Mathematical models in population genetics II

Anand Bhaskar

Evolutionary Biology and Theory of Computing Bootcamp

January 21, 2014

Quick recap

- Large discrete-time randomly mating Wright-Fisher population $\rightarrow$ continuous-time coalescent process for the genealogy of a sample of size n drawn at present
- Continuous-time Markov chain with transitions from k ancestral lineages to $k-1$ at rate $\binom{k}{2}$
- Mutations can be superimposed on the genealogical tree to generate allelic types in the sample

Quick recap

- Large discrete-time randomly mating Wright-Fisher population $\rightarrow$ continuous-time coalescent process for the genealogy of a sample of size n drawn at present
- Continuous-time Markov chain with transitions from k ancestral lineages to $k-1$ at rate $\binom{k}{2}$
- Mutations can be superimposed on the genealogical tree to generate allelic types in the sample

But ...

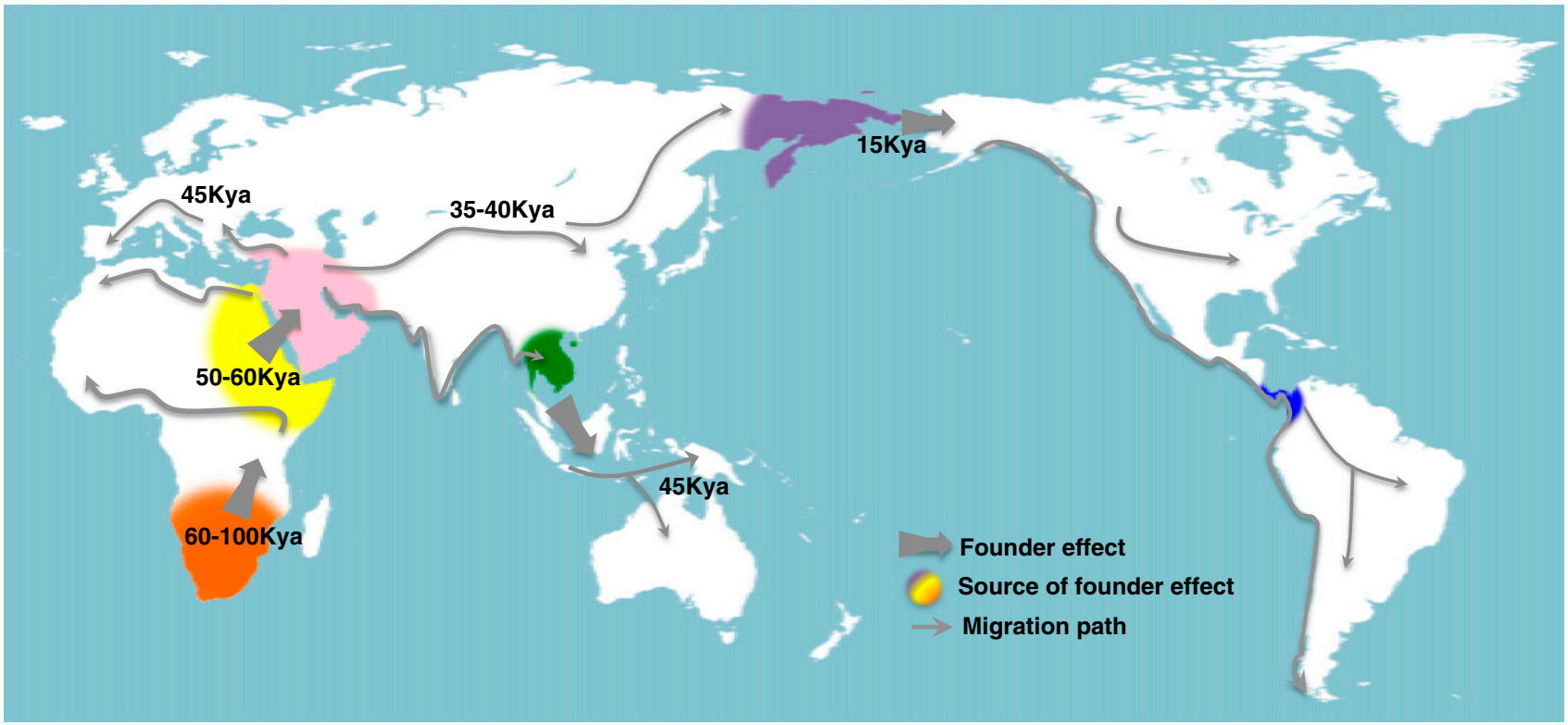
- What if the population size is changing or has a more complicated structure?
- How to extend these models to more than one genomic position/locus?

What might you do with all this theory?

- Generative model for sequence data that captures the most important biological mechanisms
- Infer biological parameters of the population: mutation rates, genome-wide recombination maps
- Infer demographic structure of the population
- Regions of the genome under selective pressure

Demographic structure

Demography – example



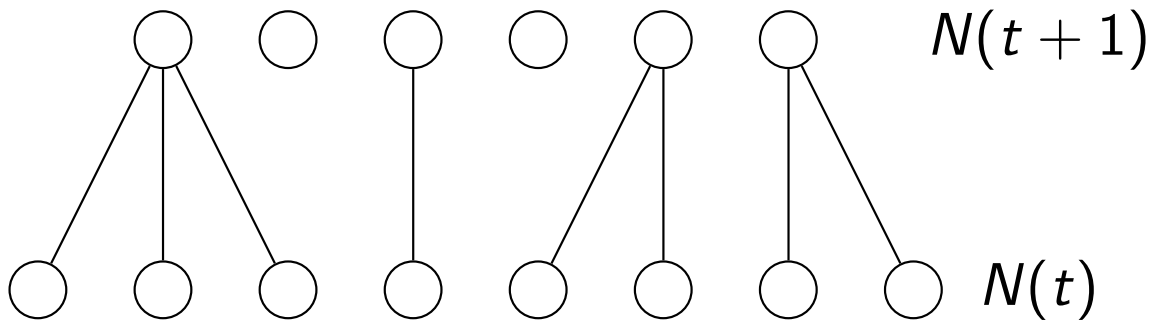
(Henn *et al.*, 2012)

Importance of modelling demography

- Demographic processes influence genetic variation
- Population stratification can confound association studies
- Correct null model
- Forensics
- Historical interest

Wright-Fisher model with variable population size

- Randomly mating population of size $N(t)$ in generation t
- $t = 0$ is present, t increasing in the past



- At generation t, $N(t)$ offspring generated, each picks a parent among $N(t+1)$ parents independently and uniformly

Coalescent with variable population size

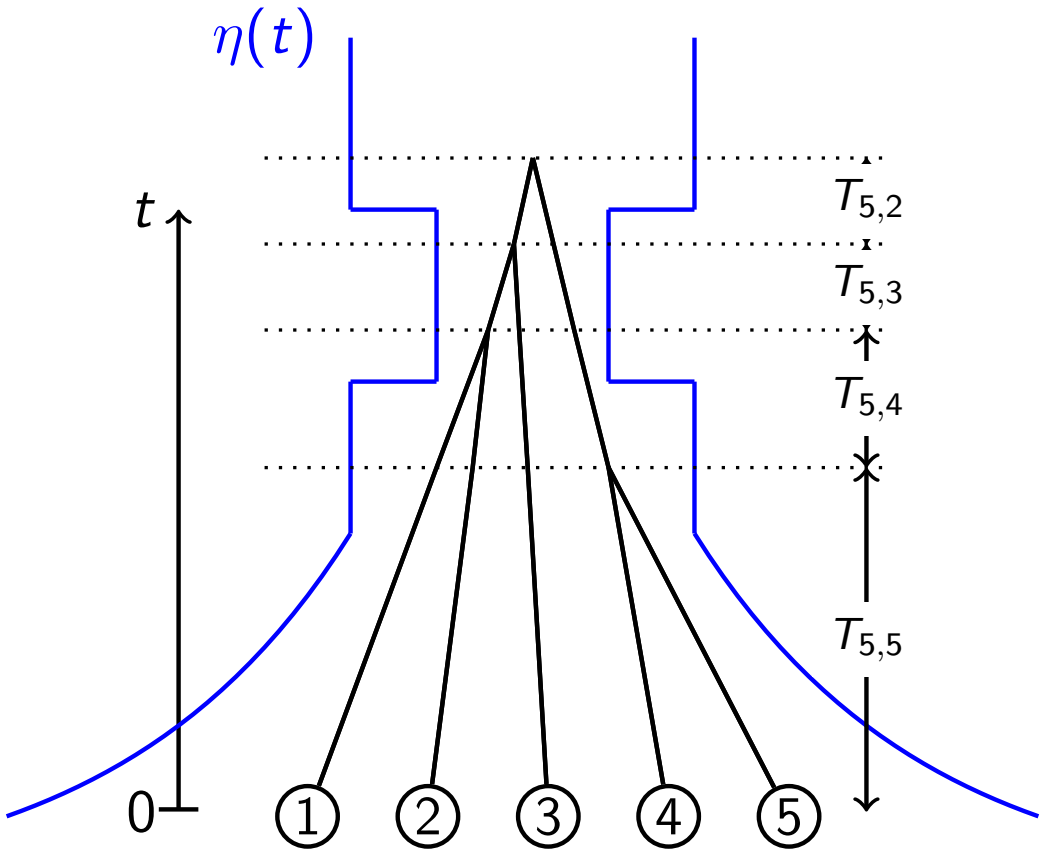
- Take $N(t)$ to ∞ at same rate for each t
- Choose rescaling parameter $\mathcal{N}$ for time such that

$$\eta(t) = \lim_{\substack{\mathcal{N}\to\infty, \\ N(t)\to\infty}} \frac{N(\lceil \mathcal{N} t \rceil)}{\mathcal{N}}$$

  exists and is positive for all $t \geq 0$
- Sample of size n randomly drawn at time 0
- Each pair of lineages coalescences according to exponential distribution with time-variable rate $\frac{1}{\eta(t)}$

Coalescent with variable population size

Let $T_{n,k}$ be the waiting time while there are k ancestral lineages for a sample of size n drawn at time 0.



Coalescent with variable population size

- CDF of $T_{n,n}$,

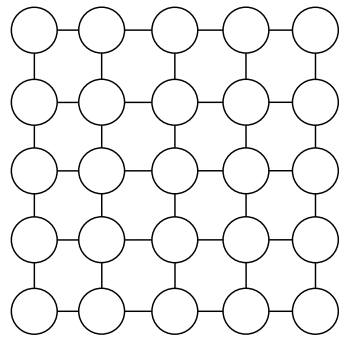
$$\mathbb{P}(T_{n,n} \leq t) = \int_0^t \frac{\binom{n}{2}}{\eta(\tau)} \exp\left(-\int_0^\tau \frac{\binom{n}{2}}{\eta(x)} dx\right) d\tau$$

- $\{T_{n,j}\}_{j=2}^n$ are independent random variables when $\eta(t) \equiv$ constant, but this is not true in general
- CDF of $T_{n,k}$ has a more complicated form for general k due to the varying population size
- But the CDF of $T_{n,k}$ conditional on $\sum_{j=k+1}^n T_{n,j}$ is simply,

$$\mathbb{P}(T_{n,k} \leq t \mid \sum_{j=k+1}^{n} T_{n,j} = t') = \int_{t'}^{t'+t} \frac{\binom{k}{2}}{\eta(\tau)} \exp\left(-\int_{t'}^\tau \frac{\binom{k}{2}}{\eta(x)} dx\right) d\tau$$

Wright-Fisher model with structure

- g subpopulations (aka demes), with Wright-Fisher random mating in each deme



- Population size N_α in deme α
- Occasionally demes exchange individuals (migrations)
- Per-generation probability an offspring in deme α has parent in deme β given by $c_{\alpha\beta}$

Structured coalescent

- Take N_α to ∞ at same rate for each α
- Let $\mathcal{N} = \sum_\alpha N_\alpha$ be the rescaling parameter for time
- Sample $\mathbf{n} = (n_\alpha)_\alpha$ drawn at time 0
- Each pair of lineages in deme α coalesces at rate $\frac{1}{N_\alpha/\mathcal{N}}$
- Migration of lineages from deme α to β at rate $\frac{m_{\alpha\beta}}{2}$, where $m_{\alpha\beta} = 2\mathcal{N}c_{\alpha\beta}$

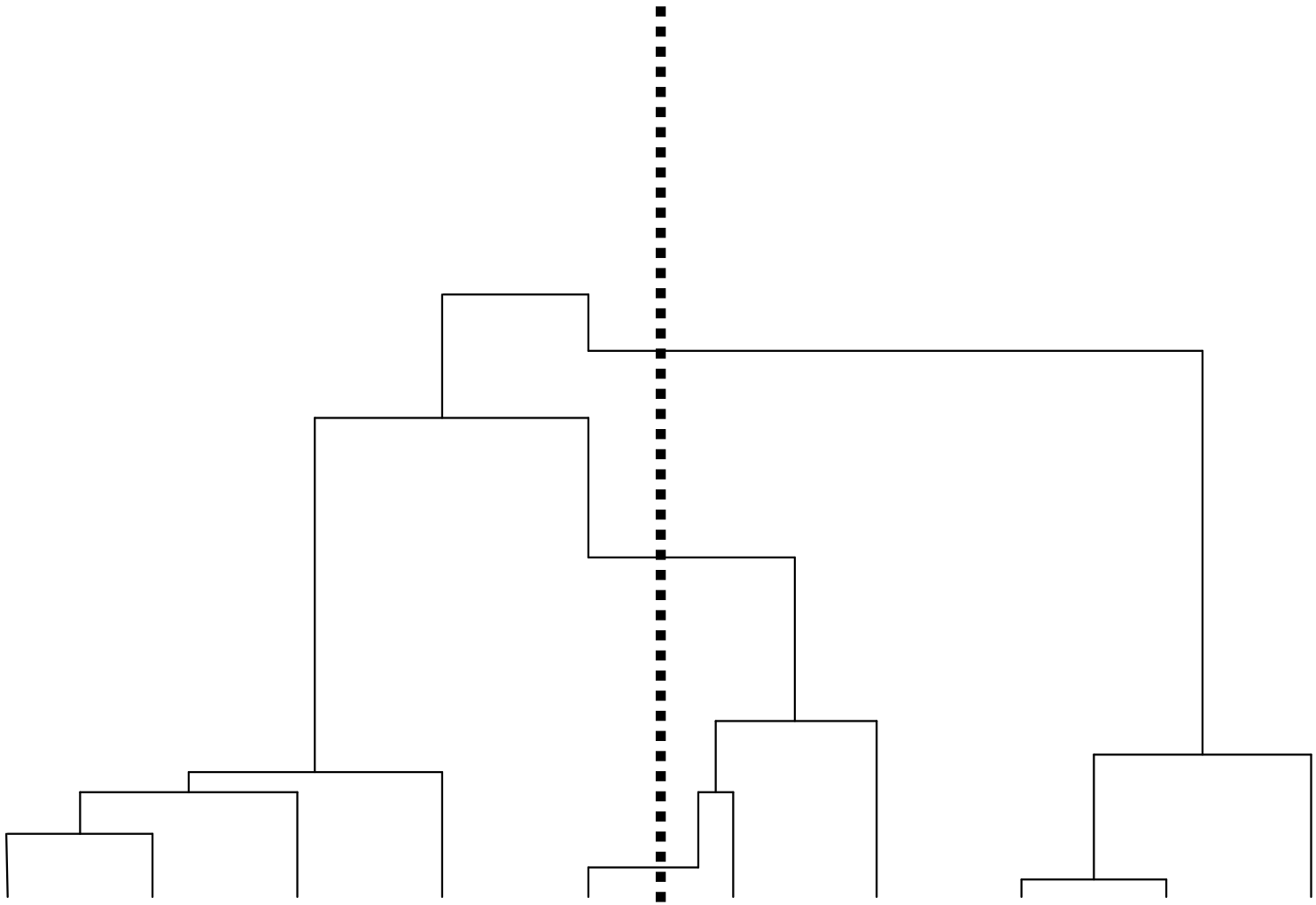
Structured coalescent

- Suppose we have two demes labelled α and β
- State of the Markov chain at time t given by $\mathbf{n}(t) = (n_\alpha(t), n_\beta(t))$
- Transitions out of state $\mathbf{n} = (n_\alpha, n_\beta)$

$$(n_\alpha, n_\beta) \to \begin{cases} (n_\alpha - 1, n_\beta) & \text{at rate } \binom{n_\alpha}{2} \frac{1}{N_\alpha/\mathcal{N}} \\ (n_\alpha, n_\beta - 1) & \text{at rate } \binom{n_\beta}{2} \frac{1}{N_\beta/\mathcal{N}} \\ (n_\alpha - 1, n_\beta + 1) & \text{at rate } \frac{n_\alpha m_{\alpha\beta}}{2} \\ (n_\alpha + 1, n_\beta - 1) & \text{at rate } \frac{n_\beta m_{\beta\alpha}}{2} \end{cases}$$

Structured coalescent

Example genealogy for 2 demes and sample $\mathbf{n} = (5, 5)$ at time 0



Measure of population structure

- $p_w(\theta) =$ probability that two individuals sampled from the same deme are IBD
- $p_b(\theta) =$ probability that two individuals sampled from different demes are IBD

$$p_w = \mathbb{E}[e^{-\theta T_w}], \quad p_b = \mathbb{E}[e^{-\theta T_b}]$$

where T_w (resp. T_b) are the time to coalescence for two individuals sampled from the same (resp. different) deme

- Structure in the population summarized by F_{ST}, defined as

$$F_{ST} = \frac{p_w(\theta) - \overline{p}(\theta)}{1 - \overline{p}(\theta)},$$

where $\overline{p}(\theta)$ is the probability of IBD when two individuals are sampled at random (across all demes)

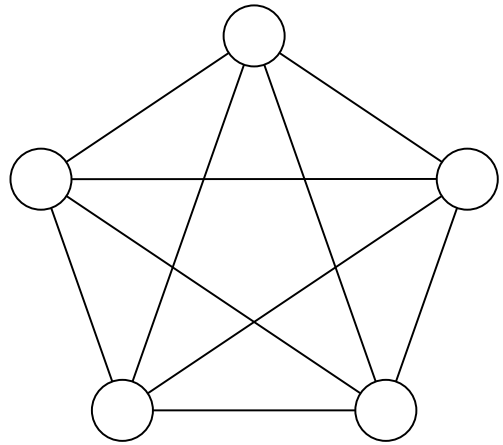
Measure of population structure

- $0 \leq F_{ST} \leq 1$, $F_{ST} = 0$ when no substructure, $F_{ST} = 1$ when populations isolated
- For small θ,

$$F_{ST} \approx 1 - \frac{\mathbb{E}[T_w]}{\mathbb{E}[T]},$$

  where T is the time to coalescence for two individuals sampled at random (across all demes)
- For human population, some studies estimate $F_{ST} = 0.12$.

Simple example – symmetric island model

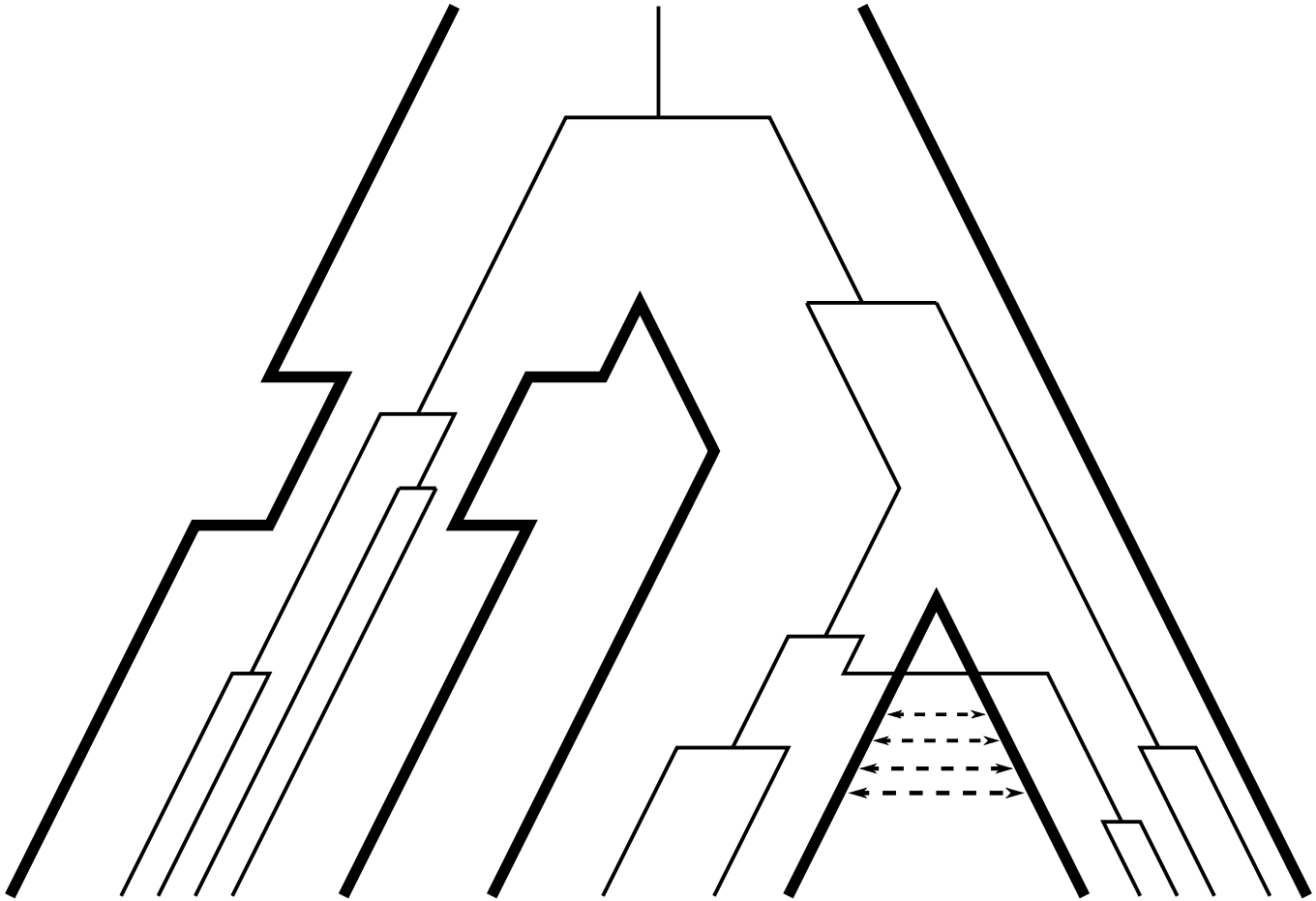


- Suppose we have g demes of equal size
- Migration from any deme α to β is given by $m_{\alpha\beta} = \frac{m}{g-1}$
- Let $\theta/2$ be the mutation rate
- Conditioning on the most recent genealogical event, can write recurrences for $p_w(\theta)$ and $p_b(\theta)$

$$p_w(\theta) = \frac{1}{1+\theta+m} + \frac{m}{1+\theta+m} p_b(\theta)$$

$$p_b(\theta) = \frac{m/(g-1)}{\theta+m} p_w(\theta) + \frac{m(g-2)/(g-1)}{\theta+m} p_b(\theta).$$

General population structure

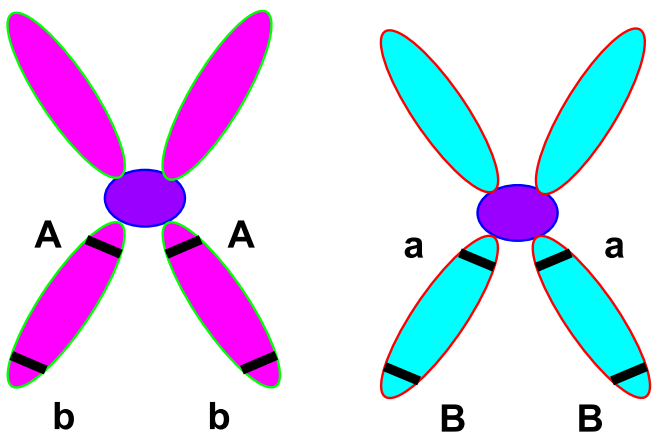


Recombination

Recombination

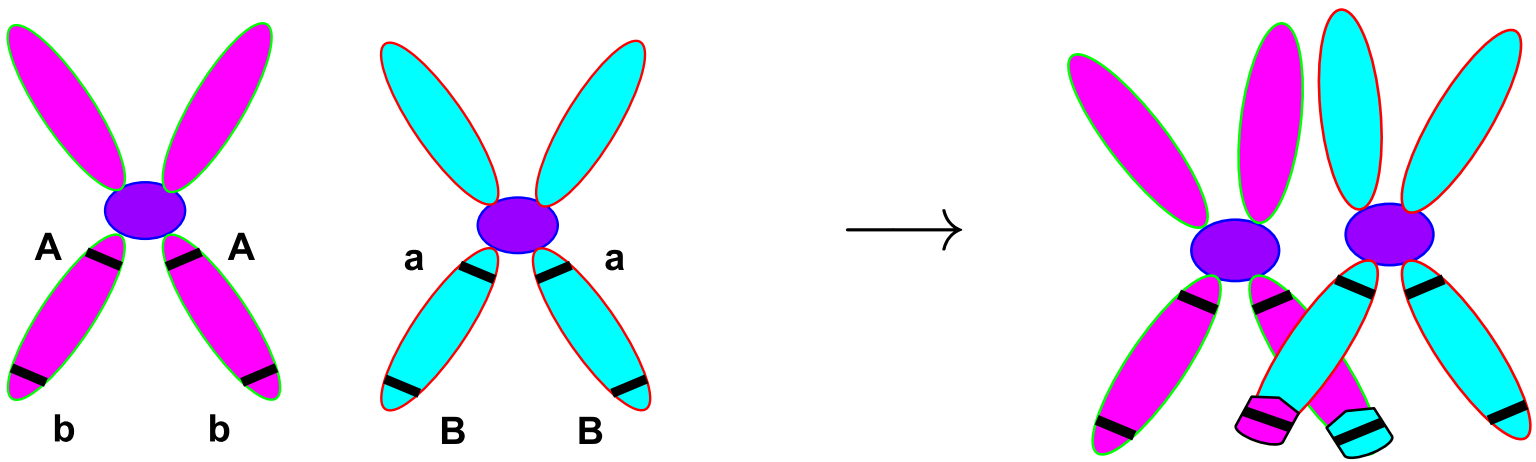
Recombination is a major evolutionary mechanism responsible for generating genetic variation in sexual organisms

- Humans are *diploid* organisms
- We have two copies of every chromosome — a maternal and a paternal copy. These are called *homologous* chromosomes
- In the synthesis phase of meiosis, each chromosome gets duplicated so that it is comprised of two identical sister chromatids joined at the centromere



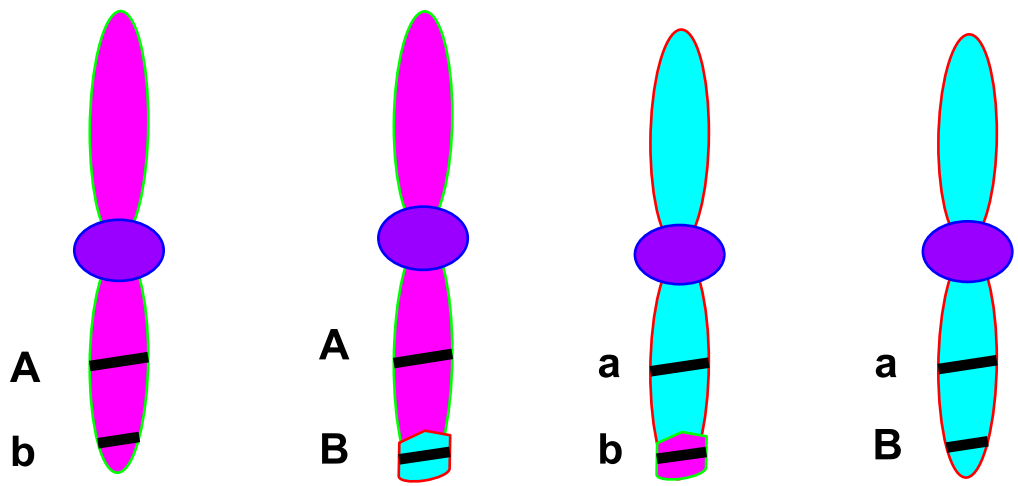
Recombination

During the subsequent prophase I stage of meiosis, homologous chromosomes come into contact and DNA is exchanged between chromatids on homologous chromosomes.



Recombination

At the end of meiosis, 4 *haploid* daughter cells are produced. Some of these daughter cells have *different* haplotypes from either of the parental haplotypes.

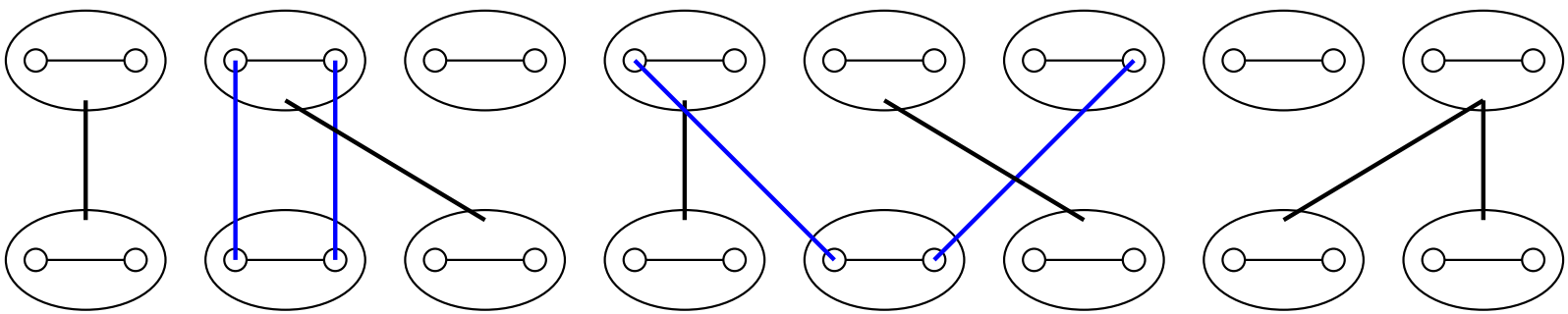


Importance of recombination

- Creates new genetic variation by mixing alleles between different haplotypes
- Breaks down genealogical correlation between two positions on the same chromosome
- Implications for many computational problems in population genetics, including
  - Phasing genotype data into haplotype data
  - Imputing missing data
  - Disease-association mapping
  - Inferring local ancestry of admixed populations
  - Detecting signatures of natural selection

Wright-Fisher model with recombination

- Consider a population of N individuals at two loci
- For each offspring individual, with probability r, there is recombination between the loci and a parent is chosen for each locus independently and uniformly at random
- With probability $1 - r$, an individual chooses the same parent uniformly at random for both loci

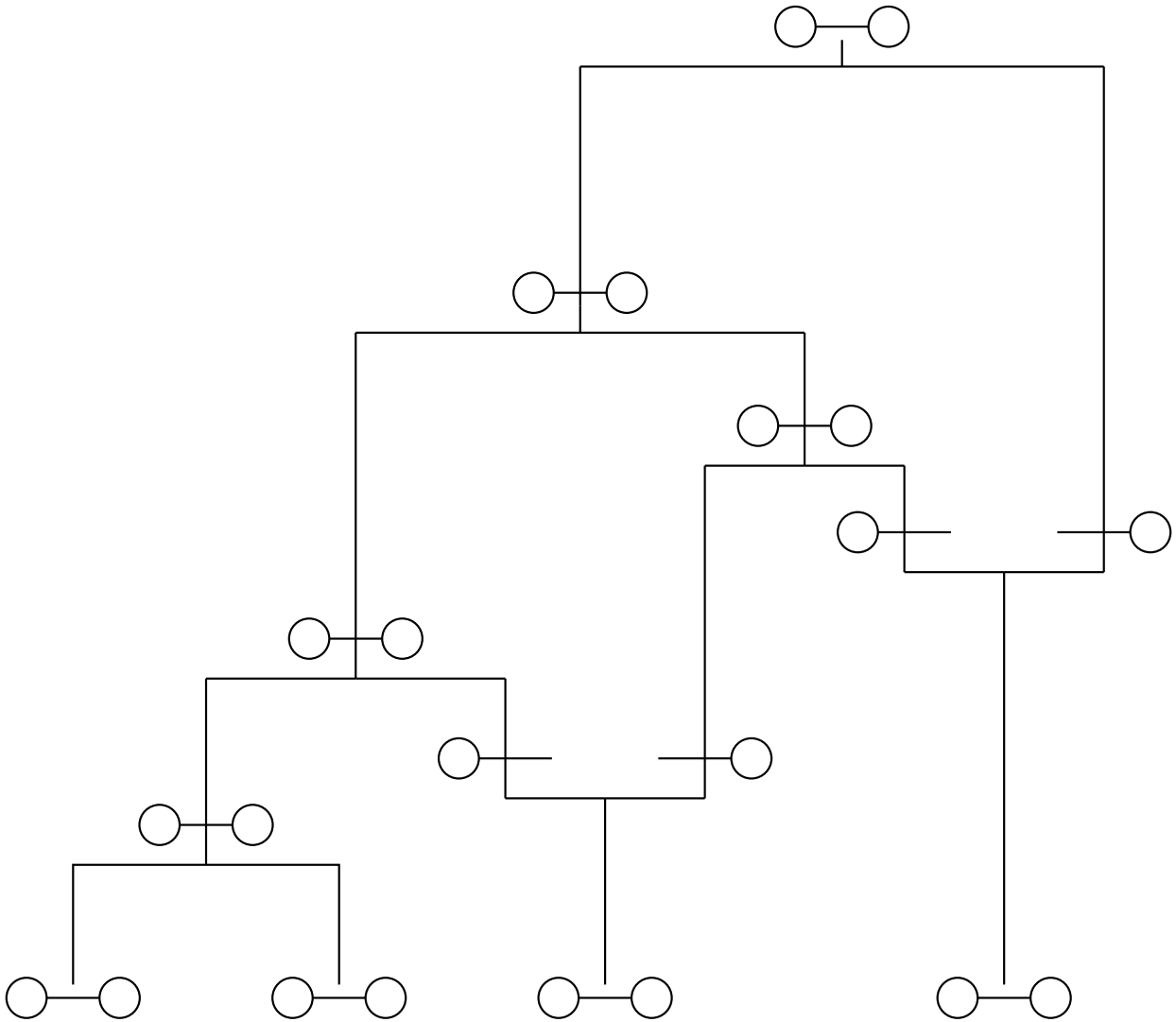


Coalescent with recombination – desiderata

- Suppose we have two loci A and B separated by recombination rate $\rho/2$
- We want a continuous-time process to model the genealogical structures at both loci
- Marginally, the genealogy at each locus must be given by the coalescent process we saw earlier
- If there was no recombination ($\rho = 0$), genealogies at both loci should be identical
- If there was free recombination ($\rho = \infty$), genealogies at both loci should be independent
- In general, the genealogies at both loci will be correlated due to non-trivial recombination

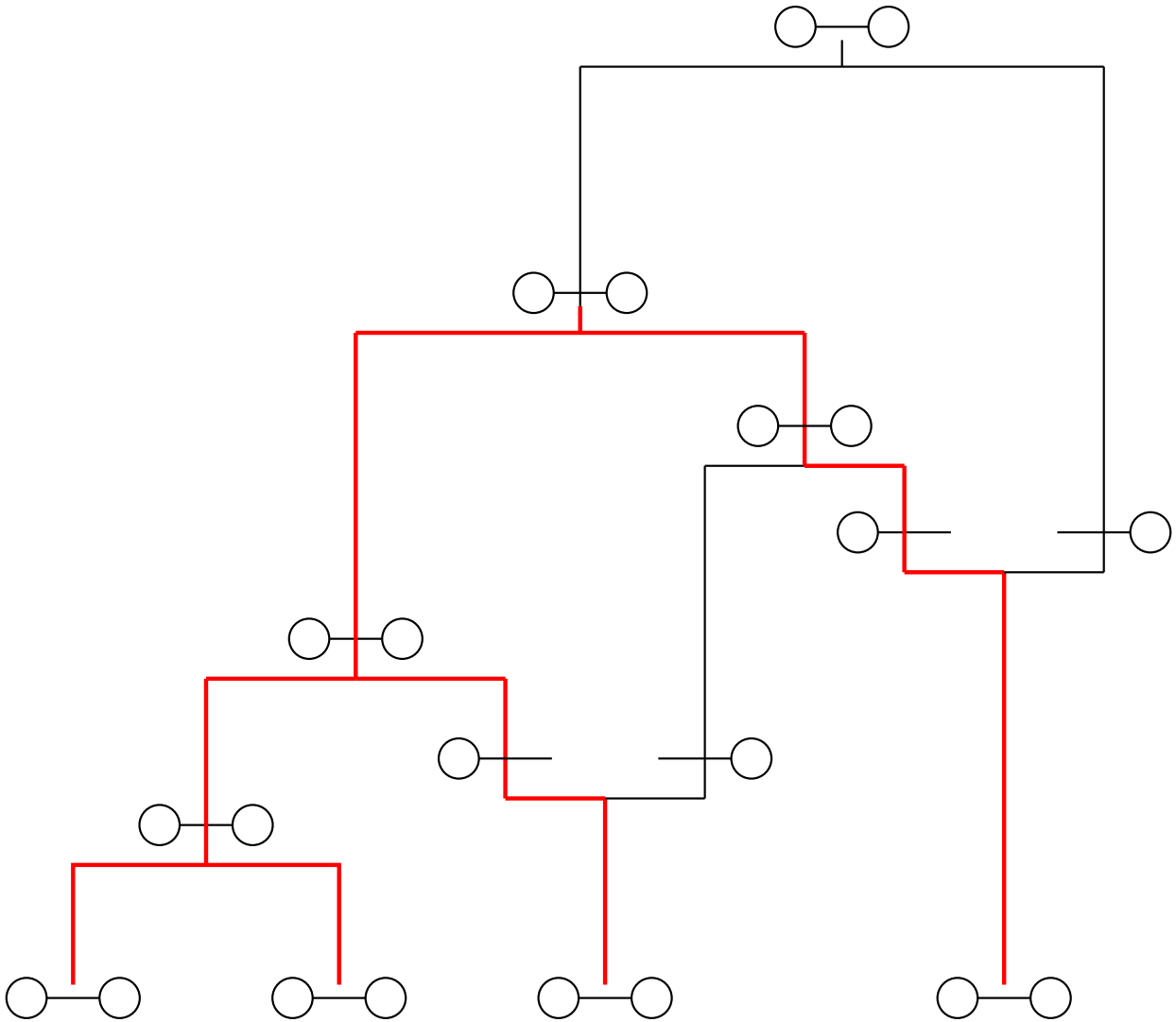
Ancestral recombination graph

Sample of size 4 at two loci. Example joint genealogy at loci A and B



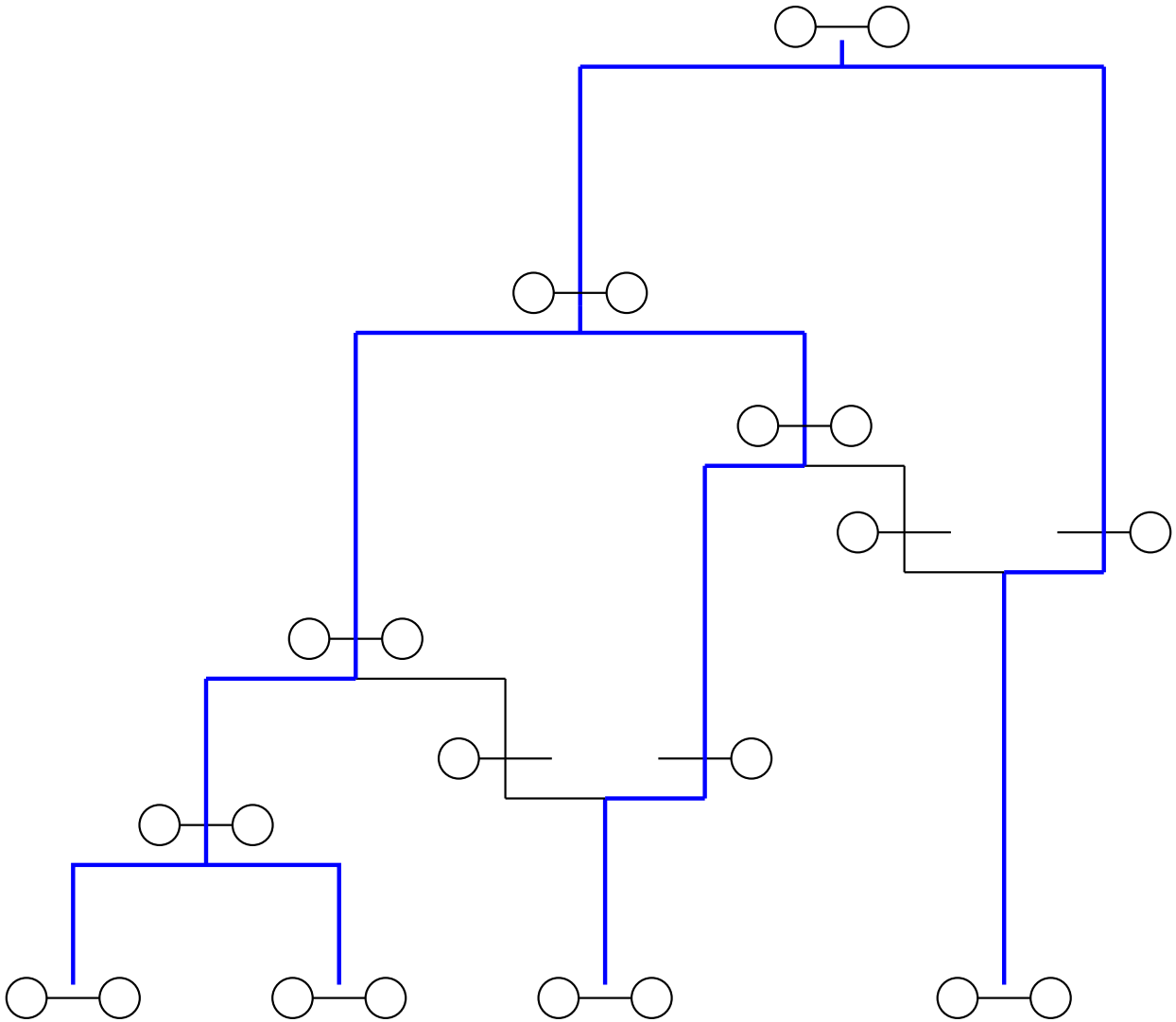
Ancestral recombination graph

Sample of size 4 at two loci. Marginal genealogy at locus A



Ancestral recombination graph

Sample of size 4 at two loci. Marginal genealogy at locus B

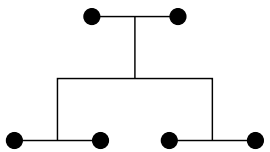


Coalescent with recombination

- r per-generation per-individual probability of recombination between A and B
- Population-scaled recombination rate $\rho = 2Nr$
- The configuration at time t in the ancestral process Markov chain will be specified by $\mathbf{n}(t) = (a(t), b(t), c(t), d(t))$
- $a(t)$ (resp. $b(t)$) is the number of ancestors at time t that contribute genetic material to the original sample at locus A (resp. locus B) *only*
- $c(t)$ is the number of ancestors at time t that contribute genetic material to the original sample at *both locus A and locus B*
- $d(t)$ is the number of ancestors at time t that *do not contribute any genetic material* to the original sample

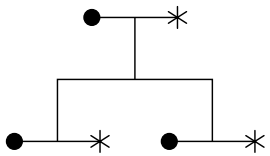
Coalescent with recombination

Suppose the current sample configuration is (a, b, c, d)
There are 10 kinds of coalescence events:



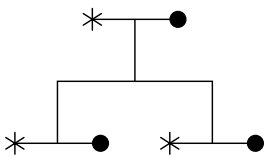
$(a, b, c, d) \to (a, b, c-1, d)$

Rate: $\binom{c}{2}$



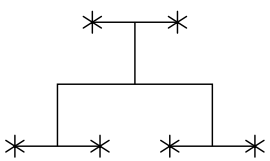
$(a, b, c, d) \to (a-1, b, c, d)$

Rate: $\binom{a}{2}$



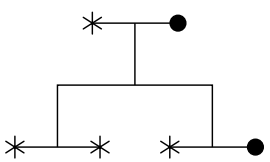
$(a, b, c, d) \to (a, b-1, c, d)$

Rate: $\binom{b}{2}$



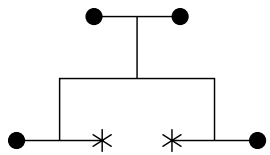
$(a, b, c, d) \to (a, b, c, d-1)$

Rate: $\binom{d}{2}$



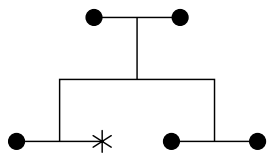
$(a, b, c, d) \to (a, b, c, d-1)$

Rate: bd



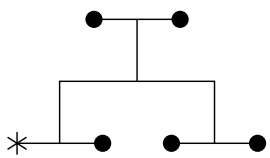
$(a, b, c, d) \to (a-1, b-1, c+1, d)$

Rate: ab



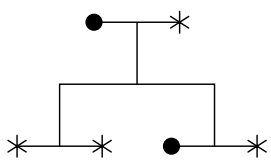
$(a, b, c, d) \to (a-1, b, c, d)$

Rate: ac



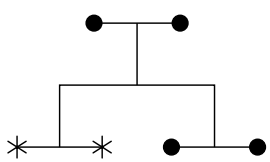
$(a, b, c, d) \to (a, b-1, c, d)$

Rate: bc



$(a, b, c, d) \to (a, b, c, d-1)$

Rate: ad

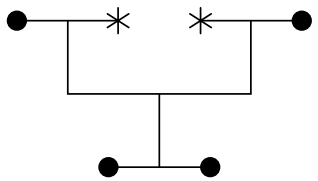


$(a, b, c, d) \to (a, b, c, d-1)$

Rate: cd

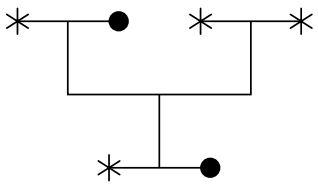
Coalescent with recombination

There are 4 kinds of recombination events:



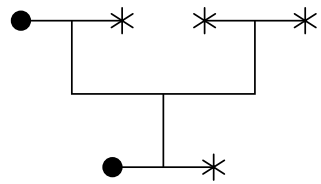
$(a, b, c, d) \to (a+1, b+1, c-1, d)$

Rate: $\frac{c\rho}{2}$



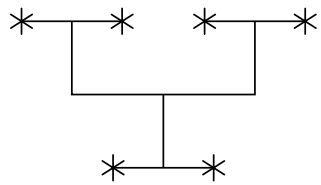
$(a, b, c, d) \to (a, b, c, d+1)$

Rate: $\frac{b\rho}{2}$



$(a, b, c, d) \to (a, b, c, d+1)$

Rate: $\frac{a\rho}{2}$



$(a, b, c, d) \to (a, b, c, d+1)$

Rate: $\frac{c\rho}{2}$

Coalescent with recombination

- Summarizing the transitions out of state (a, b, c, d)

$$(a,b,c,d) \to \begin{cases} (a,b,c-1,d) & \text{at rate } \binom{c}{2} \\ (a-1,b-1,c+1,d) & \text{at rate } ab \\ (a-1,b,c,d) & \text{at rate } \binom{a}{2} + ac \\ (a,b-1,c,d) & \text{at rate } \binom{b}{2} + bc \\ (a,b,c,d-1) & \text{at rate } (a+b+c)d + \binom{d}{2} \\ (a+1,b+1,c-1,d) & \text{at rate } \frac{c\rho}{2} \\ (a,b,c,d+1) & \text{at rate } \frac{(a+b+d)\rho}{2} \end{cases}$$

- Absorbing states $\{(0,0,1,d) \mid d \geq 0\}$
- Can impose mutations on the genealogy at each locus as usual

Coalescent with recombination – reduced representation

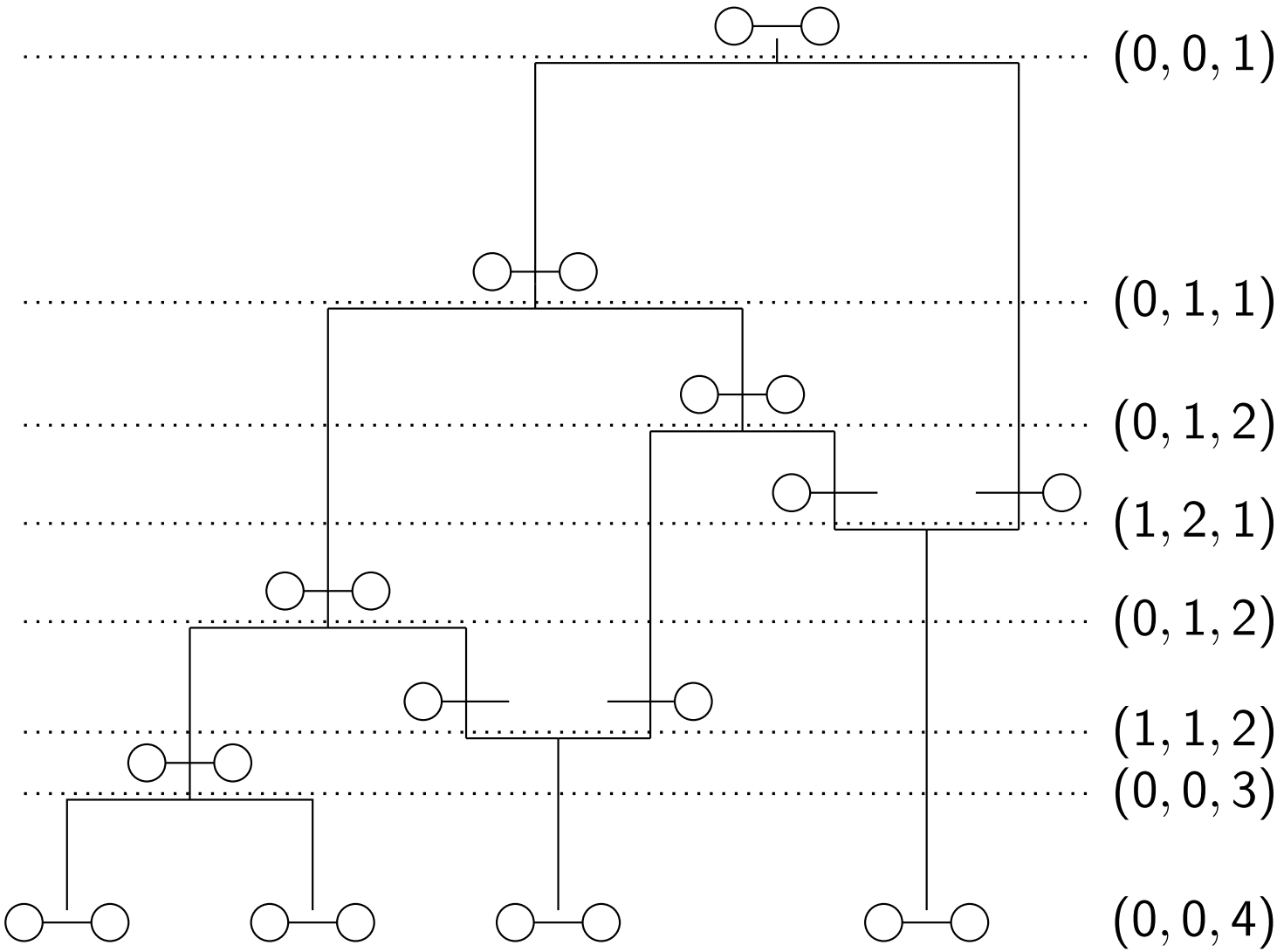
- The fourth component of the state representation (the 'd' component) need not be tracked since these ancestors do not contribute any genetic material to the original sample, and hence have no influence
- Let the reduced state be (a, b, c). The transitions out of this state are

$$(a,b,c) \to \begin{cases} (a,b,c-1) & \text{at rate } \binom{c}{2} \\ (a-1,b-1,c+1) & \text{at rate } ab \\ (a-1,b,c) & \text{at rate } \binom{a}{2} + ac \\ (a,b-1,c) & \text{at rate } \binom{b}{2} + bc \\ (a+1,b+1,c-1) & \text{at rate } \frac{c\rho}{2} \end{cases}$$

- Absorbing state is $(0, 0, 1)$

Ancestral recombination graph

Example joint genealogy at loci A and B



Coalescent with recombination

- The coalescent process at just locus A is embedded in the coalescent with recombination for loci A and B
- For state (a, b, c), there are $a+c$ ancestors that contribute genetic material at locus A
- Transitions out of states in $S_m = \{(a, b, c) \mid a + c = m\}$ are

$$(a, b, c) \to \begin{cases} (a, b, c-1) & \text{at rate } \binom{c}{2} \\ (a-1, b, c) & \text{at rate } \binom{a}{2} + ac \end{cases}$$

- Total rate of transitions out of S_m is $\binom{a}{2} + \binom{c}{2} + ac = \binom{m}{2}$, which agrees with the transition rates of the coalescent process at locus A

Computational challenges due to recombination

- The two-locus model can be extended to multiple loci with different recombination rates between them
- For K loci, the state space of the Markov chain would have $2^K - 1$ components for the number of ancestors that contribute genetic material to the sample at different subsets of the K loci
- State space size $O((n + 2^K)^{(2^K - 1)})$
- Even simulating data under these models can be expensive for large K and large ρ
- Computing probability of an observed sample under this model is prohibitive in practice even for $K = 2$, two alleles per locus and a sample size of a few hundred haplotypes

What about natural selection?

- Easy to incorporate an allelic advantage in the discrete Wright-Fisher model
- Can construct a continuous time coalescent model
- Genealogical structure – ancestral selection graph
- However, easier to develop a forwards-in-time continuum model (coming up)