Inference on population trees by approximating Wright-Fisher diffusions

Jukka Sirén
University of Helsinki

19th June 2012

Acknowledgment

Joint work with

- Jukka Corander
  University of Helsinki
- Pekka Marttinen
  Aalto University
- Bill Hanage
  Harvard University

Population tree – Species tree

- Both describe historical relations between groups, which are (approximately) isolated.

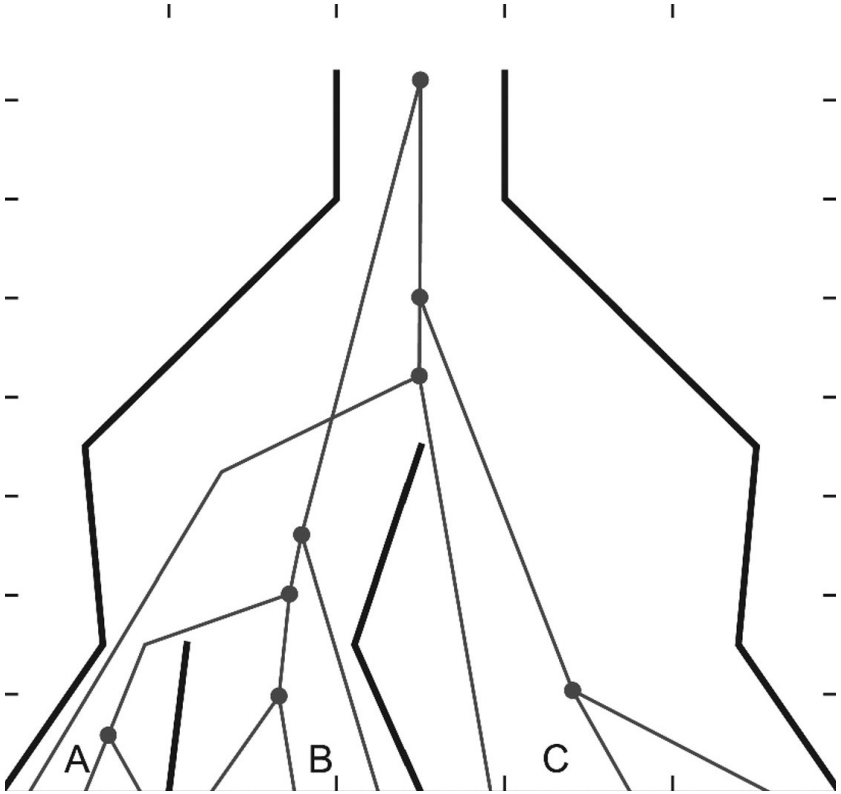
Population tree – Species tree

- Both describe historical relations between groups, which are (approximately) isolated.
- Difference comes from the main source of genetic variation: mutation vs drift.

Population tree – Species tree

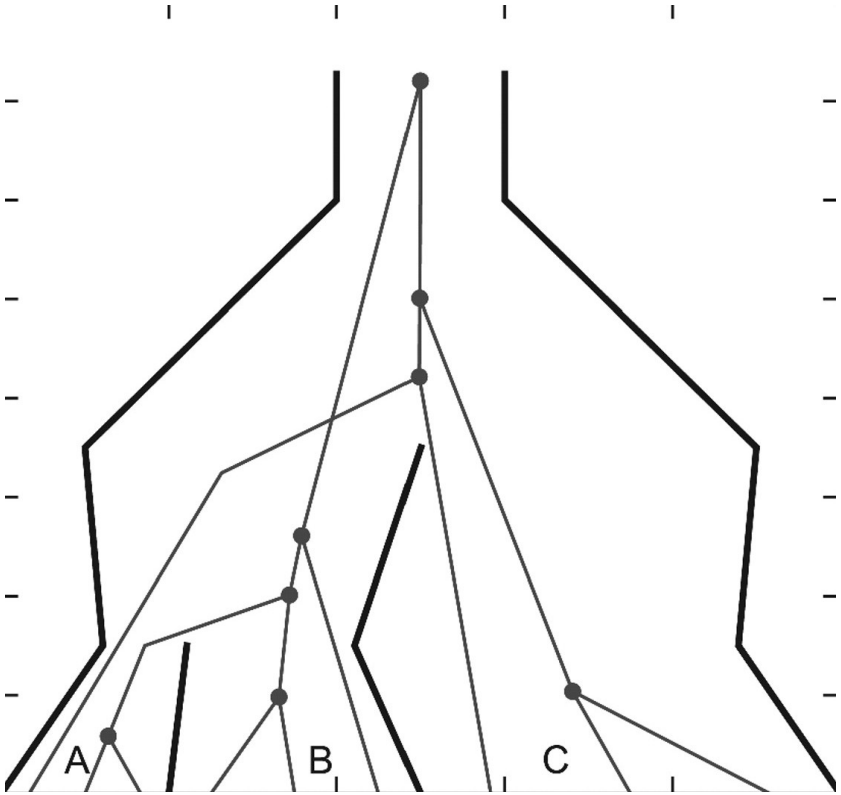
- Both describe historical relations between groups, which are (approximately) isolated.
- Difference comes from the main source of genetic variation: mutation vs drift.
- This work concentrates on methods which are applicable to data with large numbers of markers (SNP,AFLP,MLST etc.)

Multi-species coalescent



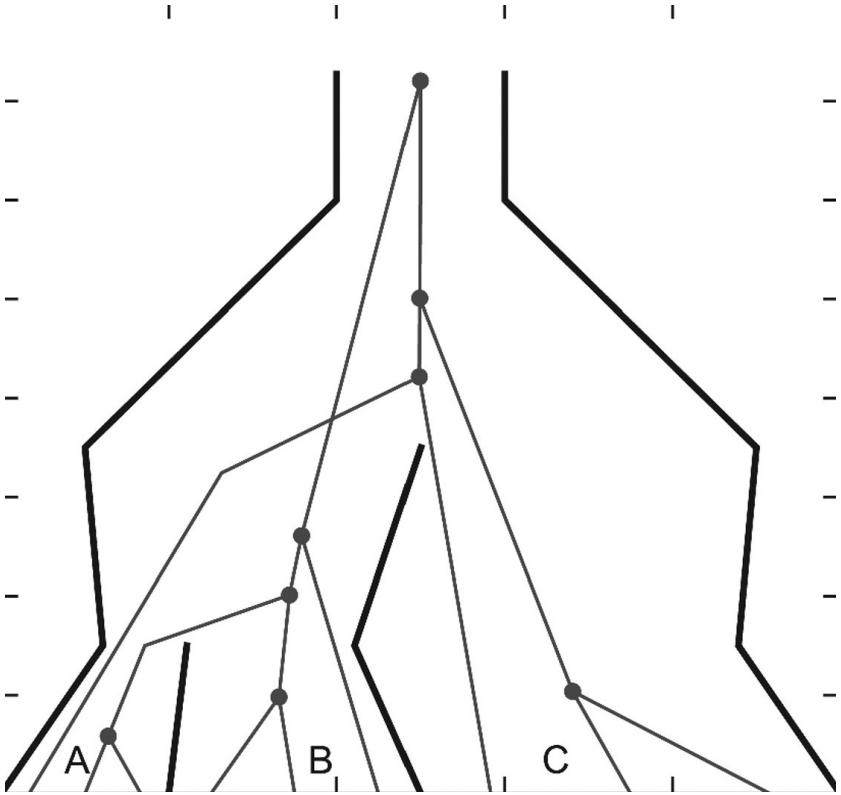
- Computationally very expensive.

Multi-species coalescent



- For biallelic markers efficient algorithms to compute the likelihood of a species tree have been recently introduced. (RoyChoudhury et al 2008 Genetics, Bryant et al 2012 MBE)

Multi-species coalescent



- For biallelic markers efficient algorithms to compute the likelihood of a species tree have been recently introduced. (RoyChoudhury et al 2008 Genetics, Bryant et al 2012 MBE)
- But they are still limited to relatively small data sets.

Wright-Fisher model

- An alternative to the coalescent is provided by the Wright-Fisher model.

Wright-Fisher model

- An alternative to the coalescent is provided by the Wright-Fisher model.
- If X_t is the count of an allele A in a population of size N at time t, then

$$X_{t+1} \mid X_t \sim Bin(N, X_t/N).$$

Wright-Fisher model

- An alternative to the coalescent is provided by the Wright-Fisher model.
- If X_t is the count of an allele A in a population of size N at time t, then

$$X_{t+1} \mid X_t \sim Bin(N, X_t/N).$$

- By scaling with $\psi_t = X_t/N$ and $\tau = t/N$, we get diffusion approximation

$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim N(\psi_\tau, \epsilon\psi_\tau(1-\psi_\tau)).$$

Wright-Fisher model

- An alternative to the coalescent is provided by the Wright-Fisher model.
- If X_t is the count of an allele A in a population of size N at time t, then
$$X_{t+1} \mid X_t \sim Bin(N, X_t/N).$$
- By scaling with $\psi_t = X_t/N$ and $\tau = t/N$, we get diffusion approximation
$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim N(\psi_\tau, \epsilon\psi_\tau(1-\psi_\tau)).$$
- Used in phylogenetics by Cavalli-Sforza, Edwards, Felsenstein and others in the 60's and 70's.

Wright-Fisher model

- Diffusion approximation

$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim N(\psi_\tau, \epsilon\psi_\tau(1-\psi_\tau).$$

- The Gaussian approximation is good only for small values of ϵ. (Atoms on boundaries, scaling).

Wright-Fisher model

- Diffusion approximation

$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim N(\psi_\tau, \epsilon\psi_\tau(1-\psi_\tau).$$

- The Gaussian approximation is good only for small values of ϵ. (Atoms on boundaries, scaling).
- Alternative, Balding-Nichols model:
  Beta-distribution instead of Gaussian.

Wright-Fisher model

- Diffusion approximation

$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim N(\psi_\tau, \epsilon\psi_\tau(1-\psi_\tau).$$

- The Gaussian approximation is good only for small values of ϵ. (Atoms on boundaries, scaling).
- Alternative, Balding-Nichols model: Beta-distribution instead of Gaussian.
- Should be a better approximation.

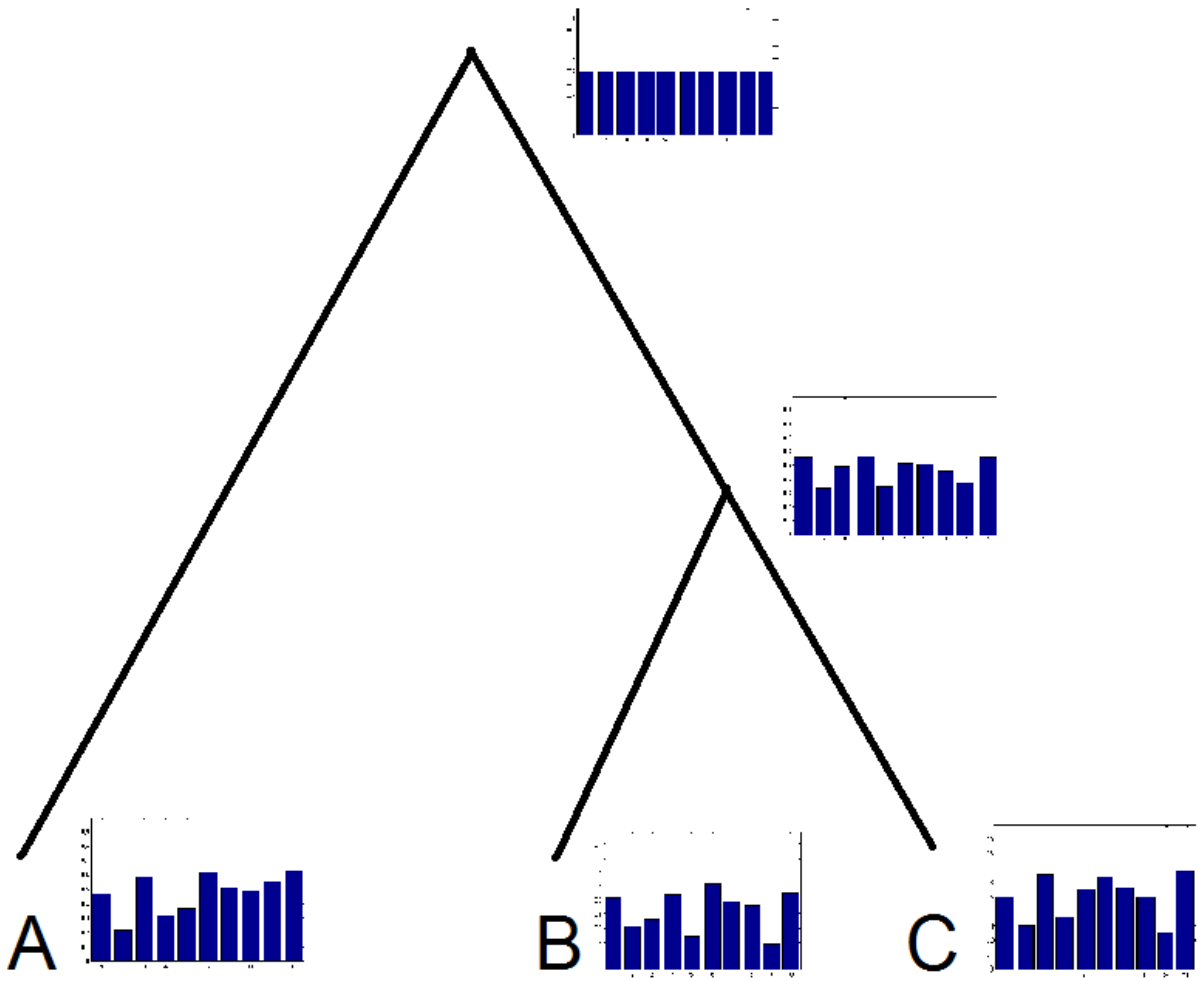
Wright-Fisher model

- Diffusion approximation

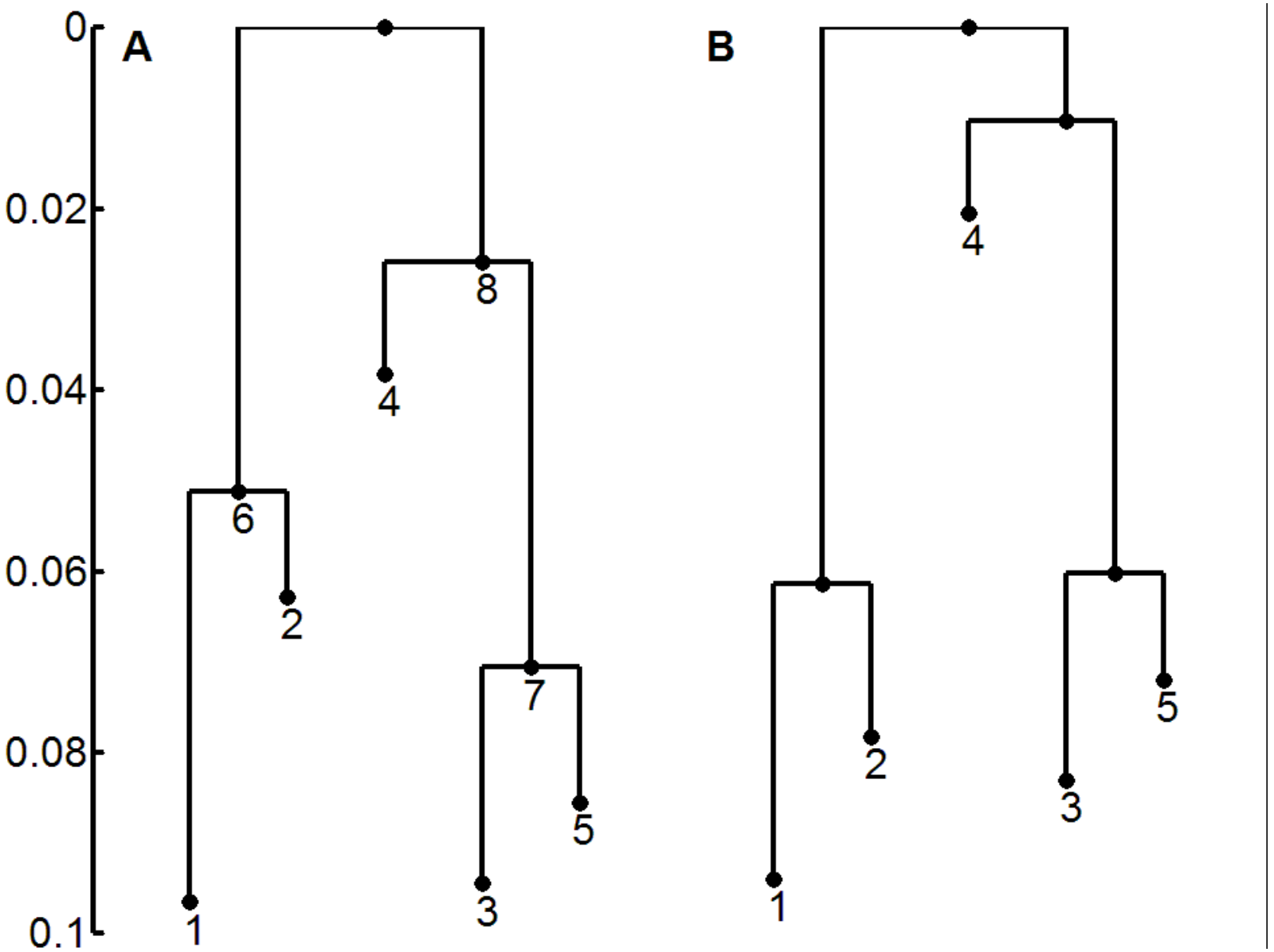
$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim N(\psi_\tau, \epsilon\psi_\tau(1-\psi_\tau).$$

- The Gaussian approximation is good only for small values of ϵ. (Atoms on boundaries, scaling).
- Alternative, Balding-Nichols model: Beta-distribution instead of Gaussian.
- Should be a better approximation.
- Also, computational advantages due conjugacy as the sampling from populations often binomial.

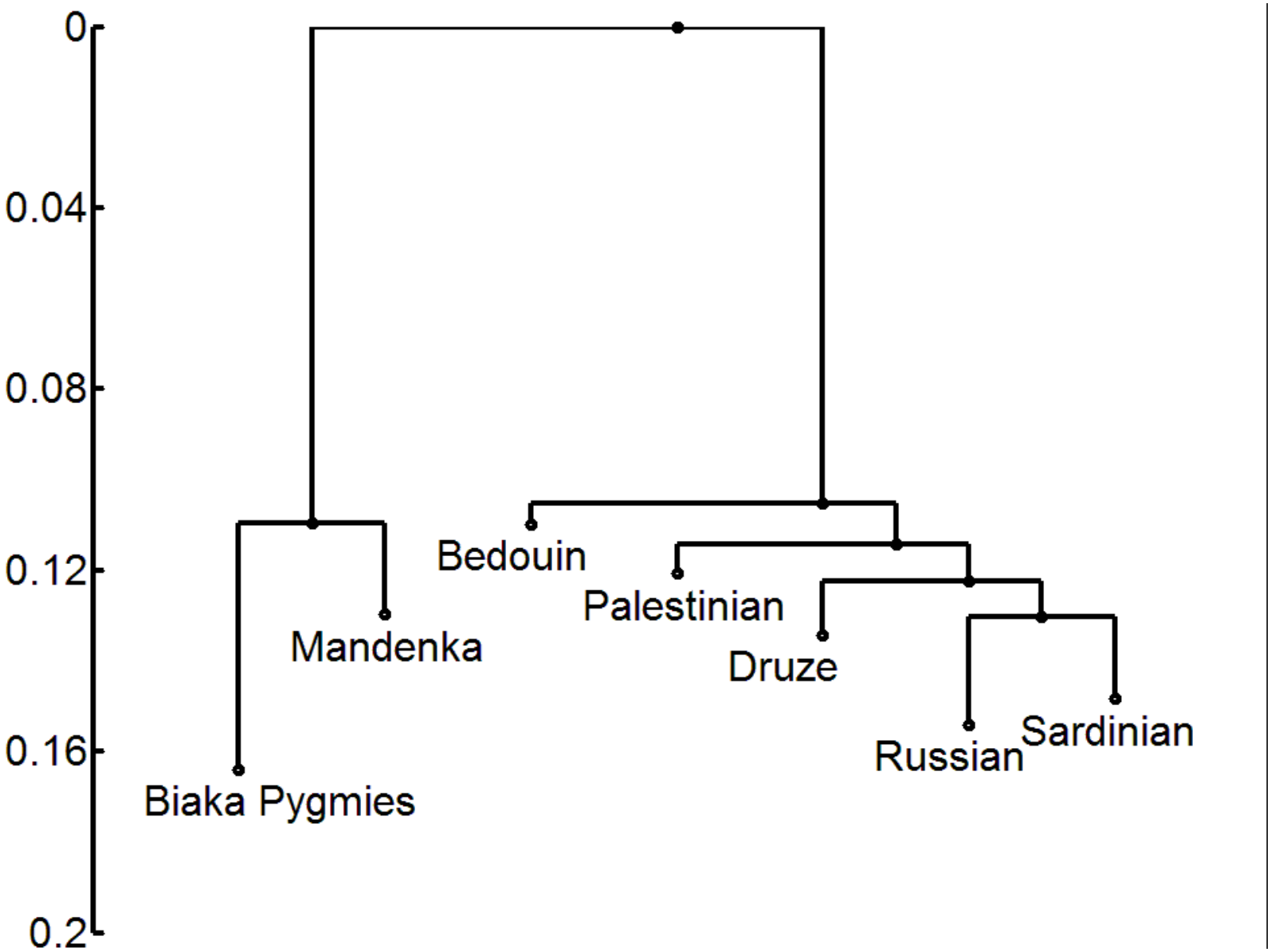
Full model



Tree used in simulation (A) and estimated tree (B)



Tree of human populations from SNPs



WF-model with mutation

- Extend the simple WF-model with mutation occuring at rate u from A and at rate v to A.

WF-model with mutation

- Extend the simple WF-model with mutation occuring at rate u from A and at rate v to A.
- The one-generation distribution then changes to

$$X_{t+1} \mid X_t \sim \mathit{Binomial}(N, \eta_t),$$

where $\eta_t = N^{-1}((1-u)X_t + v(N - X_t))$.

WF-model with mutation

- Extend the simple WF-model with mutation occuring at rate u from A and at rate v to A.
- The one-generation distribution then changes to

$$X_{t+1} \mid X_t \sim \mathit{Binomial}(N, \eta_t),$$

where $\eta_t = N^{-1}((1-u)X_t + v(N - X_t))$.

- Again, diffusion approximation is obtained as

$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim N(\eta_\tau, \epsilon\eta_\tau(1-\eta_\tau)).$$

WF-model with mutation

- Extend the simple WF-model with mutation occuring at rate u from A and at rate v to A.
- The one-generation distribution then changes to

$$X_{t+1} \mid X_t \sim \mathit{Binomial}(N, \eta_t),$$

  where $\eta_t = N^{-1}((1-u)X_t + v(N - X_t))$.
- Again, diffusion approximation is obtained as

$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim N(\eta_\tau, \epsilon\eta_\tau(1-\eta_\tau)).$$

- Poor approximation, as the mean and variance do not scale linearly or nearly linearly with time.

WF-model with mutation

- Solution:
  Compute the actual expectation and variance of the WF process:

$$
\begin{aligned}
E(X_t) &= E(E(X_t \mid X_{t-1})) = \dots \\
Var(X_t) &= E(Var(X_t \mid X_{t-1})) + Var(E(X_t \mid X_{t-1})) = \dots
\end{aligned}
$$

WF-model with mutation

- Solution:
  Compute the actual expectation and variance of the WF process:

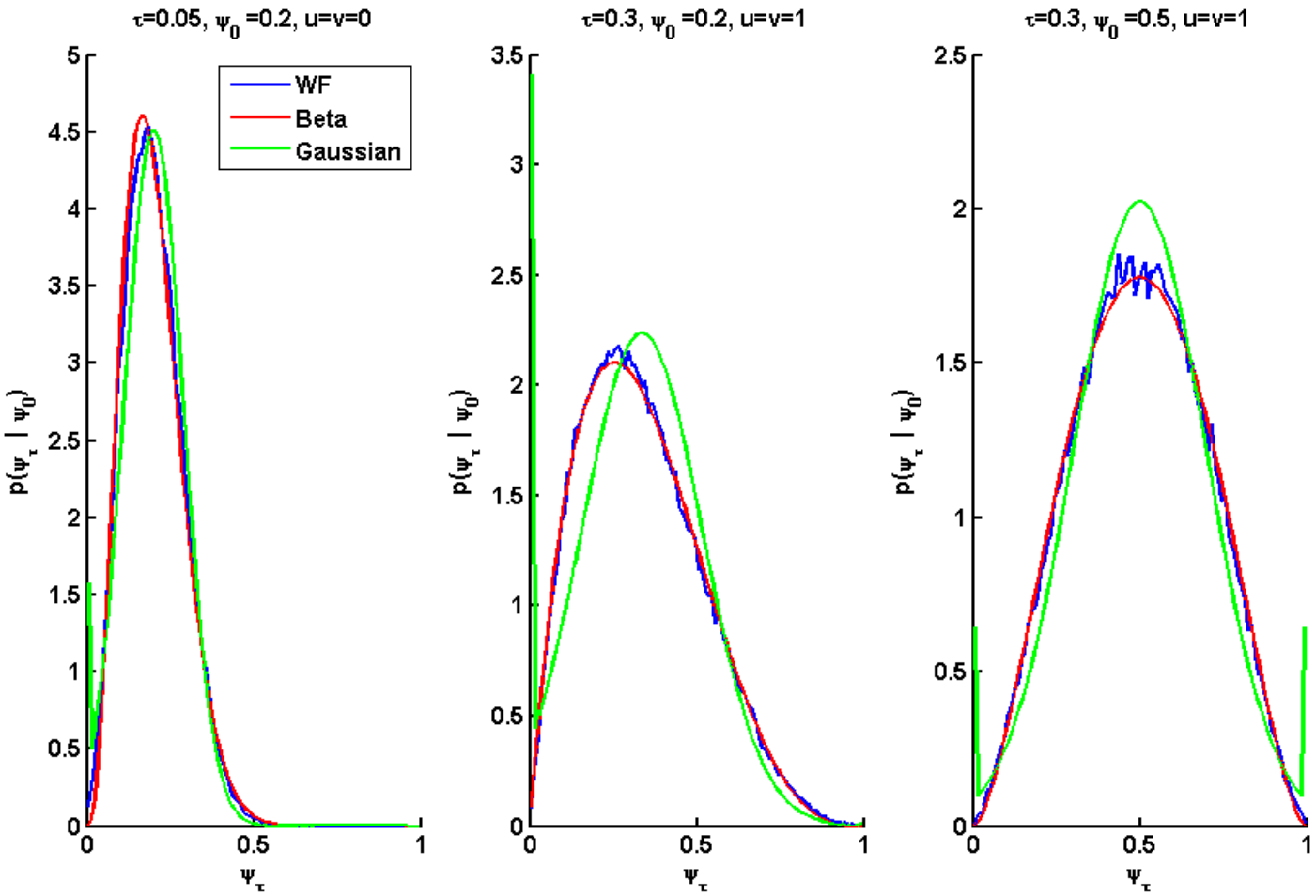
$$\begin{aligned} E(X_t) &= E(E(X_t \mid X_{t-1})) = \dots \\ Var(X_t) &= E(Var(X_t \mid X_{t-1})) + Var(E(X_t \mid X_{t-1})) = \dots \end{aligned}$$

- Then we can approximate the change with

$$\psi_{\tau+\epsilon} \mid \psi_\tau \sim Beta(\alpha_{\tau,\epsilon}, \beta_{\tau,\epsilon}),$$

  where $\alpha_{\tau,\epsilon}$ and $\beta_{\tau,\epsilon}$ are chosen to get the correct mean and variance.

Comparison of approximations



$N = 1e4$

Infinite alleles model

- The idea can be generalized to more complex situations such as (truncated) infinite alleles model.

Infinite alleles model

- The idea can be generalized to more complex situations such as (truncated) infinite alleles model.
- $r-1$ distinct alleles are followed and rth allele type represents all other alleles.

Infinite alleles model

- The idea can be generalized to more complex situations such as (truncated) infinite alleles model.
- $r-1$ distinct alleles are followed and rth allele type represents all other alleles.
- Consequently, mutation occurs only from the $r-1$ alleles to the rth allele (rate u).

Infinite alleles model

- The idea can be generalized to more complex situations such as (truncated) infinite alleles model.
- $r-1$ distinct alleles are followed and rth allele type represents all other alleles.
- Consequently, mutation occurs only from the $r-1$ alleles to the rth allele (rate u).
- Let $X_{i,t}$ denote the number of alleles of type i at time t, with population size N.

Infinite alleles model

- The idea can be generalized to more complex situations such as (truncated) infinite alleles model.
- $r-1$ distinct alleles are followed and rth allele type represents all other alleles.
- Consequently, mutation occurs only from the $r-1$ alleles to the rth allele (rate u).
- Let $X_{i,t}$ denote the number of alleles of type i at time t, with population size N.
- The allele counts have a multinomial distribution conditional on the alleles of the previous generation

$$X_{1t}, \ldots, X_{rt} | X_{1(t-1)}, \ldots, X_{r(t-1)} \sim \mathit{Multinomial}(N, \eta_t),$$

where η_t is a r dimensional vector with entries

$$\eta_{jt} = \begin{cases} 1-(1-u)\left(1-\psi_{r(t-1)}\right) & \text{if } j = r \text{ and} \\ (1-u)\psi_{j(t-1)} & \text{otherwise.} \end{cases}$$

Infinite alleles model

- Similarly, as in the biallelic case, we can explicitly compute the mean and covariance as

$$E(X_{it}) = E(E(X_{it} \mid X_{i(t-1)})) = \dots$$

and

$$\begin{aligned} Cov(X_{it}, X_{jt}) =& E(Cov(X_{it}, X_{jt} \mid X_{i(t-1)}, X_{j(t-1)})) \\ &+ Cov(E(X_{it}, X_{jt} \mid X_{i(t-1)}, X_{j(t-1)})) \\ =& \dots \end{aligned}$$

Beta-Dirichlet model

- Because of the covariance structure, a Dirichlet is not an adequate approximation.

Beta-Dirichlet model

- Because of the covariance structure, a Dirichlet is not an adequate approximation.
- Instead, we first model the mutation with

$$\psi_{r\tau} \sim \mathit{Beta}(\gamma\mu_{r\tau}, \gamma(1-\mu_{r\tau})). \quad (1)$$

Beta-Dirichlet model

- Because of the covariance structure, a Dirichlet is not an adequate approximation.
- Instead, we first model the mutation with

$$\psi_{r\tau} \sim \mathit{Beta}(\gamma\mu_{r\tau}, \gamma(1-\mu_{r\tau})). \tag{1}$$

- And then the drift with

$$(1-\psi_{r\tau})\,\psi_{0\tau} \mid \psi_{r\tau} \sim \mathit{Dirichlet}(\phi\psi_{10}, \ldots, \phi\psi_{(r-1)0}). \tag{2}$$

Beta-Dirichlet model

- Because of the covariance structure, a Dirichlet is not an adequate approximation.
- Instead, we first model the mutation with

$$\psi_{r\tau} \sim \mathit{Beta}(\gamma\mu_{r\tau}, \gamma(1-\mu_{r\tau})). \tag{1}$$

- And then the drift with

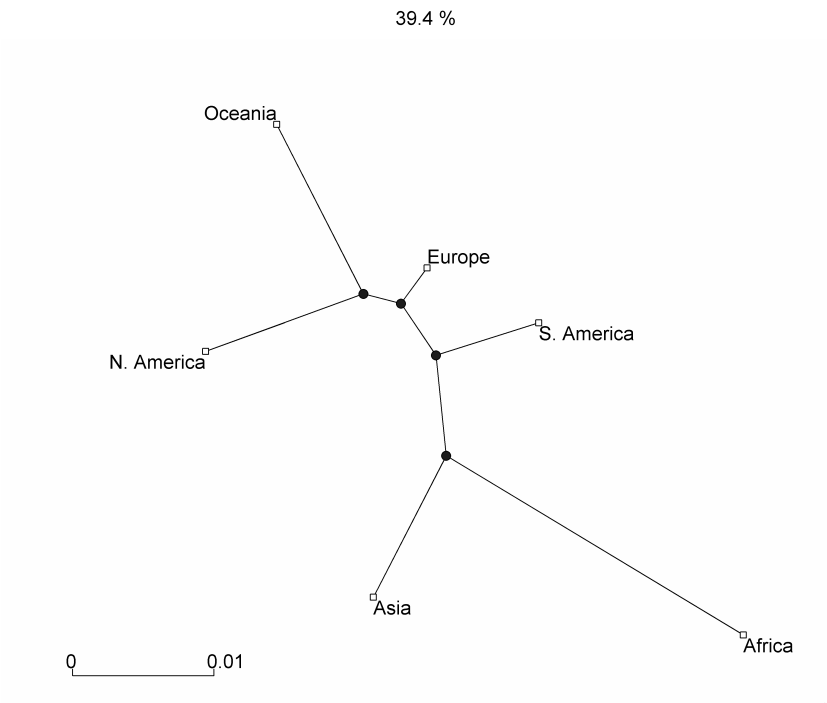
$$(1-\psi_{r\tau})\,\psi_{0\tau} \mid \psi_{r\tau} \sim \mathit{Dirichlet}(\phi\psi_{10}, \ldots, \phi\psi_{(r-1)0}). \tag{2}$$

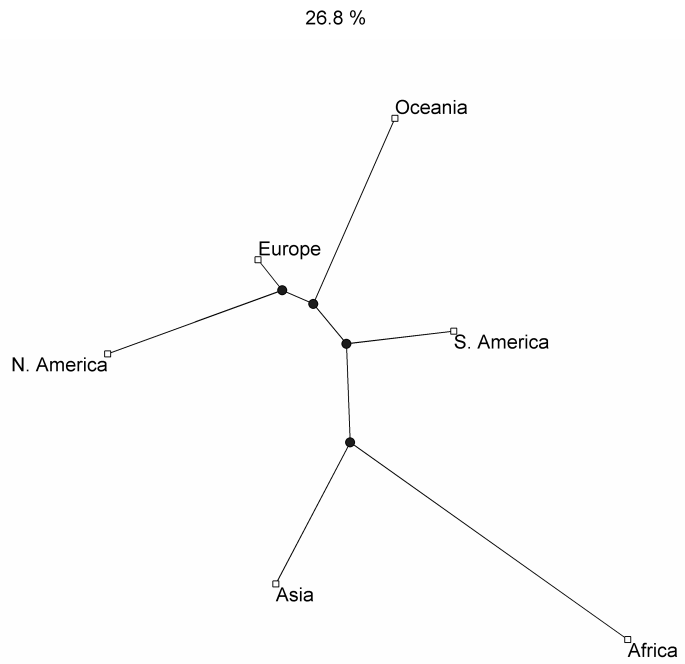
- To get same expectations and covariance structure as with the infinite alleles model, we use

$$\phi = \frac{(m+1)e^{-(m+1)\tau}}{1-e^{-(m+1)\tau}} \text{ and}$$

$$\gamma = \frac{m\left(1-e^{-(m+1)\tau}\right)}{\left(1-e^{-(m+1)\tau}\right) - (m+1)e^{-m\tau}\left(1-e^{-\tau}\right)}.$$

Global population structure of S. pneumoniae





Computation?

- Posterior distribution of the topolgy and branch specific parameters (time, mutation rate).

Computation?

- Posterior distribution of the topolgy and branch specific parameters (time, mutation rate).
- Different strategies with different models.

Computation?

- Posterior distribution of the topolgy and branch specific parameters (time, mutation rate).
- Different strategies with different models.
- In general case, MCMC.

Computation?

- Posterior distribution of the topolgy and branch specific parameters (time, mutation rate).
- Different strategies with different models.
- In general case, MCMC.
- For biallelic loci combination of Laplace approximization, AMIS and numerical maximization algorithms.

Future work

- Comparison with the coalescent approach.
- Computational strategies.
- Migration? Graphs instead of trees?
- More complex mutational models? Microsatellites?

References

Sirén J, Marttinen P, Corander J. 2011.
Reconstructing population histories from single nucleotide polymorphism data.
Molecular Biology and Evolution. **28**:673-683.

Sirén J. 2012.
Statistical models for inferring the structure and history of populations from genetic data.
PhD thesis. University of Helsinki.

Sirén J, Hanage WP, Corander J. 2012.
Inference on Population Histories by Approximating Infinite Alleles Diffusion.
Under revision.