Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models

Rasmussen and Stadler (*eLife, 2019*)

Seth D. Temple

STAT 591; APMA 990
Seattle, WA; Vancouver, BC
March 29, 2023

UNIVERSITY of WASHINGTON

Setting the scene: MTBD

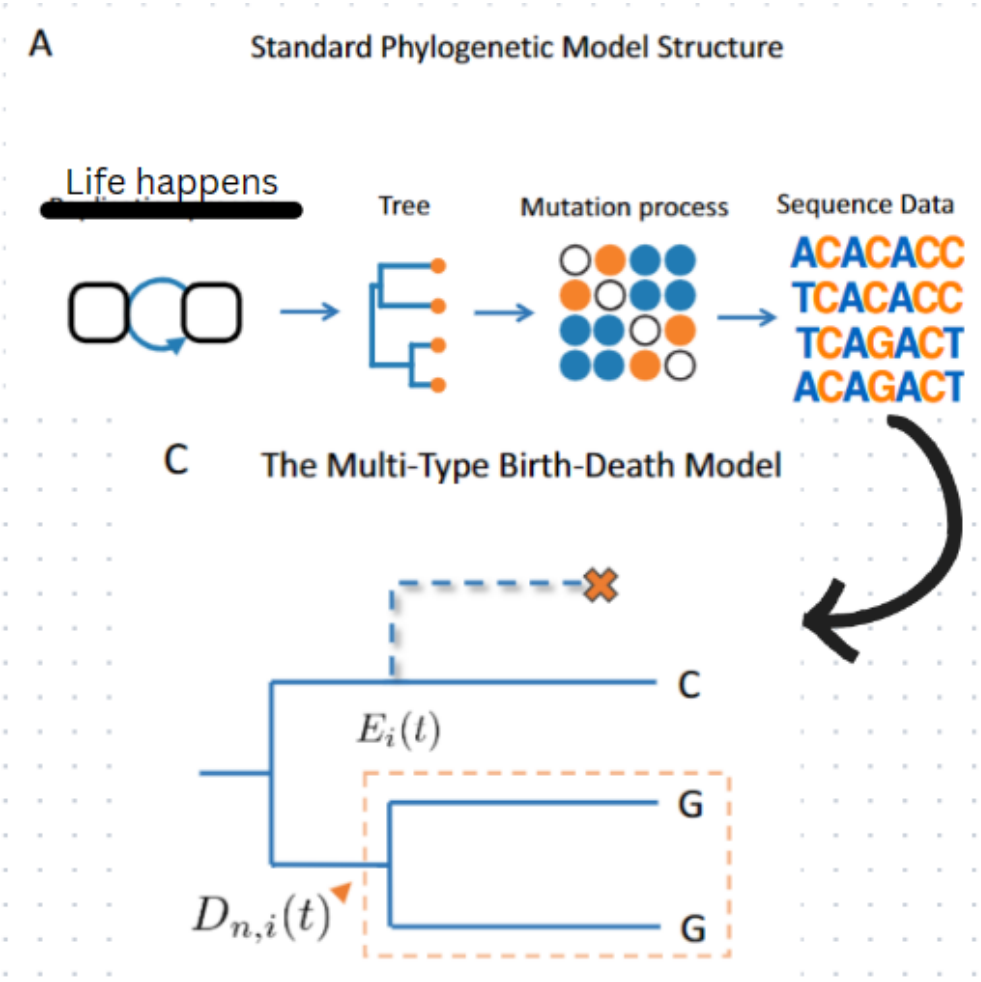


Figure: Doctored version of Fig 1 (Rasmussen, 2019)

Contemporaneous challenges

1. Fitness variation of muts affects tree process for microbes (viruses) $\rightarrow$ **feedback loop**
   - Common arg for indep. tree, mut processes in macro speciation is separate time scales
   - Distr. of fitness effects (DFE) is central ??? in evo bio

2. MTBD model (Stadler, 2013) compute scales $O(2^L)$ for binary char
   - (Stadler, 2013; Barrido-Sottani, 2018) study three **phenotypes** for HIV transmission

3. Existing methods to study selection in phylogenetics use **dN/dS** w/ codon subs. models
   - (Harris slides, 2022; Temple, Waples, & Browning, 2023+) indicate other methods required for recent selection

Sneak peak: MFBD

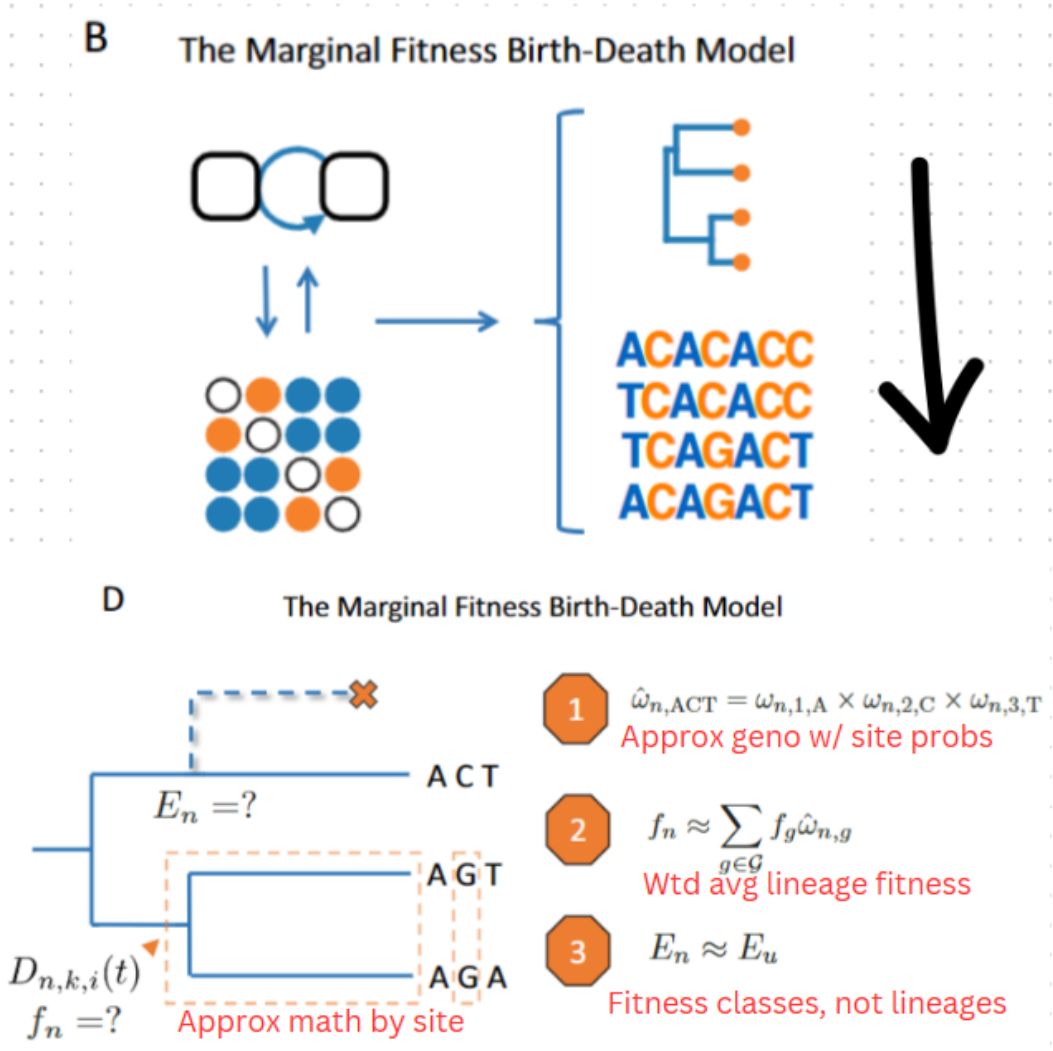


Figure: Doctored version of Fig 1 (Rasmussen, 2019)

Remaining Agenda

1. Mathematical notation
2. MTBD: multi type birth death
   - 2.1 Two coupled ODEs
   - 2.2 Computing up tree to root
   - 2.3 Some equations explained
3. MFBD: marginal fitness birth death
   - 3.1 Key assumptions
   - 3.2 Some equations explained
4. Simulation studies
   - 4.1 Exact MTBD versus approx MFBD
   - 4.2 Distribution of fitness effects
5. Ebola analysis (epistasis)
6. Takeaways from influenza analysis

Notation

- $_{n,m}$: lineages
- $_{i,j}$: state types (genotypes, alleles)
- $_{k,l}$: sites
- λ, d : birth, death rates
- γ : transition rates
- f, σ : fitness effects (gtype, site-specific)
- s, ρ : sampling rate at (death, $t = 0$)
- $D(t)$: density subtree observations (descending)
- $E(t)$: density state type leaves no observations (extinct)

More definitions at
sdtemple.github.io/misc.html
sdtemple.github.io/files/mfbd-handout-sdtemple.pdf

MTBD: D descending functions

$$\frac{d}{dt}D_{n,i}(t) = -(\lambda_i + \sum_{j=1}^{M} \gamma_{i,j} + d_i)D_{n,i}(t)$$ (1) No birth, no state transitions, no death

$$+2\lambda_i E_i(t) D_{n,i}(t)$$ (2) Birth of unsampled

$$+\sum_{j=1}^{M} \gamma_{i,j} D_{n,j}(t)$$ (3) State transition

You can develop/motivate this model as a continuous time Markov chain (Temple, 2021)

Here + is separate in a total law sense.
The $\times$ / $\cdot$ is independence in contemporaneous event sense.

MTBD: E extinct functions

$$\frac{d}{dt}E_i(t) = (1 - s_i)d_i$$

(1) not sampled at death

$$-(\lambda_i + \sum_{j=1}^{M} \gamma_{i,j} + d_i)E_i(t)$$

$$+\lambda_i E_i(t)^2$$

(2) birth, both went extinct

$$+\sum_{j=1}^{M} \gamma_{i,j}E_j(t).$$

Those lines not annotated have similar interpretation as last.

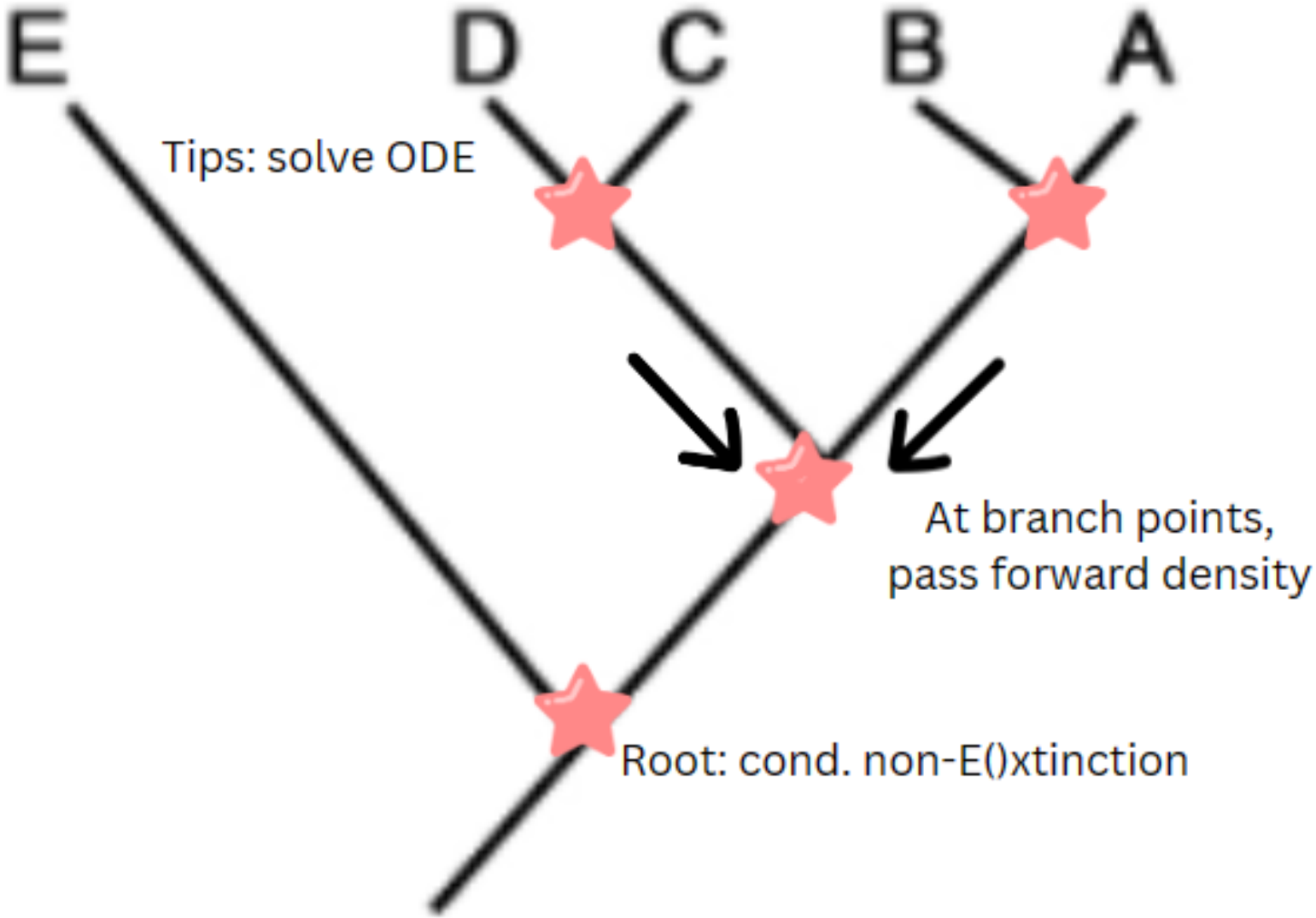
Computing from tips to root

Based on pedigree peeling (Cannings, 1976) [Thompson], tree pruning (Felsenstein, 1981)

1. At tip of lineage $_n$, **initalize ODE**
   $D_{n,i}(t) = d_i s_i$ if t at death, else $= \rho_i$
   (Also, initialize for $E(\cdot)$)
2. Solve ODEs
3. Moving up (back in real time), at branch point,
   $D_{a,i} = 2\lambda_i D_{m,i}(t) D_{n,i}(t)$
4. At root, form likelihood | non-extinction
   (akin to cond. sweep in SLiM)

$$D_n = \sum_{i=1}^{I} q_i D_{n,i}(t_{root}) / [1 - E_i(t_{root})]$$

Computing from tips to root



MFBD: key assumptions

1. Approx genotype probs with marginal site probs
2. Marginalize over genotype fitness w/ wtd geno probs to get site-specific fitness effects
3. Study discrete space of fitness classes for E densities
4. Trans rates $i \to j$ at site k don't depend on genetic background

Figure: 2.,3. explained next slide.
1. is nearly same as MTBD but w/ fitness-dependent births.

1. Update $D_{n,k,i}$ by numerically integrating (***Equation 19***) over time step Δt.
2. Update the marginal site probabilities $\omega_{n,k,i}$ using (***Equation 20***)
3. Update the expected marginal fitness values $\hat{f}_{n,k,i}$ using (***Equation 13***) or (***Equ***

MFBD: genotype probabilities, fitness

2.

\prod implies independence

$$\hat{\omega}_{n,g} = \frac{\prod_{k=1}^{L} \omega_{n,k,g_k}}{\sum_{g \in \mathcal{G}} \prod_{k=1}^{L} \omega_{n,k,g_k}}.$$

\sum over genotypes is a total law

3.

$$\hat{f}_{n,k,i} = \sum_{\{g \in \mathcal{G}: g_k = i\}} f_g \hat{\omega}_{n,g}.$$

weighted average, infer f's

If fitness multiplicative

$$\hat{f}_{n,k,i} = \sigma_{ki} \prod_{l=1, l \neq k}^{L} \sum_{j=1}^{M} \sigma_{lj} \omega_{n,l,j},$$

Sim study: exact MTBD or approx MFBD

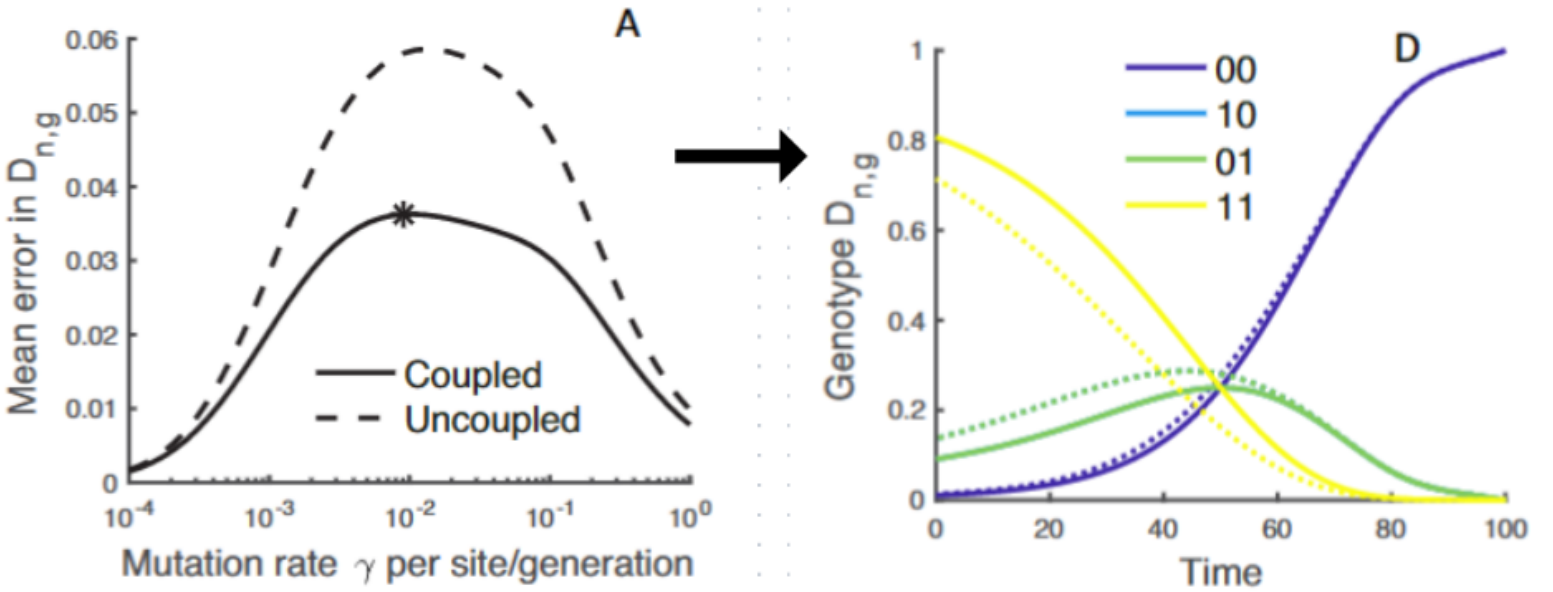


Figure: Comparing MFBD versus MTBD truth, where mean is over genotypes and time-integrated. Genotypes with favored allele 1, i.e., 01, 10, 11, more frequent as $t \to 0$

Sim study: exact MTBD or approx MFBD

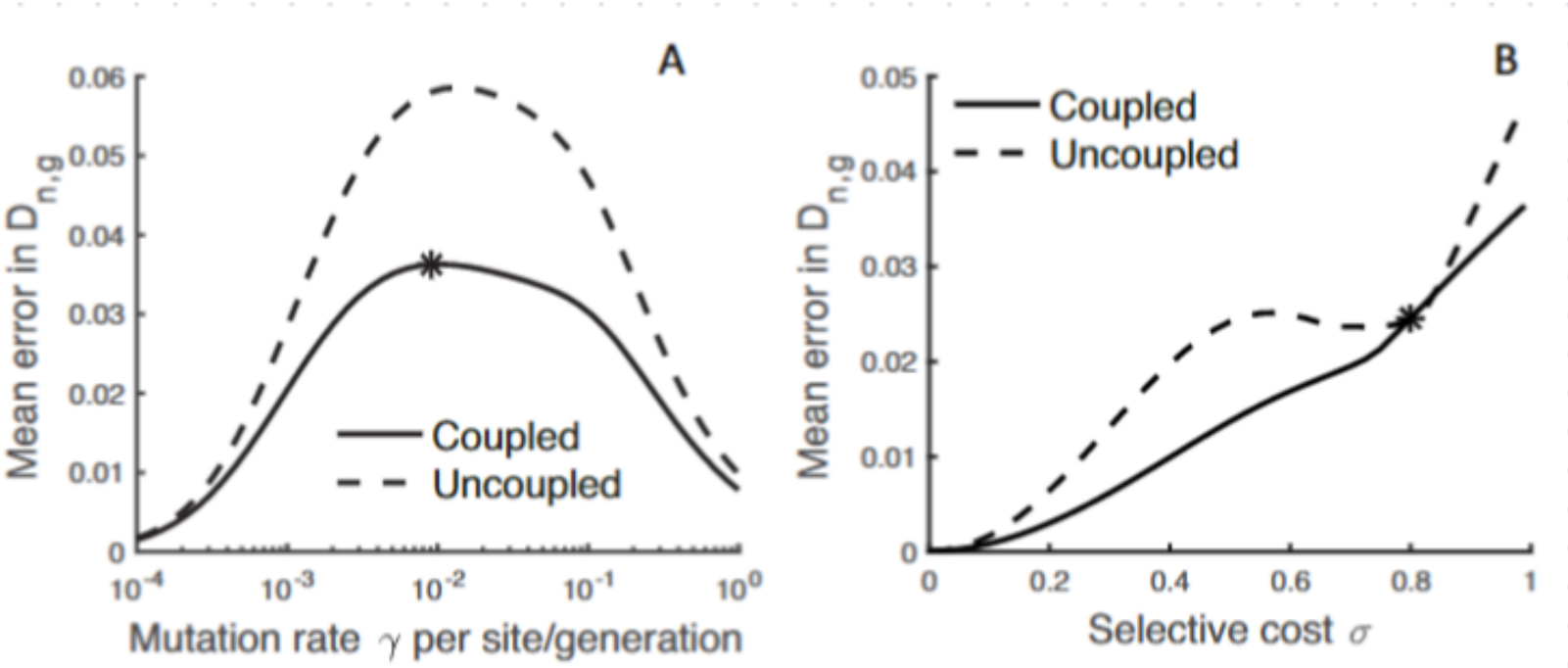


Figure: Comparing MFBD versus MTBD truth, where mean is over genotypes and time-integrated. Uncoupled (dashed) is naive method and not of interest. Qualitatively similar finding when study epistatis.

Sim study: site-specific fitness

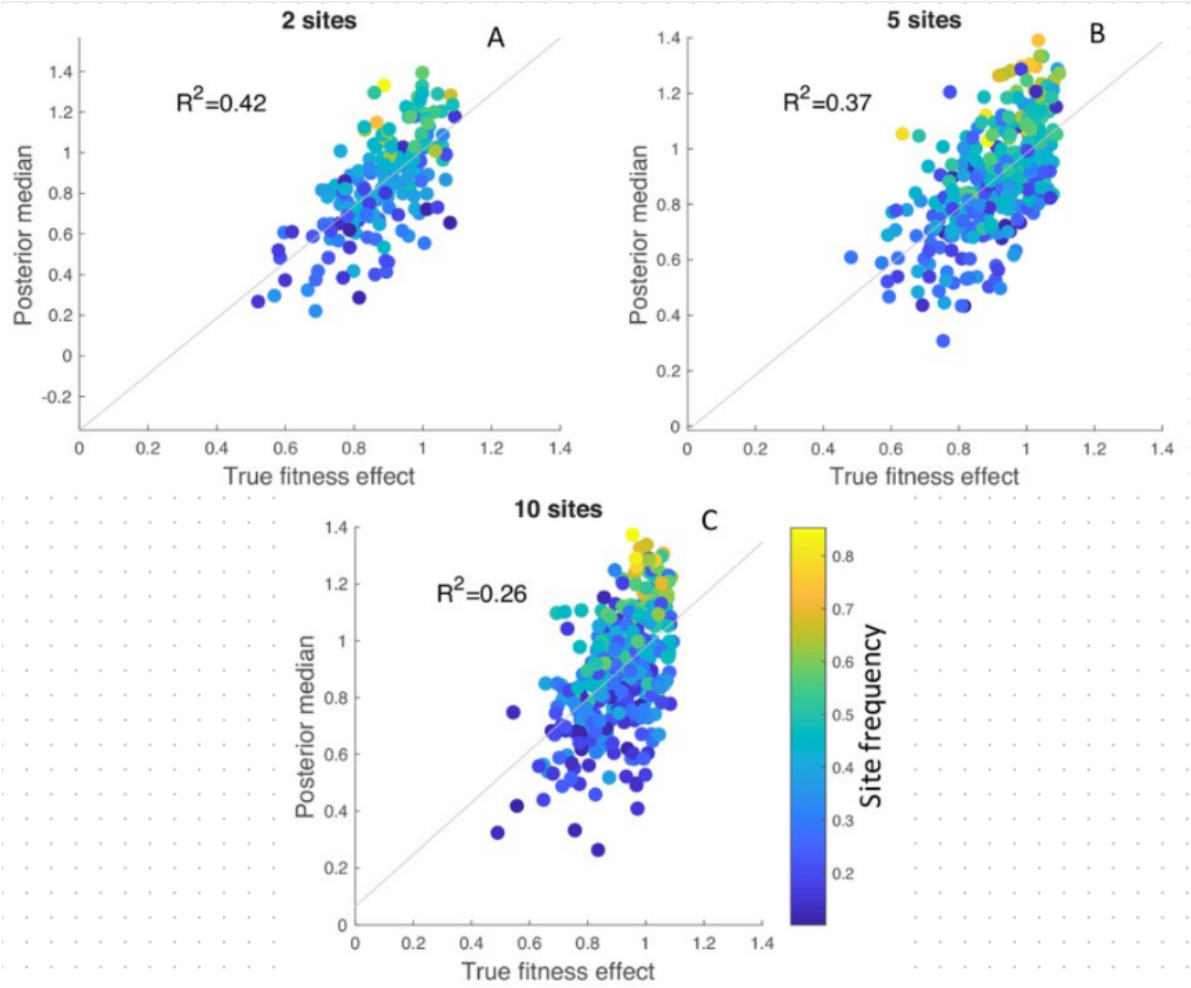


Figure: Site-specific fitness estimates and true fitness decreasing with num sites

Ebola data analysis

- Outbreak in West Africa 2014-2016
- 1610 viral samples

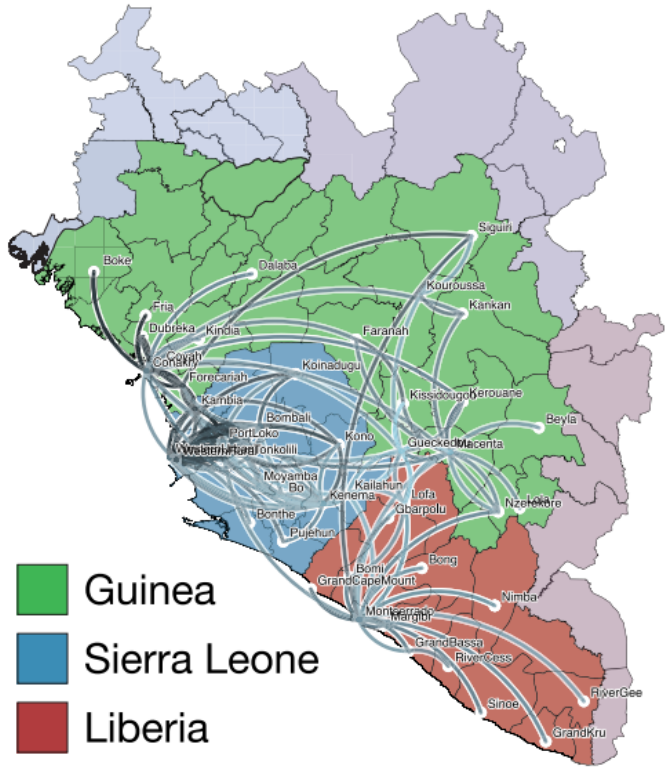


Figure: Epi dynamics from (Suchard, 2018).
Also studied by (Stockdale, 2021), (Temple, 2021), others.

Ebola results: superimposed phylo tree

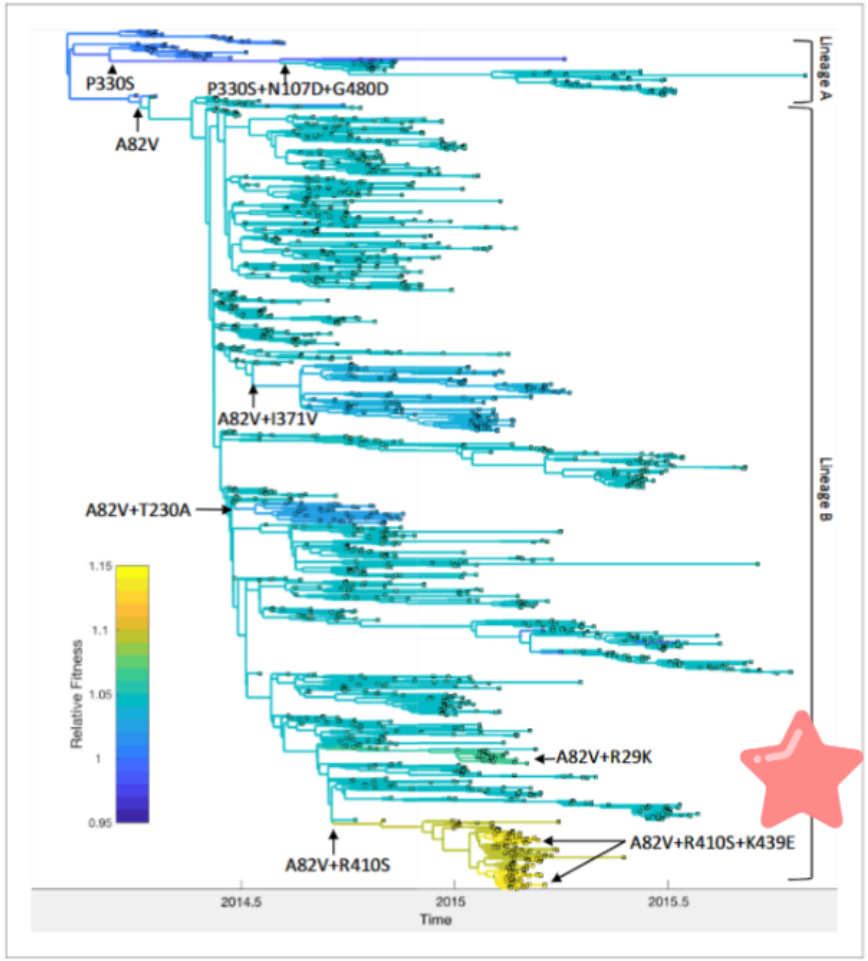


Figure: Inferred tree and MFBD params for Ebola 2014-2016. Red star highlights recent triple mutation of highest fitness, evidencing epistasis.

Ebola results: genotypes table

Table 1. Estimated posterior median fitness and 95% CI for the Ebola GP mut

Genotype	Sample freq	Base model
Makona	0.036	1.00
A82V	0.720	1.05 (1.04–1.07)
P330S	0.002	0.98 (0.82–1.14)
P330S+N107D+G480D	0.037	1.04 (0.98–1.12)
A82V+R410S	0.044	1.09 (1.00–1.18)
A82V+R410S+K439E	0.035	1.14 (1.01–1.26)
A82V+R29K	0.019	1.06 (0.93–1.19)
A82V+T230A	0.026	1.03 (0.93–1.11)
A82V+I371V	0.067	1.03 (0.98–1.09)

Figure: Doctored Table 1 from (Rasmussen, 2019). Used genotype instead of site MFBD model b/c other study suggested epistasis. Genotype ranks preserved when geography incorporated.
Fitness between hosts in pop **attenuated** compared to Urbanowski cell culture infectivity

Influenza data analysis $\rightarrow$ takeaways

- BEAST2 MCMC never converged $\rightarrow$ caution
- Pop level fitness est *in vivo* did not correlate well with deep mutational scanning *in vitro* (yeast?)
  - But, ad hoc effort to incorporate DMS outside info improved likelihood substantially
- Rapid turnover in flu viruses (we all get sick) (Volz, 2013)
  - Phylogeography may be important
- Many muts occur once in phylogeny, same background as other muts in HA protein
  - “identifiability ... akin to ... collinearity in ... regression”

Concluding remarks

- Marginal fitness birth death model is **approx, scalable** way to study fitness effects in rapidly evolving microbes
  - Builds upon multi type birth death model by approximating multi site genotypes with site-specific contributions
- Real data analysis **teaches that** epidemiological fitness *in vivo* population $\neq$ cellular infectivity *in vitro*
- Class: build up likelihoods w/ coupled ODEs, model correlated phylogenetic obs with trees
- Stats lesson: take a simplifying assumption, and test how well things work
- (Opinion: baking MFBD into tree explore in BEAST2 still does not fully study HUGE tree space. Know from coalescent theory how unwieldy trees are.)

Pause: Q & A

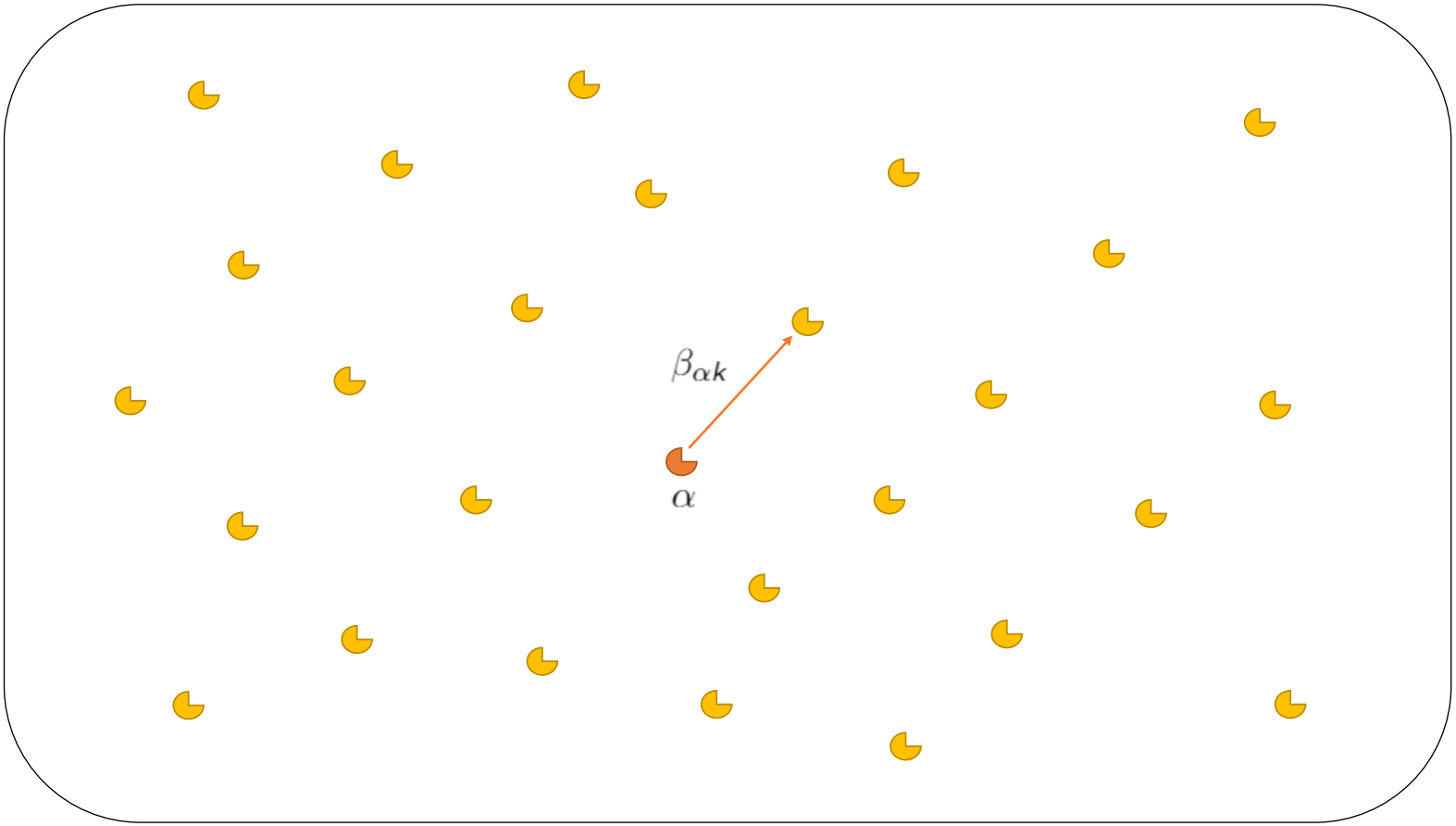
Appendices and references following

More relevant materials at sdtemple.github.io, including:

- Commentary on unifying BDS and SIR (MacPherson et al., 2021)
- Report on PBLA (Stockdale et al., 2021): https://github.com/sdtemple/pblas

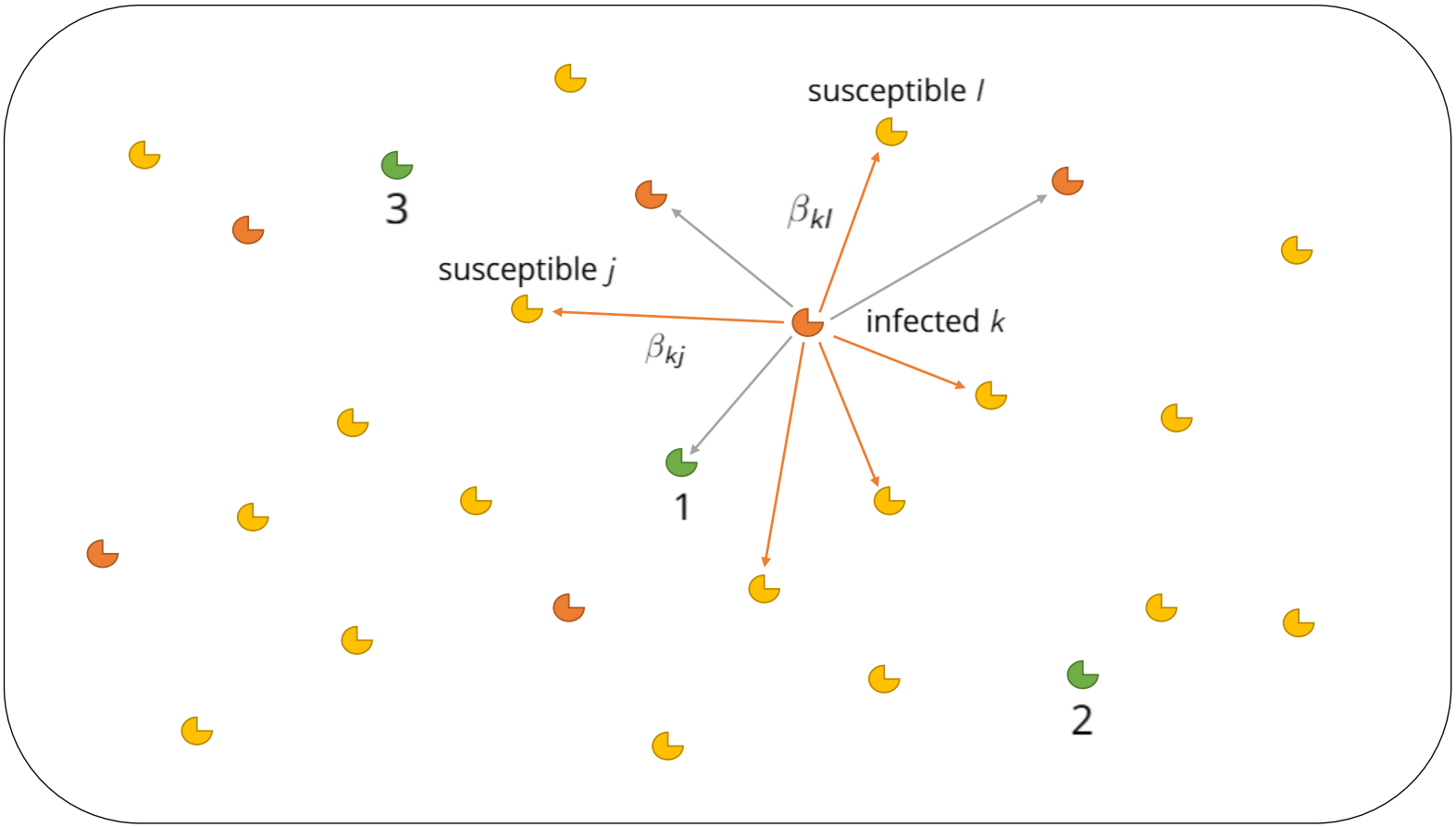
Stochastic epidemic (Stockdale, 2021; Temple, 2021)

Infection rates β_{kj} and removals after $r_j - i_j \sim \text{Gamma}(m_j, \gamma_j)$



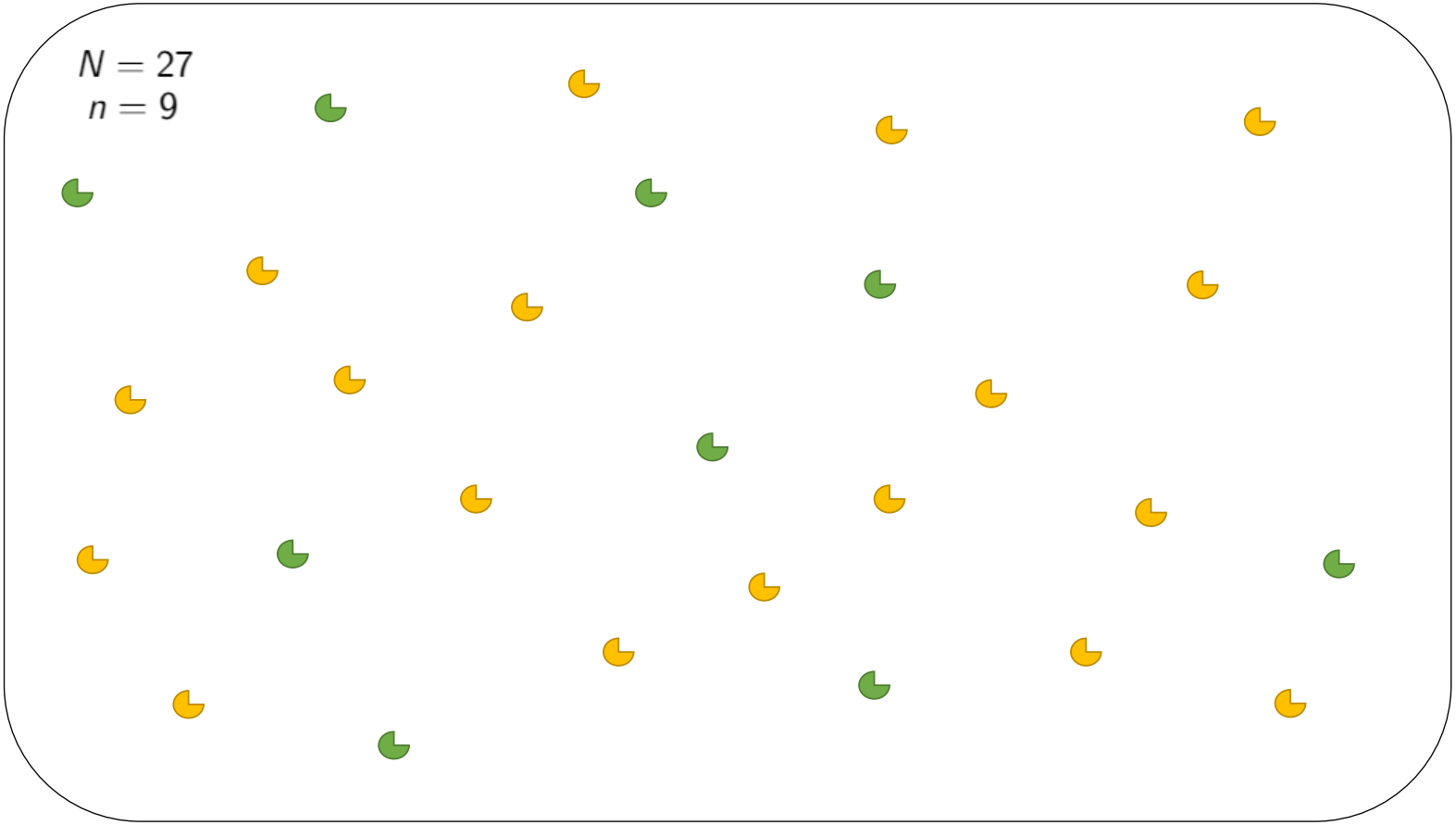
Stochastic epidemic

At time t, $S(t)$ susceptibles, $I(t)$ infecteds, and $R(t)$ removeds, with $N = S(t) + I(t) + R(t)$.



Stochastic epidemic

Epidemic ends when $I(t) = 0$.



Stochastic epidemic

To simulate an epidemic, we exploit **Poisson processes (PPs)**. Define a ***race*** as the minimum of (exponential) rvs.

Algorithm (Epidemic Simulator)

1. $S(0) = N - 1, I(0) = 1$
2. Until $I(t) = 0$:
   - 2.1 **Race $S(t)I(t)$ PPs with rate β and $I(t)$ PPs with rate γ,** where t_1 is the winning race time.
   - 2.2 If a γ-PP wins, $I(t_1) = I(t) - 1$ and $R(t_1) = R(t) + 1$.
   - 2.3 If a β-PP wins, $S(t_1) = S(t) - 1$ and $I(t_1) = I(t) + 1$.
   - 2.4 Update $t = t_1$.

This is way to frame (Gillespie, 1977) for SIR model

Stochastic epidemic model

$\{S(t), I(t)\}$ is a continuous-time Markov chain (CTMC) with "jumps" based on an underlying Poisson process.

- $\tau_{kj} := r_k \wedge i_j - i_k \wedge i_j$
  - Time k tries to infect j
- $\psi_j = \exp(-\sum_{k \neq j}^{n} \beta_{kj} \tau_{kj})$
  - Probability j not infected before i_j
  - $\psi_{kj} = \exp(-\beta_{kj} \tau_{kj})$ is marginal term
- $\chi_j = \sum_{k \neq j}^{n} \beta_{kj} 1_{\{i_k < i_j < r_k\}}$
  - Probability j infected at i_j
- $\phi_j = \exp(-\sum_{k=n+1}^{N} \beta_{jk} (r_j - i_j))$
  - Probability j doesn't infect never-infecteds

References I


M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, et al. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10 virus evolution 4 (1): vey016. *DOI: http://doi. org/10.1093*, 1093.


D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81(25):2340–2361, Dec. 1977.


C. Cannings, E. A. Thompson, and M. H. Skolnick. Probability functions on complex pedigrees. *Adv. Appl. Probab.*, 10(1):26–61, Mar. 1978.


J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. en. *J. Mol. Evol.*, 17(6):368–376, 1981.

References II

- J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.*, 19(A):27–43, 1982.
- J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.*, 13(3):235–248, Sept. 1982.
- R. C. Griffiths and S. Tavare. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.*, 46(2):131–159, Oct. 1994.
- A. Eyre-Walker and P. D. Keightley. The distribution of fitness effects of new mutations. en. *Nat. Rev. Genet.*, 8(8):610–618, Aug. 2007.

References III


T. Stadler and S. Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. en. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 368(1614):20120198, Mar. 2013.


E. M. Volz, K. Koelle, and T. Bedford. Viral phylodynamics. en. *PLoS Comput. Biol.*, 9(3):e1002947, Mar. 2013.


J. Barido-Sottani, T. G. Vaughan, and T. Stadler. Detection of HIV transmission clusters from phylogenetic trees using a multi-state birth-death model. en. *J. R. Soc. Interface*, 15(146), Sept. 2018.

References IV

- D. A. Rasmussen and T. Stadler. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth-death models. en. *Elife*, 8, Aug. 2019.
- A. MacPherson, S. Louca, A. McLaughlin, J. B. Joy, and M. W. Pennell. Unifying phylogenetic Birth–Death models in epidemiology and macroevolution. en. *Syst. Biol.*, June 2021.
- J. E. Stockdale, T. Kypraios, and P. D. O'Neill. Pair-based likelihood approximations for stochastic epidemic models. en. *Biostatistics*, 22(3):575–597, July 2021.

References V

S. D. Temple. PhD preliminary exam report on PBLA for stochastic epidemic models. *https://github.com/sdtemple/pblas*, 2021.

S. D. Temple, R. W. Waples, and S. R. Browning. Robust statistical inference of very recent and strong incomplete selective sweeps. *https://github.com/sdtemple/iSWEEP*, 2023.