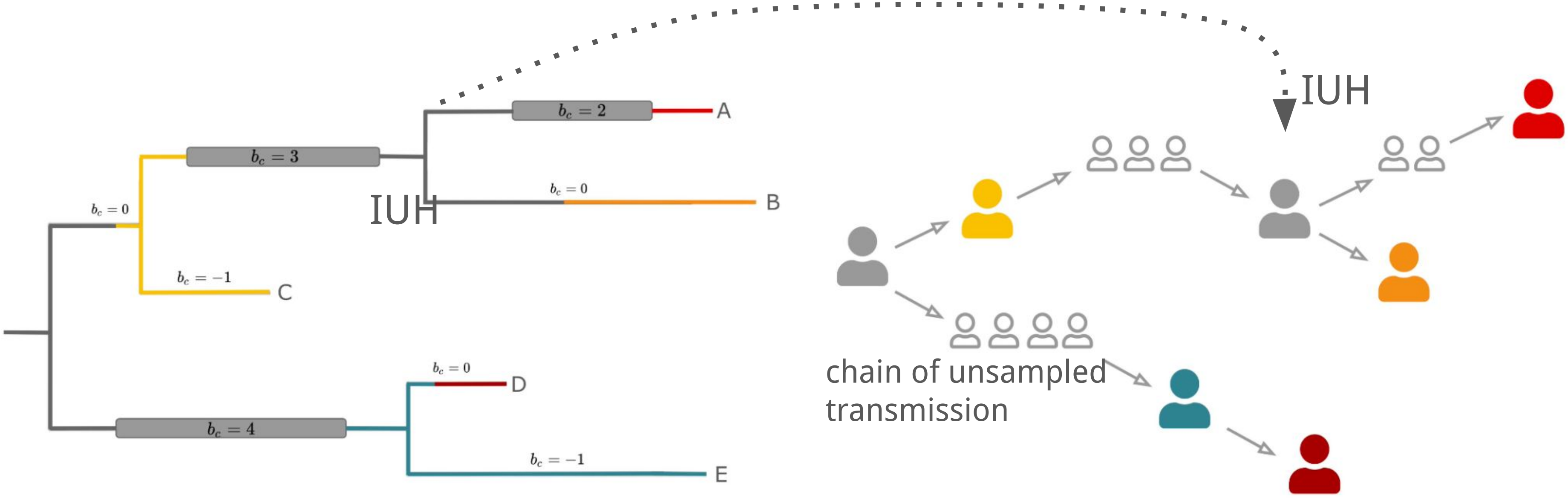
- There are 3 kinds of people in the tree: sampled individuals, unsampled individuals, and people in chains of unsampled transmission

- Nodes of the phylogeny have to be in individuals.

An annotated phylogeny (A), and the associated transmission tree (B)



Example showing an internal unsampled host (IUH)



IUH

IUH

chain of unsampled transmission

A. Phylogenetic tree with colouring and block counts

B. Corresponding transmission tree

BREATH’s colouring rules

- Each part of the phylogeny has a colour
- Sampled hosts each have a colour. There is one ‘unsampled’ colour (U).
- Each sampled host’s tip in the phylogeny, and some of its edge, has its host’s colour
- Except for the unsampled colour, colours must be connected (like in TransPhylo)
- Colour changes denote transmission events (as in TransPhylo)

Bayesian decomposition

Built from the usual: P(A|B)P(B) = P(B|A)P(A)

$$
\begin{aligned}
P(G, T, \theta, N_e g, w | D, \tau_s) \\
&\propto P(D|G, T, N_e g, \theta, w) P(G|T, \theta, N_e g, w, \tau_s) P(T|\theta, \tau_s) P(N_e g) P(\theta) P(w) \\
&= P(D|G, w) P(G|T, \theta, N_e g, \tau_s) P(T|\theta, \tau_s) P(\theta) P(w) P(N_e g)
\end{aligned}
$$

G: genealogy

T: transmission tree

D: data

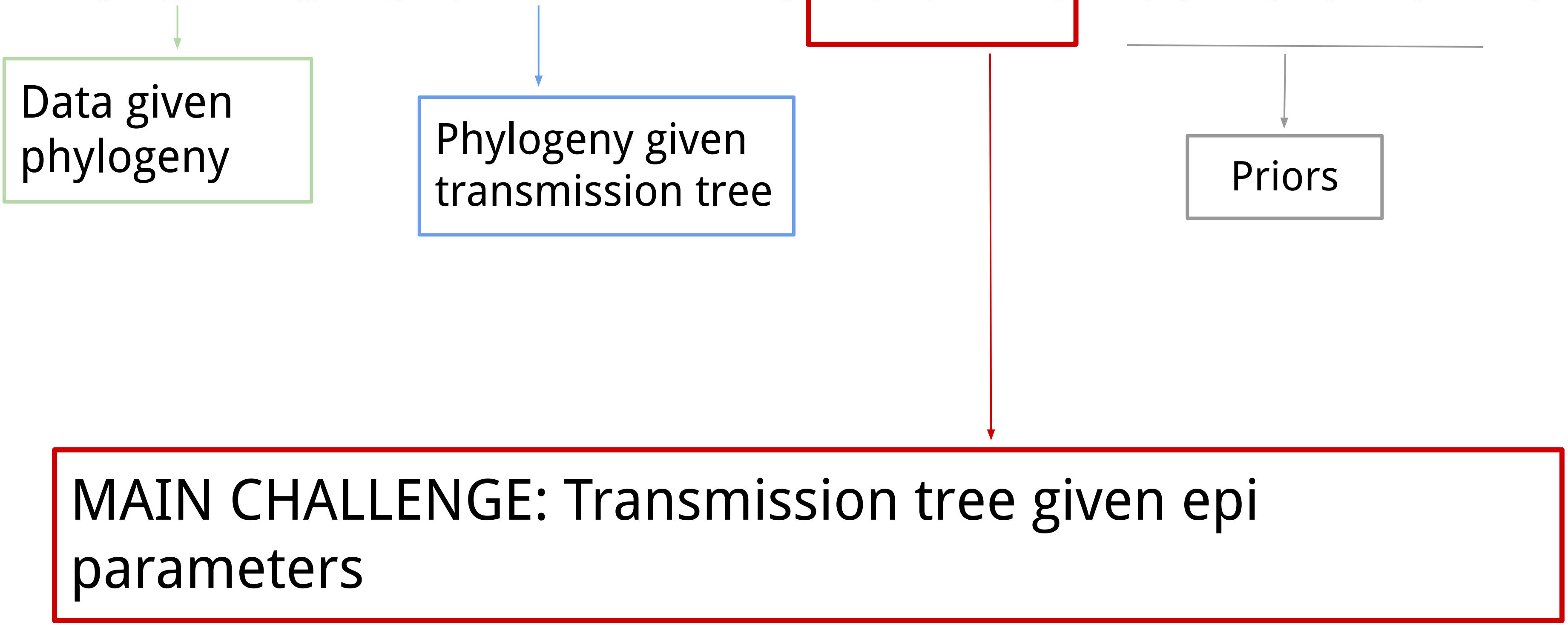
$N_e g$: within-host coalescent parameter

$\boldsymbol{\theta}$: epidemiological parameters

w: molecular clock model

$\boldsymbol{\tau}_s$: sampling times

$$P(D|G,w)P(G|T,\theta,N_eg,\tau_s)\boxed{P(T|\theta,\tau_s)}P(\theta)P(w)P(N_eg)$$



Data given phylogeny

Phylogeny given transmission tree

Priors

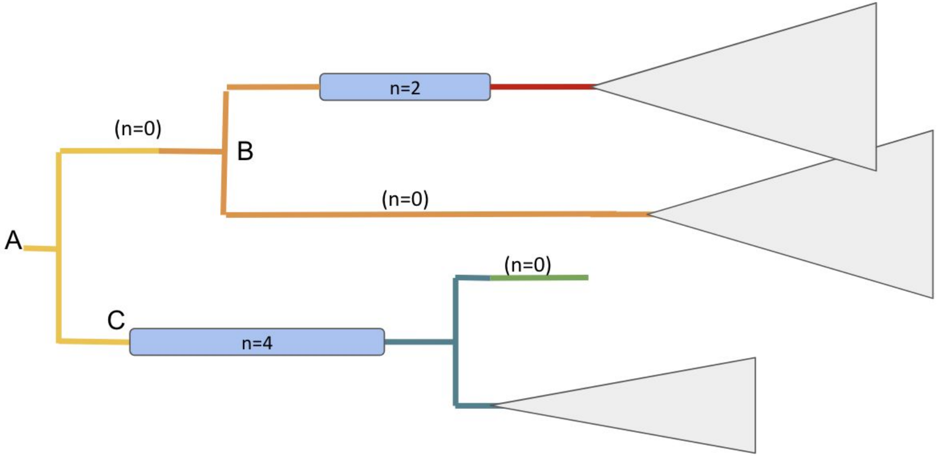
MAIN CHALLENGE: Transmission tree given epi parameters

Transmission tree likelihood

B’s likelihood depends on B’s time infection, when A infected B.

T_x is the tree descending from x.

Let c(x) be the “children” of x (the onward descendants).



$$L(A,B,C,\dots) = L(A)\ L(T_B|A)L(T_C|A) = L(A)\ \ L(B|A)L(T_{c(B)})\ \ L(C|A)L(T_{c(C)})$$

Get a product: one term for each case and for each block

The dimension of the augmented object does not change

The phylogeny has *n* tips. It is binary, and always has *2n-1* branches.

Each branch in the phylogeny has 3 additional parameters:

- the block count b_c : number of additional unsampled individuals.
  - $b_c > 0$: there is a chain of length b_c . In this case $b_c = n$.
  - $b_c = 0$: there are no unsampled people in a chain on this edge, but there is a colour change (so there is a transmission event)
  - $b_c = -1$: there is no colour change and no transmission (same host)
- the start time of the chain of unsampled individuals
- the end time of the chain of unsampled individuals.

When $b_c = -1$, the start and end times are equal and can be set to the midpoint of the branch (likelihood does not depend on the point).

Recurrent events model with intensity functions

We model two kinds of events: sampling, and transmission to another individual.

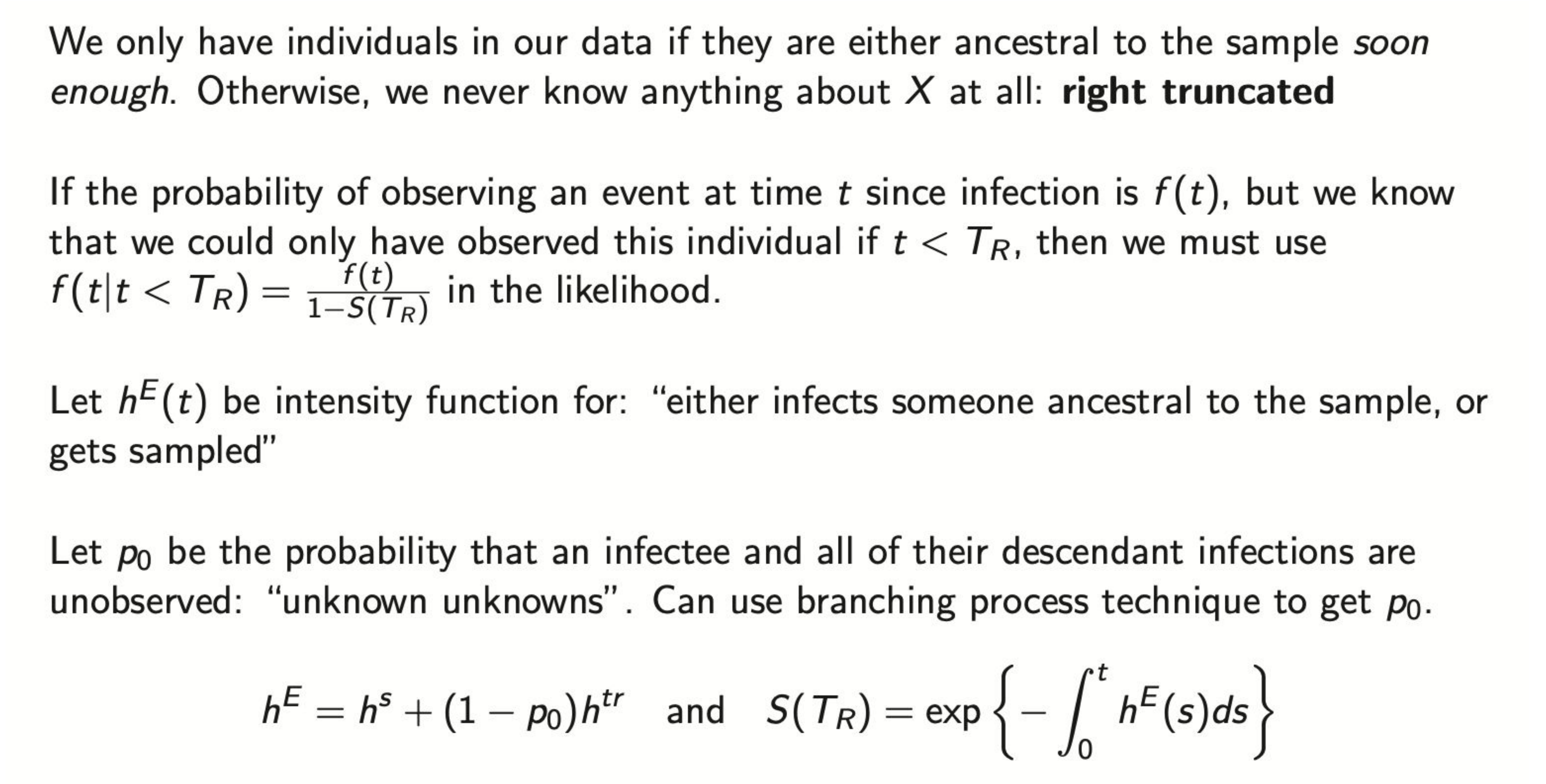
Each has an intensity function, h^s and h^{tr} respectively, with

$$h(t) = \lim_{dt \to 0} \frac{P(\text{event occurs in time } [t, t + dt])}{dt}.$$

The probability density for k events at times $t_1 < t_2 < ... < t_k$ with intensity $h(t)$, in an interval from 0 to t, is the probability that it does not happen except at times t_j

$$\exp\left\{-\int_0^t h(s)ds\right\} \prod_{j=1}^{k} h(t_j)$$

Right truncation



We only have individuals in our data if they are either ancestral to the sample *soon enough*. Otherwise, we never know anything about X at all: **right truncated**

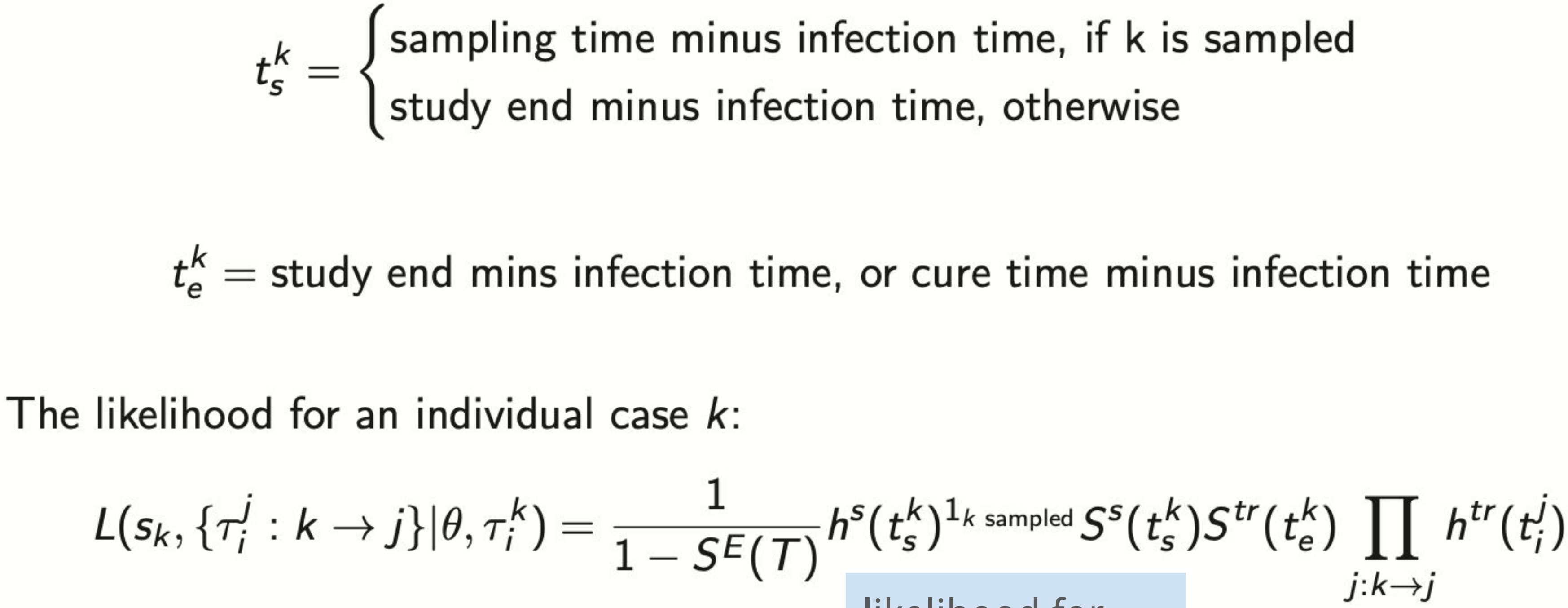
If the probability of observing an event at time t since infection is $f(t)$, but we know that we could only have observed this individual if $t < T_R$, then we must use $f(t|t < T_R) = \frac{f(t)}{1-S(T_R)}$ in the likelihood.

Let $h^E(t)$ be intensity function for: "either infects someone ancestral to the sample, or gets sampled"

Let p_0 be the probability that an infectee and all of their descendant infections are unobserved: "unknown unknowns". Can use branching process technique to get p_0.

$$h^E = h^s + (1 - p_0)h^{tr} \quad \text{and} \quad S(T_R) = \exp\left\{-\int_0^t h^E(s)ds\right\}$$

Individual cases' likelihood



$$t_s^k = \begin{cases} \text{sampling time minus infection time, if k is sampled} \\ \text{study end minus infection time, otherwise} \end{cases}$$

$$t_e^k = \text{study end mins infection time, or cure time minus infection time}$$

The likelihood for an individual case k:

$$L(s_k, \{\tau_i^j : k \to j\} | \theta, \tau_i^k) = \frac{1}{1 - S^E(T)} h^s(t_s^k)^{1_{k \text{ sampled}}} S^s(t_s^k) S^{tr}(t_e^k) \prod_{j:k \to j} h^{tr}(t_i^j)$$

right truncation

likelihood for sampling

likelihood for transmission

Likelihood for the chains of transmission

Similar idea with intensity functions. Chains of unsampled transmission end when either:

- a case gets sampled
- a case infects someone who is “multiply ancestral to the sample” (MATTS)

MATTS: at least two of the lineages in the individual have sampled descendants.

Use success probability, geometric distribution, and right truncation again

Likelihood for unsampled chains of transmission

A chain of unsampled transmission proceeds from case to case until it ends.

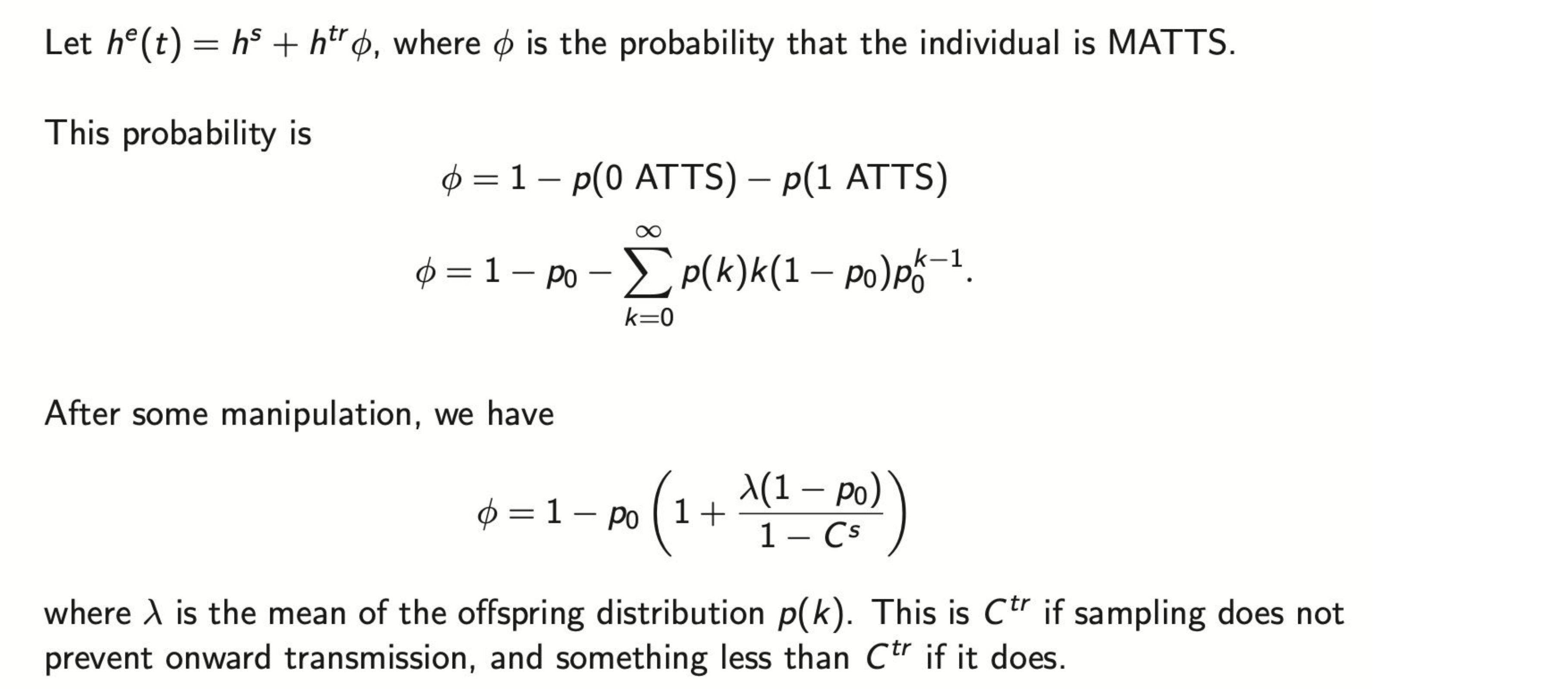
It ends when someone is either sampled, or infects someone who is multiply ancestral to the sample.

Each infectee has a probability of ending the chain (for now, ignore the finite time - we handle that by adjusting for right truncation later).

That means the distribution for the number of cases is geometric.

But we need the “success probability” for the geometric distribution.

Hazard for ending an unsampled transmission chain



Let $h^e(t) = h^s + h^{tr}\phi$, where ϕ is the probability that the individual is MATTS.

This probability is

$$\phi = 1 - p(0 \text{ ATTS}) - p(1 \text{ ATTS})$$

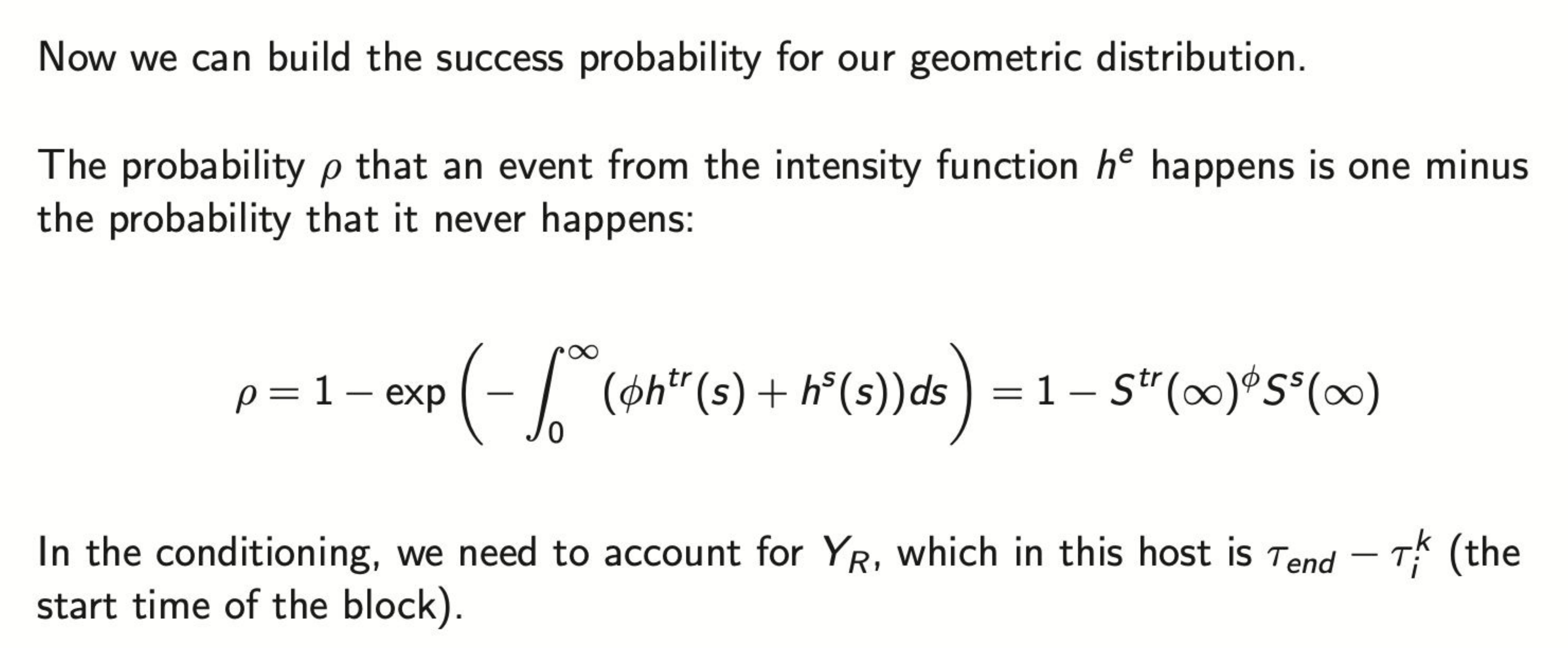
$$\phi = 1 - p_0 - \sum_{k=0}^{\infty} p(k)k(1-p_0)p_0^{k-1}.$$

After some manipulation, we have

$$\phi = 1 - p_0\left(1 + \frac{\lambda(1-p_0)}{1-C^s}\right)$$

where λ is the mean of the offspring distribution $p(k)$. This is C^{tr} if sampling does not prevent onward transmission, and something less than C^{tr} if it does.

Geometric distribution for the unsampled chains



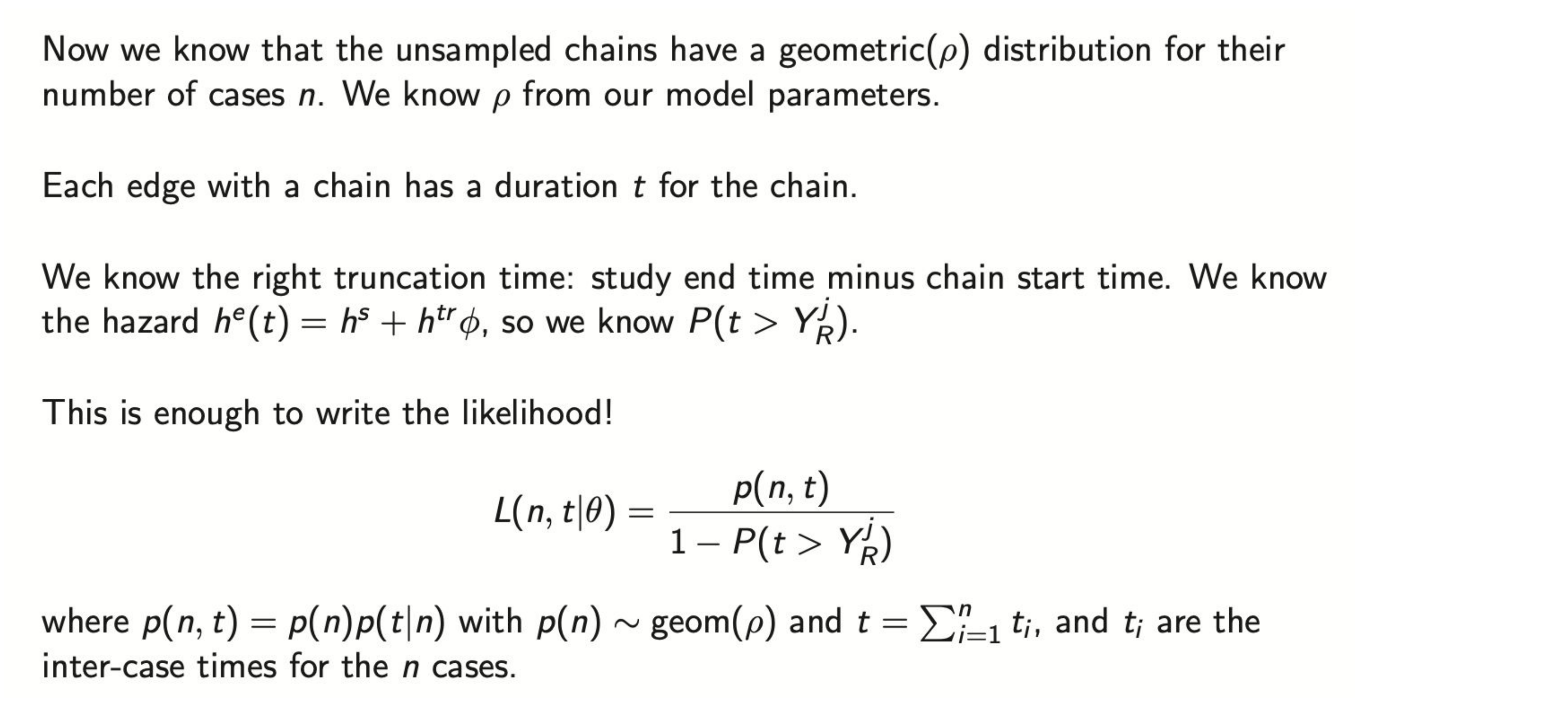
Now we can build the success probability for our geometric distribution.

The probability ρ that an event from the intensity function h^e happens is one minus the probability that it never happens:

$$\rho = 1 - \exp\left(- \int_0^\infty (\phi h^{tr}(s) + h^s(s)) ds \right) = 1 - S^{tr}(\infty)^\phi S^s(\infty)$$

In the conditioning, we need to account for Y_R, which in this host is $\tau_{end} - \tau_i^k$ (the start time of the block).

Unsampled chain likelihood



Now we know that the unsampled chains have a geometric(ρ) distribution for their number of cases n. We know ρ from our model parameters.

Each edge with a chain has a duration t for the chain.

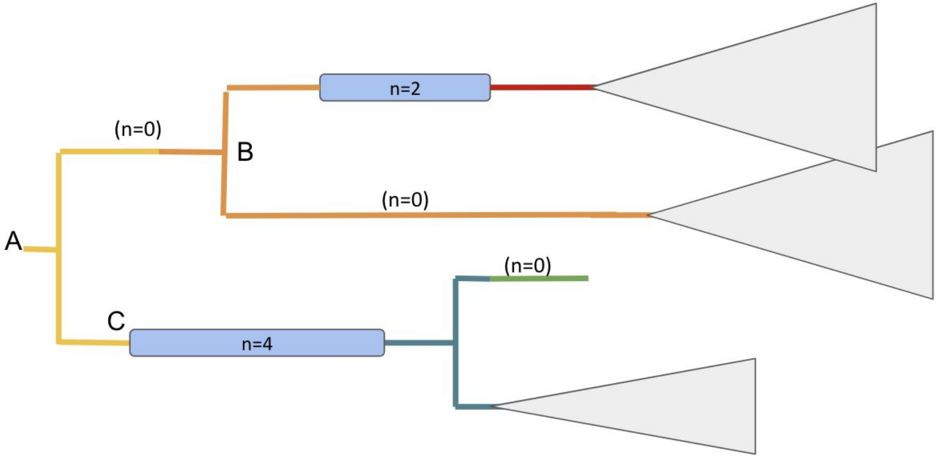
We know the right truncation time: study end time minus chain start time. We know the hazard $h^e(t) = h^s + h^{tr}\phi$, so we know $P(t > Y_R^j)$.

This is enough to write the likelihood!

$$L(n, t|\theta) = \frac{p(n,t)}{1 - P(t > Y_R^j)}$$

where $p(n,t) = p(n)p(t|n)$ with $p(n) \sim \text{geom}(\rho)$ and $t = \sum_{i=1}^{n} t_i$, and t_i are the inter-case times for the n cases.

Transmission tree likelihood is now complete



$$L(A,B,C,\ldots) = L(A)\ L(T_B|A)L(T_C|A) = L(A)\ \ L(B|A)L(T_{c(B)})\ \ L(C|A)L(T_{c(C)})$$

Now we have the ingredients: the individual cases (with their times of sampling and infecting others) , and the chains of transmission.

Implementation

BEAST2 implementation by Remco Bouckaert: the *transmission* package

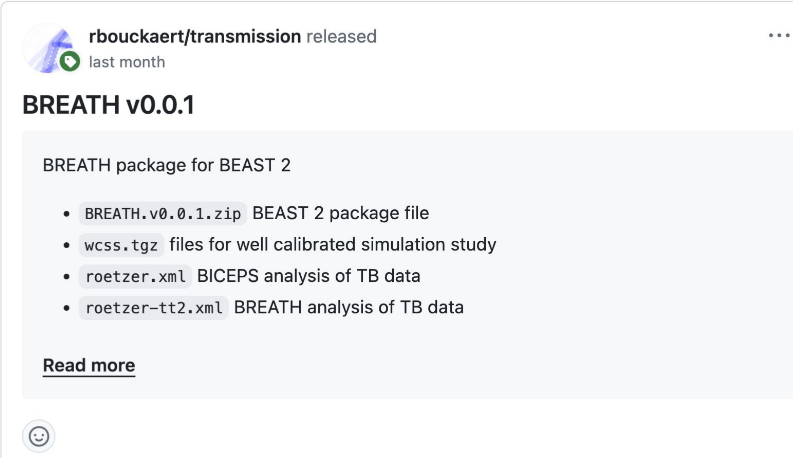
2 new moves:

Infection mover:

- pick an infection on the path between two tips
- move it elsewhere

Block operator

- 50% probability: move block boundaries
- 50% probability: remove or add infections



rbouckaert/transmission released
last month

BREATH v0.0.1

BREATH package for BEAST 2

- `BREATH.v0.0.1.zip` BEAST 2 package file
- `wcss.tgz` files for well calibrated simulation study
- `roetzer.xml` BICEPS analysis of TB data
- `roetzer-tt2.xml` BREATH analysis of TB data

Simulation model and data

Simulation model:

We simulate transmission and sampling with intensities following the model.

We simulate within-host phylogenies with a constant-rate coalescent.

This allows us to check model performance.

Data:

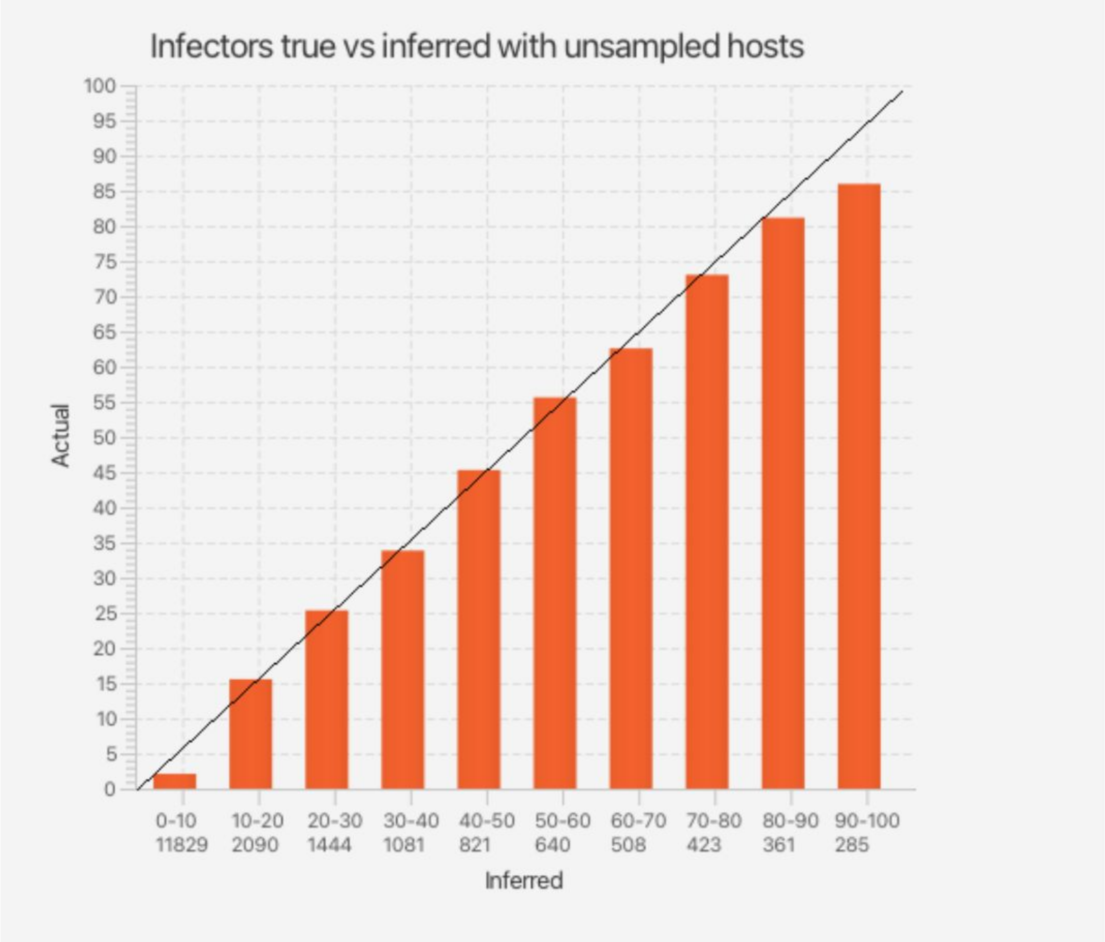
TB outbreak of 86 cases in Hamburg, Germany. Active case finding and passive surveillance. Genomic data (SNPs) and times are publicly available.

Previous method (TransPhylo) developed with these data. Roetzer et al, 2013.

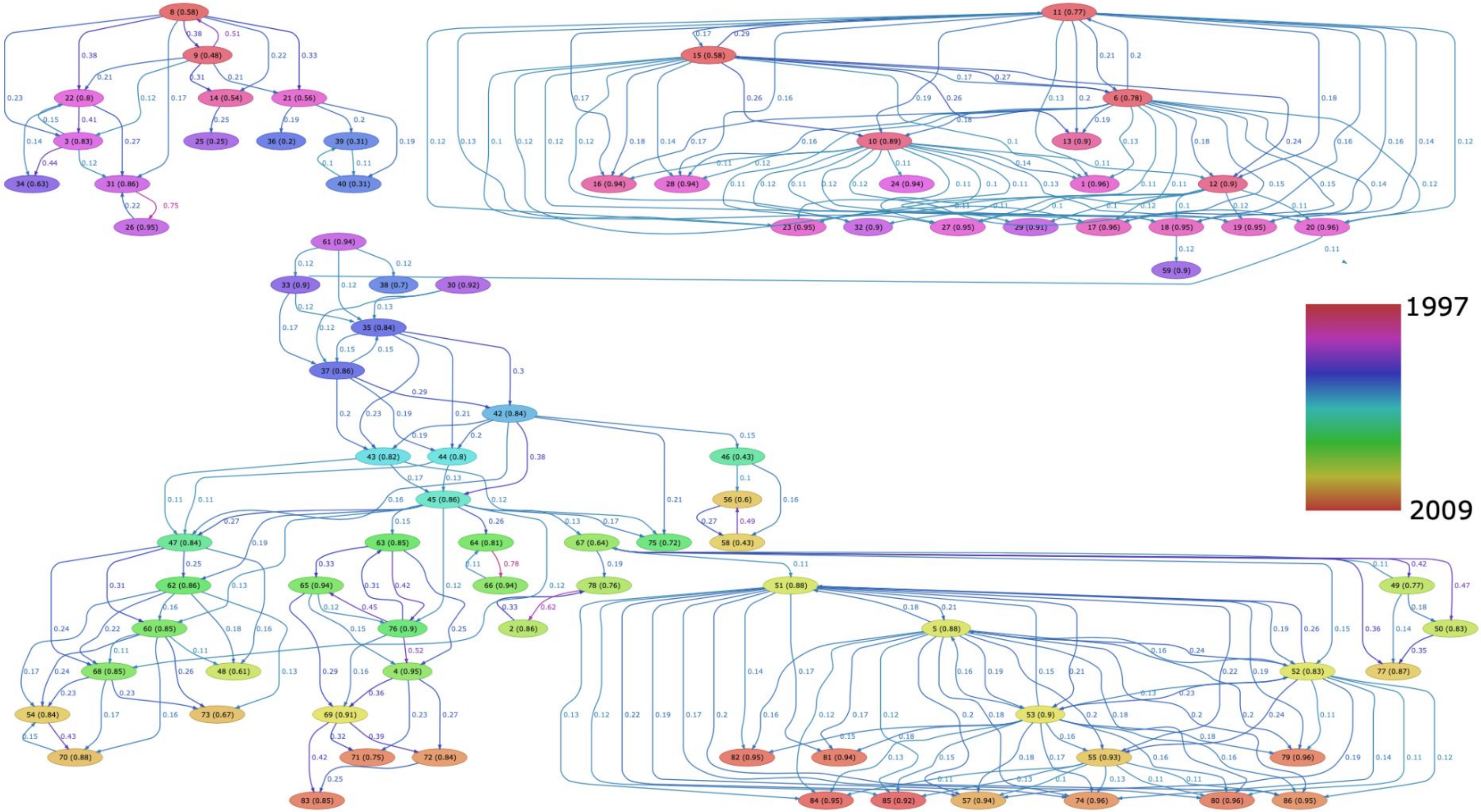
Simulation test results

What fraction of events with posterior probability x actually happened?

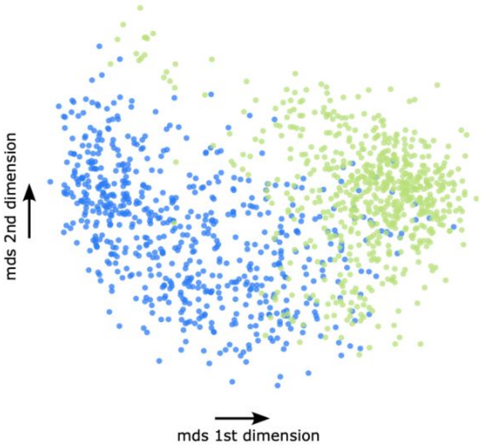
If the model is working, a fraction y = x of those events should have happened.



Outbreak analysis: transmission tree (who infected whom)

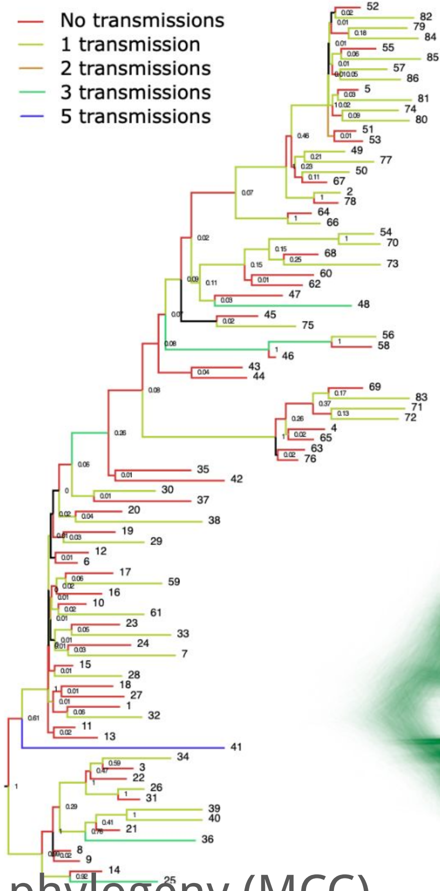


Outbreak analysis

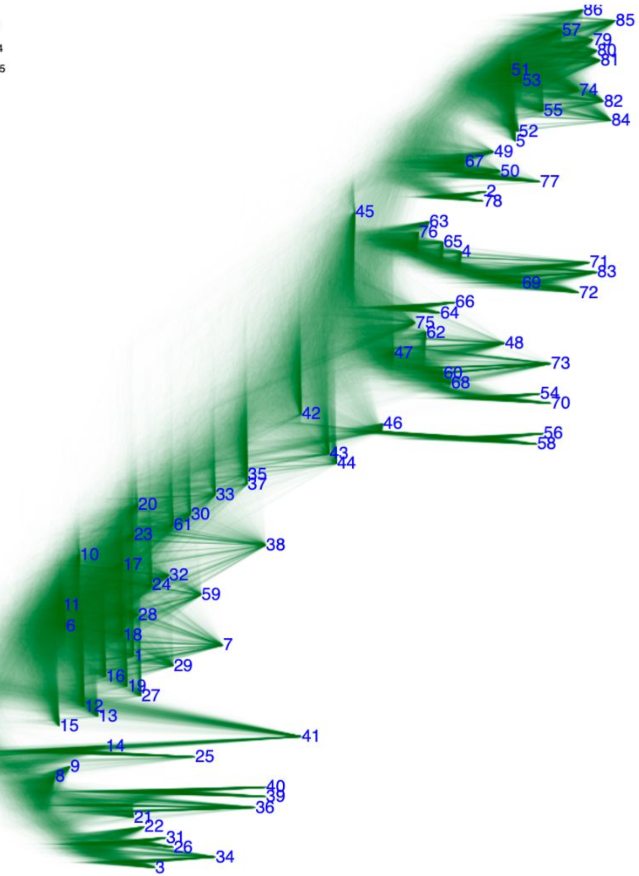


Phylogenies differ from regular coalescent models. Important to have correct process - here, *transmission*.

MDS plot

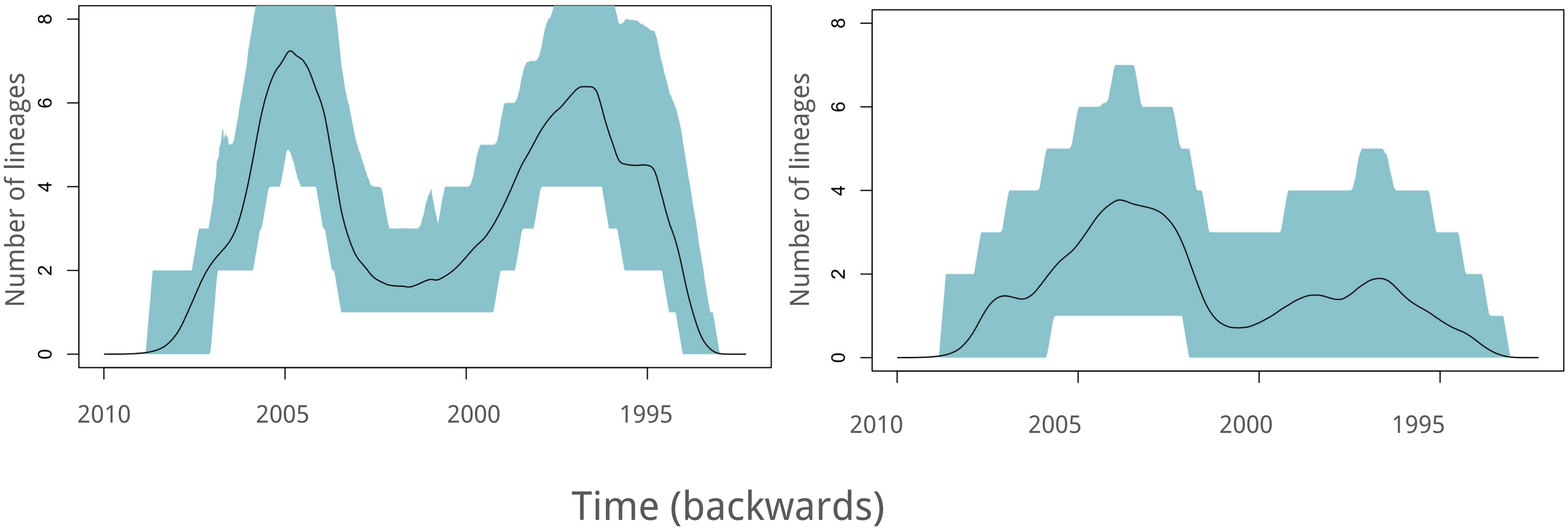


Timed phylogeny (MCC)



Densitree of posterior

Lineages through time: the outbreak nearly stopped ~7 years before it did stop



Total lineages through time

Unsampled lineages through time

BREATH’s Phylogenetic tree advantages

We compared BREATH to a coalescent model called BICEPS for the phylogeny.

BREATH has higher clade support and shorter branches -- more parsimonious.

Comparison:

- BICEPS estimates the time of origin as 1980-1994, total length ~220 years
- BREATH estimates the time of origin as 1993-1996, total length ~160 years

Interpretation:

- Person-to-person transmission is the process that created the sequence data. When we want to make a phylogeny, using this information *helps*

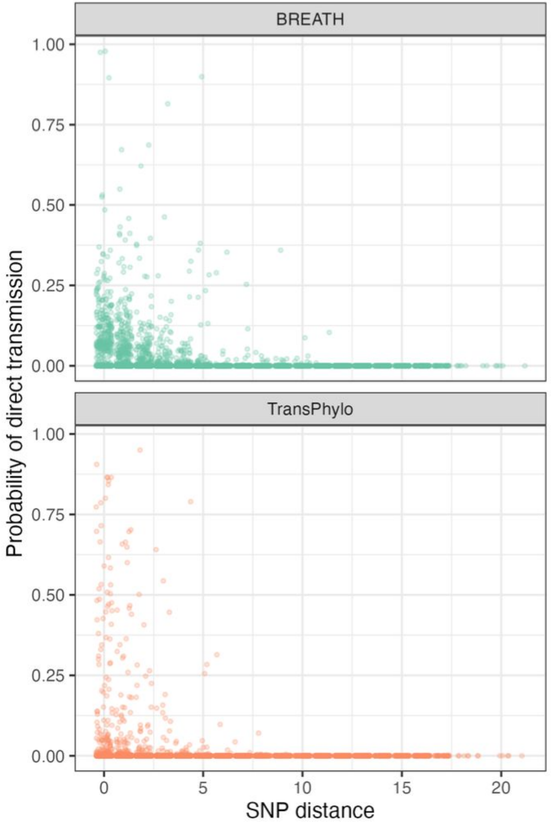
Transmission analysis compared to TransPhylo

BREATH overall has lower transmission probabilities for pairs of sampled cases than TransPhylo.

But BREATH has a more consistent pattern that nearer pairs are more likely to be transmission pairs.

Where BREATH's probabilities are high, so are TransPhylo's (but not the converse).

Interpretation: TransPhylo is sometimes overconfident because of the fixed input phylogeny



Limitations

There are a lot of parameters.

There are going to be some trade-offs. Right now we don’t *estimate* the epi parameters.

We will not be able to estimate the in-host coalescent model, the sampling fraction and the generation times.

Convergence: BREATH may take a lot of time changing the numbers in the blocks, which we don’t care too much about; this might be slow. Phylogeny construction is pretty slow. (But often we don’t have huge numbers of sequences in host-to-host outbreak settings)

Opportunities and next steps for this method

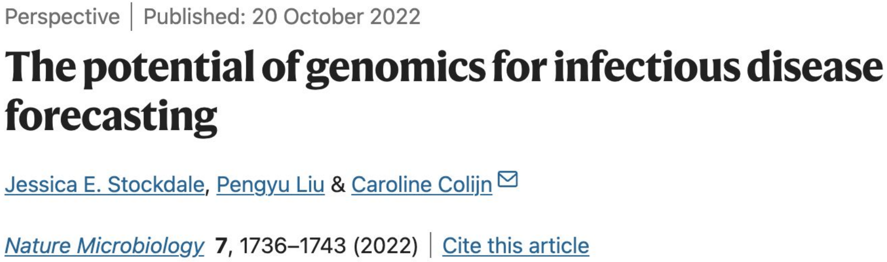
- Integrate contact data into the transmission tree prior
    - if two individuals are known not to have been in the same place at the same time, penalize transmission trees that have a direct transmission event
    - if they were in close contact, increase the likelihood of a direct transmission event
- Include the option of multiple samples per host -- either with just one infection each or multiple infections
- Estimate relative transmissibility by host and pathogen type: proportional hazards
- Extend to phylogeographic models and other applications

A bigger challenge

Gap: we have a rich *retrospective* picture from genomic data

We have very limited *prospective* (future-looking) models. Often these cannot even sustain diversity (e.g. simple SIR models with multiple strains). Yet we see diversity growing and being maintained.

Question: how to connect genomic surveillance to qualitative and quantitative models for microbial populations and their evolution?



Perspective | Published: 20 October 2022

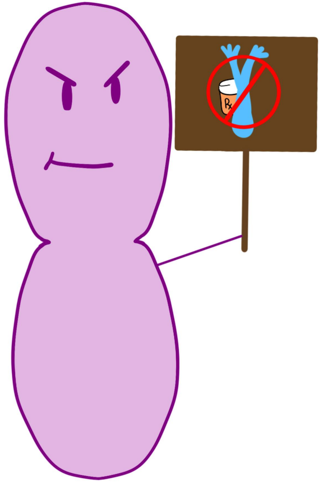
The potential of genomics for infectious disease forecasting

Jessica E. Stockdale, Pengyu Liu & Caroline Colijn

Nature Microbiology **7**, 1736–1743 (2022) | Cite this article

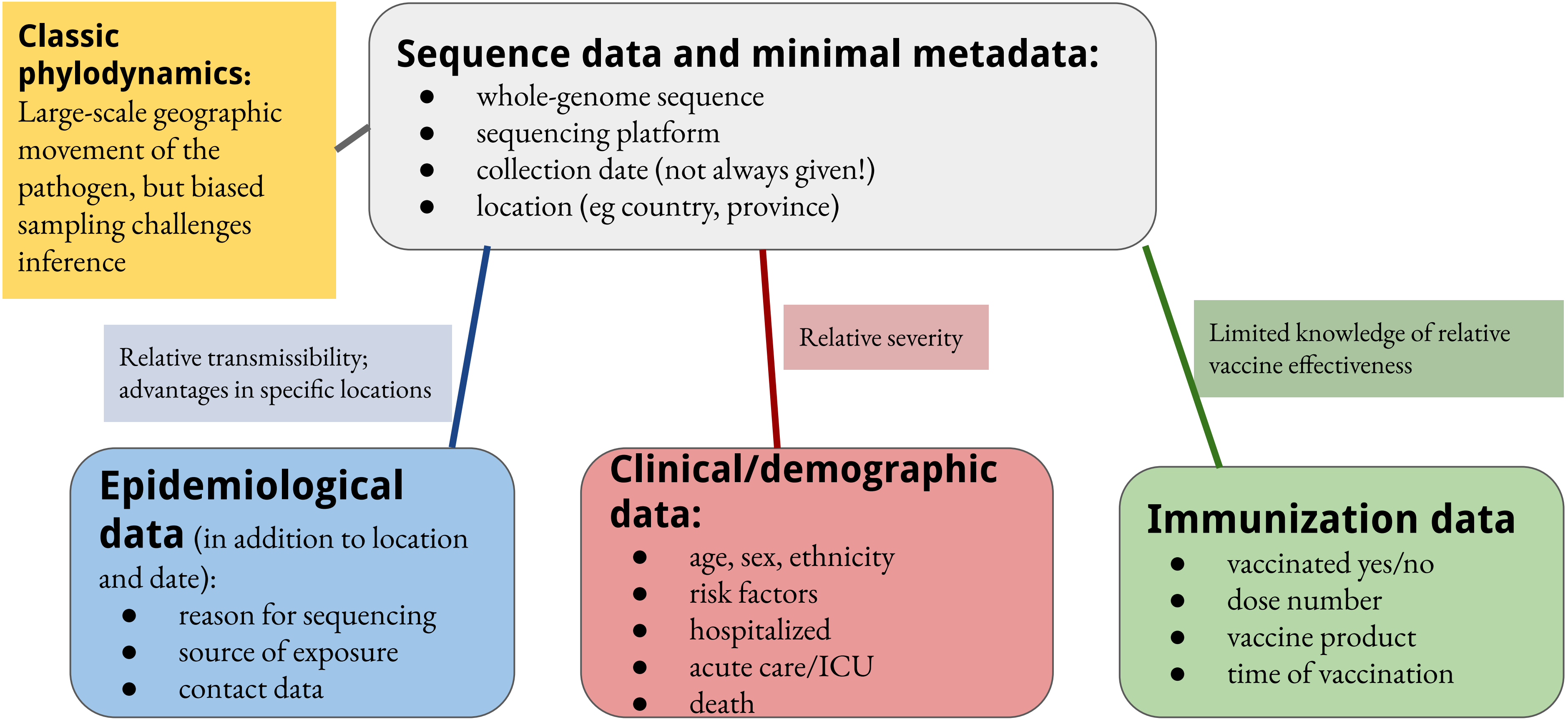
Genomic epidemiology offers many opportunities

- An opportunity for both public health and evolution/ecology: microbes as model organisms
- Public health: transmission at different scales → better interventions!
- Evolution of vaccine-resistance, AMR, and time scales of adaptation
- Evolution: generation and maintenance of diversity; selection; mechanisms of adaptation; co-evolution
- Public health - to - Evolution interface: emerging zoonosis like H5N1- host jumps, selection, phenotypic impact of mutations, recombinations (reassortment)



Penicillin resistance

Opportunities lost: the need for *linked* genomic surveillance



Classic phylodynamics: Large-scale geographic movement of the pathogen, but biased sampling challenges inference

Sequence data and minimal metadata:

- whole-genome sequence
- sequencing platform
- collection date (not always given!)
- location (eg country, province)

Relative transmissibility; advantages in specific locations

Relative severity

Limited knowledge of relative vaccine effectiveness

Epidemiological data (in addition to location and date):

- reason for sequencing
- source of exposure
- contact data

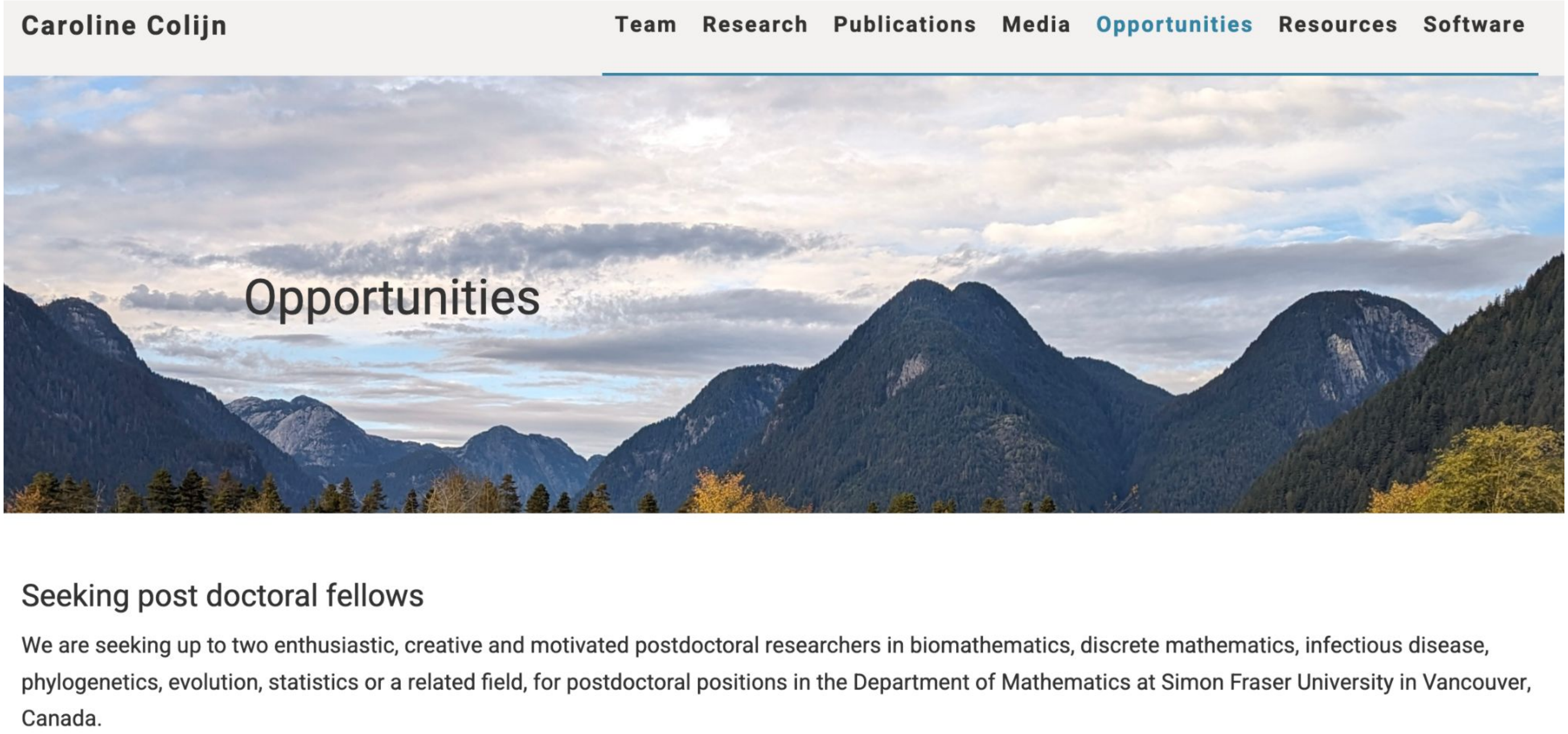
Clinical/demographic data:

- age, sex, ethnicity
- risk factors
- hospitalized
- acute care/ICU
- death

Immunization data

- vaccinated yes/no
- dose number
- vaccine product
- time of vaccination

Postdoc positions available! Email ccolijn@sfu.ca



Seeking post doctoral fellows

We are seeking up to two enthusiastic, creative and motivated postdoctoral researchers in biomathematics, discrete mathematics, infectious disease, phylogenetics, evolution, statistics or a related field, for postdoctoral positions in the Department of Mathematics at Simon Fraser University in Vancouver, Canada.

Acknowledgments

Remco Bouckaert, University of Auckland

Matthew Hall, Oxford Big Data Institute

Roetzer et al and Germany’s public health

All those generating and sharing pathogen sequence data

Motivating applications: tuberculosis transmission with sequence data, Collaboration: Ted Cohen, Yale University

Remco Bouckaert