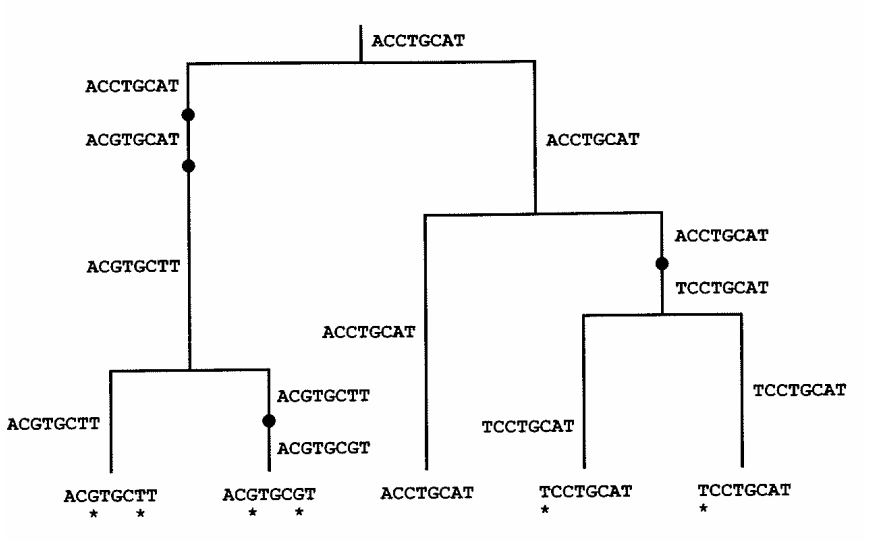
An introduction to coalescent theory

Nicolas Lartillot

May 29, 2012

Inferring population history from haplotype data



Hein, Shierup and Wiuf, 2005

- a set of *n* haplotypes randomly sampled from a population
- sequences of length *L*, known mutation rate μ
- what can we say about
  - population size (*N*) and structure?
  - demographic history?
  - selection?

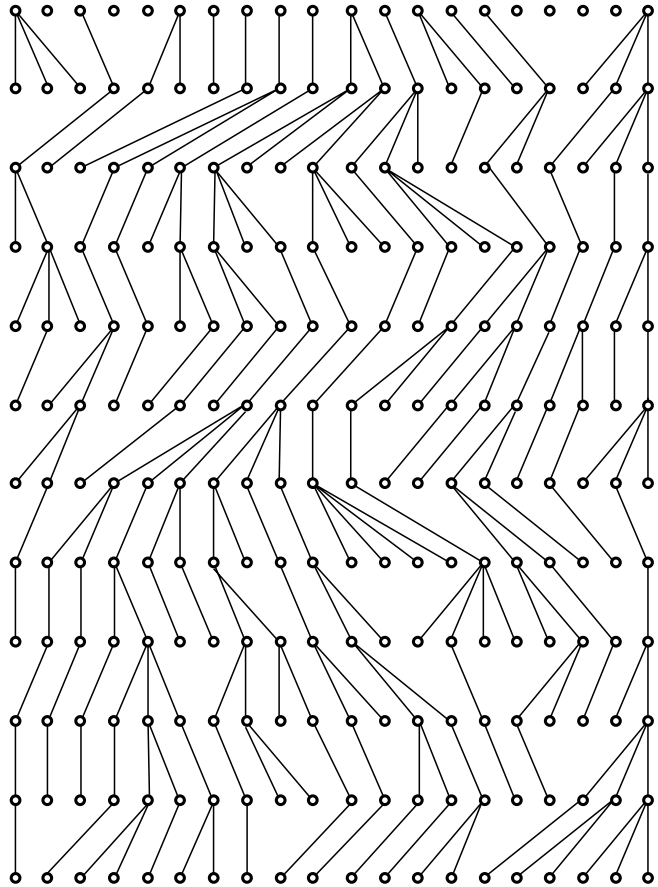
Approach

- define a model of demography and reproduction (Wright-Fisher)
- induces a law on gene genealogies (Kingman's coalescent)
- then define a model of DNA sequence mutations
- explain variation in gene sample based on combination of mutation and coalescent models.

Applications

- estimating parameters (population size, mutation rate)
- testing hypotheses (e.g. deviation from neutrality)
- building blocks for more sophisticated models (course no 2)

The Wright-Fisher model

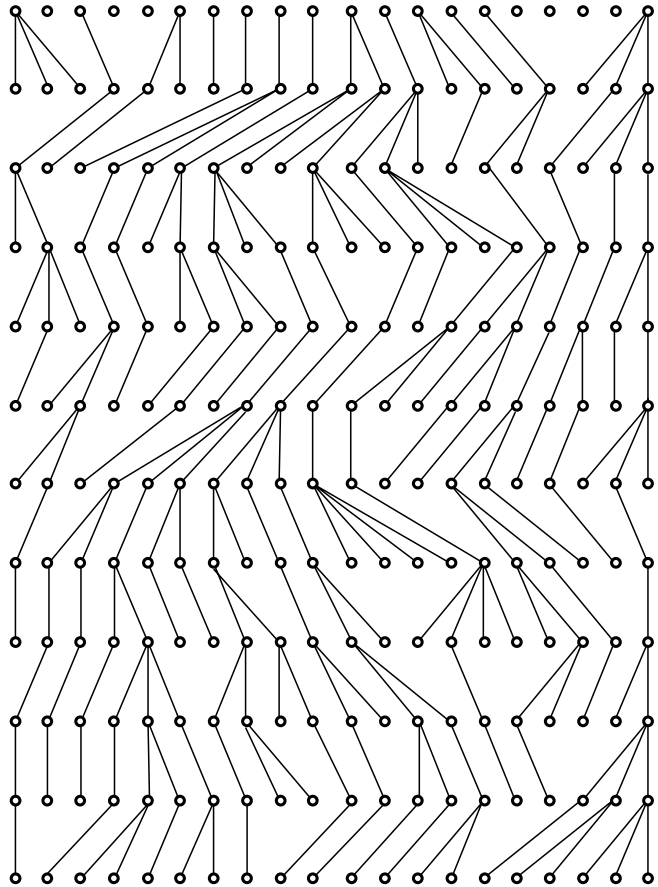


from Felsenstein

Assumptions

- panmictic population
- constant population size
- neutral

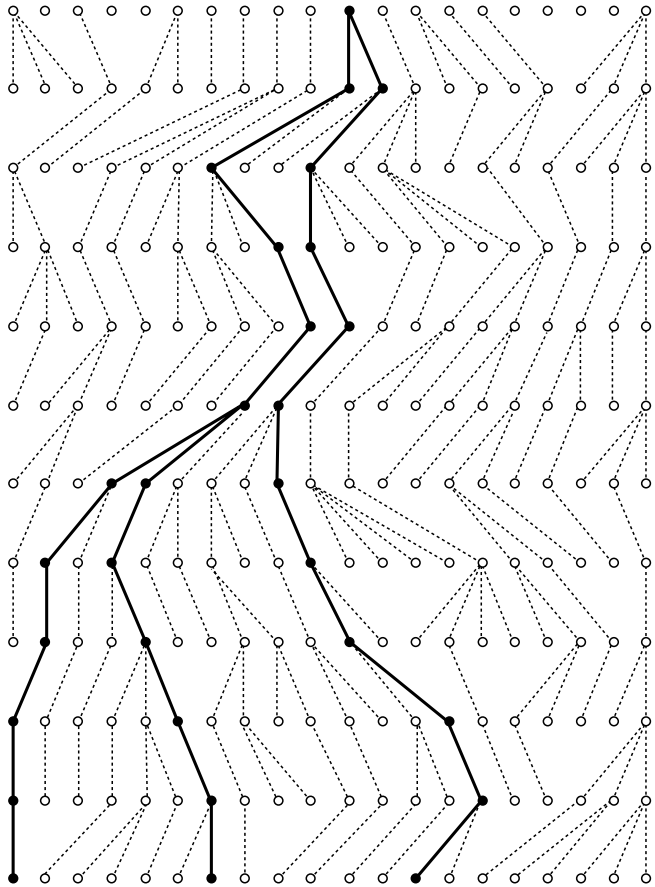
The Wright-Fisher model



from Felsenstein

- each offspring 'chooses' parent uniformly among $2N$ individuals of previous generation
- distribution of number of offspring: $Binomial(2N, 1/2N)$

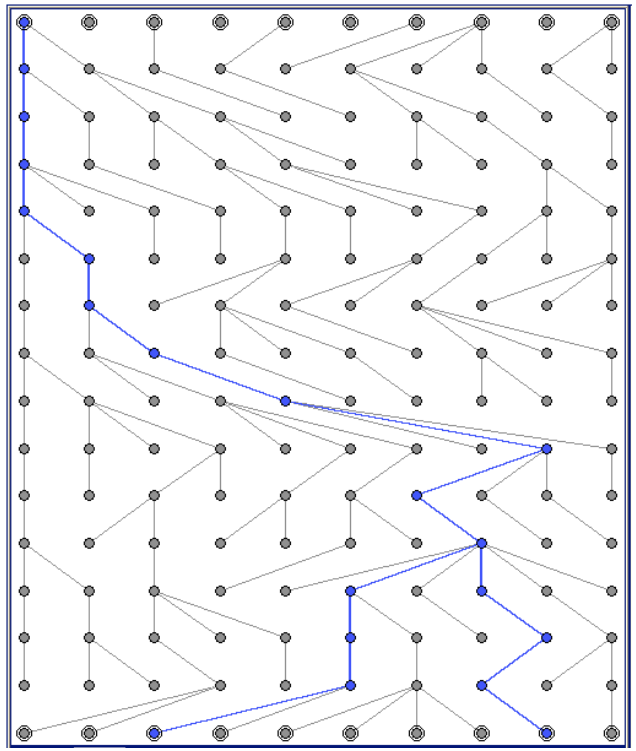
Genealogy of a sample



from Felsenstein

- n individuals taken at random (here $n = 3$)
- age of their ancestor?
- typical shape of the genealogy?

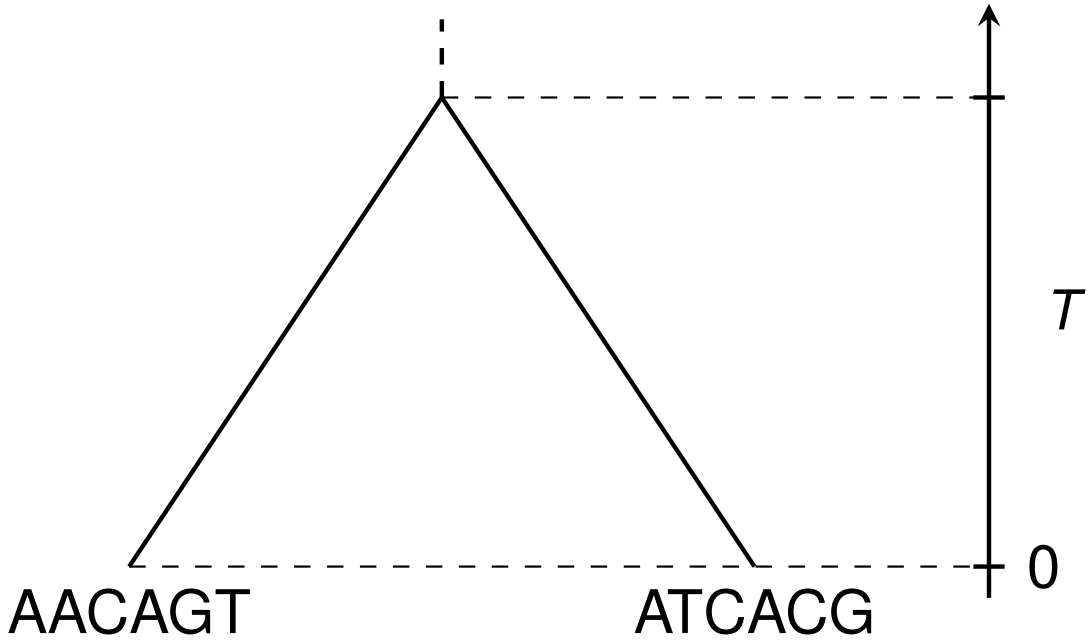
coalescence of $n = 2$ genes



www.coalescent.dk

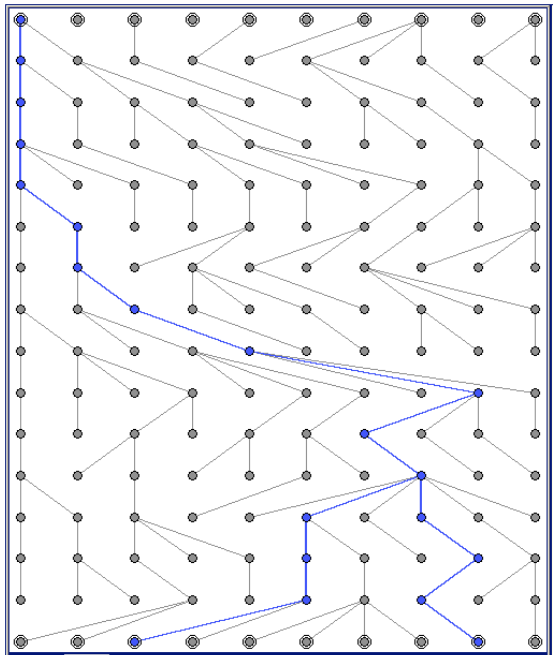
- prob. of coalescence in previous generation $1/(2N)$
- average coalescence time for 2 individuals: $\overline{T} = 2N$.

Relation between genetic diversity and coalescence time ($n = 2$)



- time since last common ancestor: T generations
- sequences of length L, known mutation rate μ
- mean fraction of sites differing between 2 individuals: $\pi = 2\mu T$.

coalescence of $n = 2$ genes



www.coalescent.dk

with mutation

- mutations at rate μ per base pair per generation
- average diversity: $\pi = 2\overline{T}\mu = 2.2N.\mu = 4N\mu = \theta$.
- θ: *scaled mutation rate* (*N* and μ are *confounded*)
- yields an estimate of *N* if μ is known and π is observed

Tajima's estimator

$n = 4$ observed DNA sequences

1	A	C	C	A	C	A	A	G
2	A	C	C	A	G	T	A	G
3	A	C	T	G	C	A	T	G
4	A	C	T	G	G	T	A	C

π_{ij}: fraction of polymorphic sites between haplotypes i and j

$$\hat{\pi} = \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij}$$

Effective population size of humans

Human-chimp divergence

- SND (single nucleotide differences): $\simeq 2\%$
- divergence time: $\simeq 6Ma$.
- thus, mutation rate: $\simeq 3.10^{-8}$

Human polymorphism

- heterozygosity: $\pi = 0.001$ (1 every 1000 bp)
- SNP (single nucleotide polymorphisms): 1 every 100 to 300 bp

$$\begin{aligned} \pi &= 4N\mu \\ N &= \pi/4/\mu \simeq 10\,000 \end{aligned}$$

- *effective* population size $<$ *census* population size

Effective population size

Genetic aspects

- autosomal: $2N$
- X chromosome: $3/2\,N$
- mitochondrial, Y chromosome: N

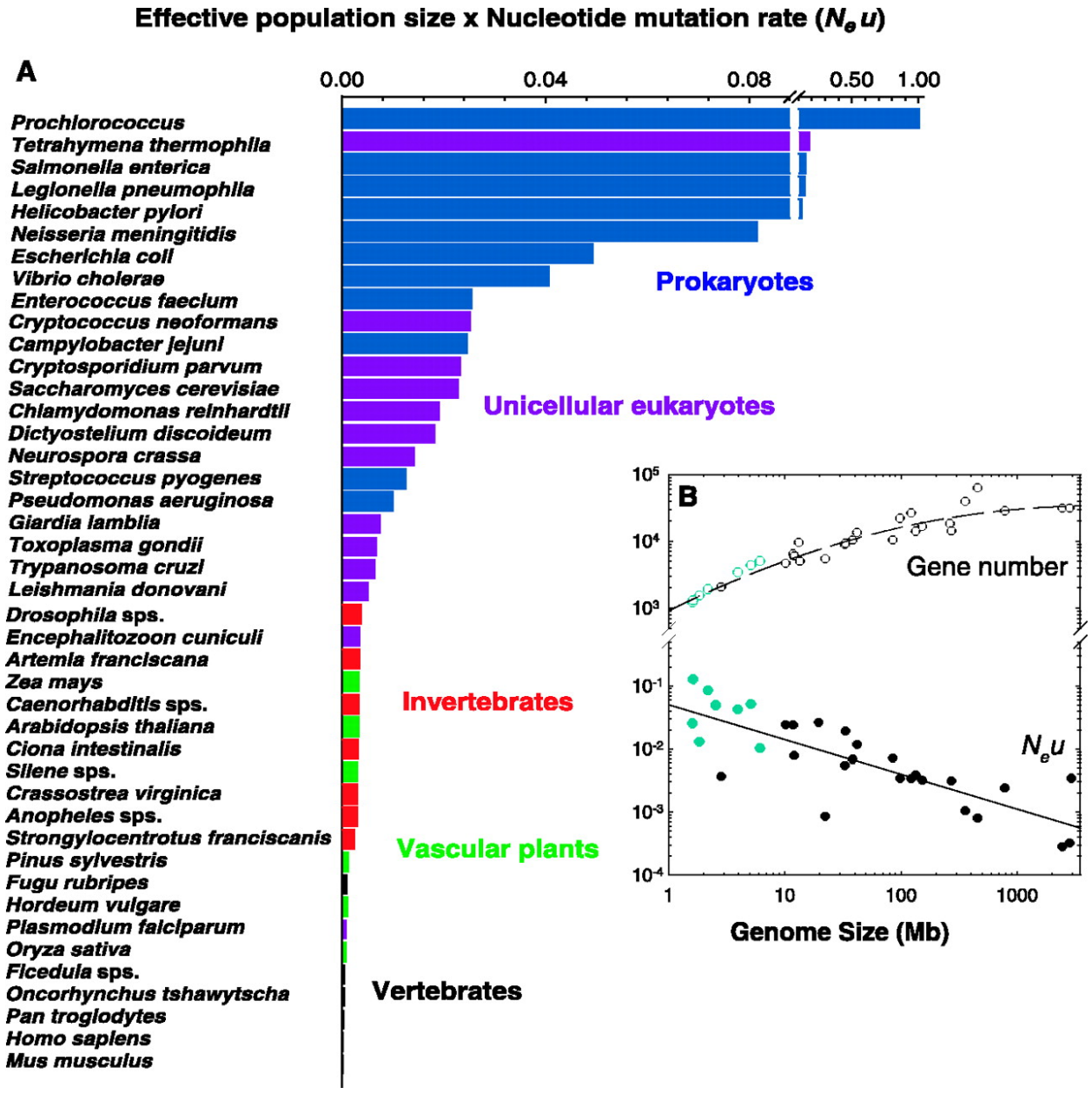
Demographic aspects

- N: harmonic mean of census size over short-term fluctuations
- frequent bottlenecks: low N
- reproductive variance (species with male dominance have low N)
- population structure (e.g. a parasite has N of its host)

Linkage and selection

- selection at linked loci reduce N at neutral loci
- purifying selection: background selection
- positive selection: selective sweeps

Nucleotide diversity across life forms



Lynch and Conery, Science 2003; 302:1401

Effective population sizes across life forms

Mutation rates (per generation)

- human : $\simeq 10^{-8}$
- fly, nematode: $\simeq 10^{-9}$
- unicellular eukaryotes and prokaryotes: $\simeq 10^{-10}$

Effective population sizes

- human, large vertebrates: 10^4
- small vertebrates: 10^5
- invertebrates, terrestrial plants: 10^6
- unicellular eukaryotes: 10^7
- prokaryotes: $> 10^8$

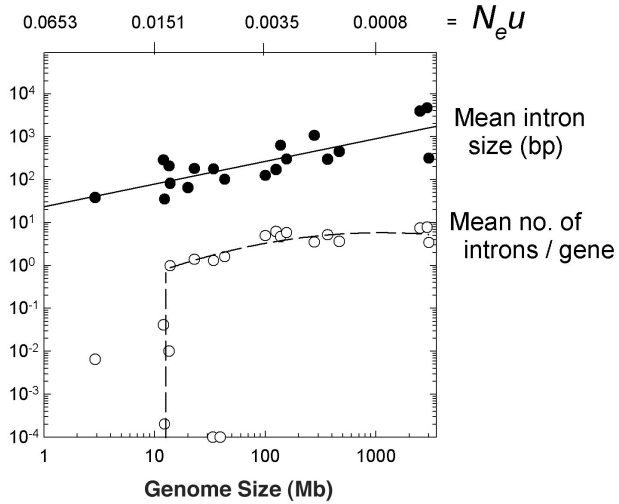
Population size and evolutionary genomics

Effective size and selection

- random drift proportional to $1/N$
- selection efficient only if $s >> 1/N$

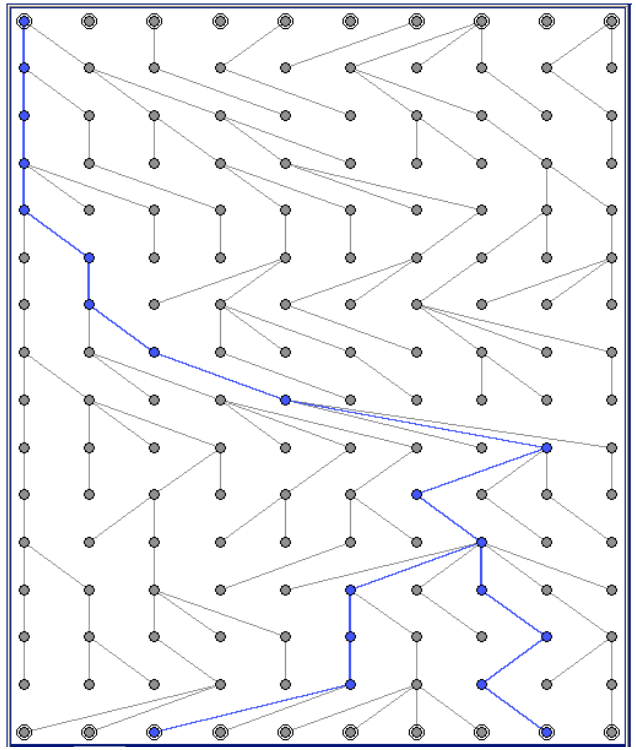
Evolutionary genomics

- small N: random drift dominates molecular evolution in humans
- many features selected in fly/yeast /E.coli not selected in humans
- genome structure influenced by population genetics parameters



Lynch and Conery, Science 2003; 302:1401

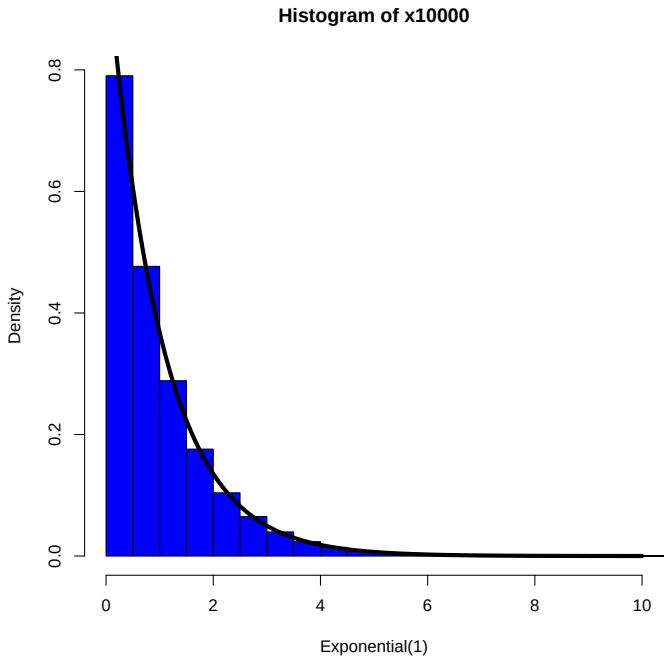
Distribution of age of ancestor



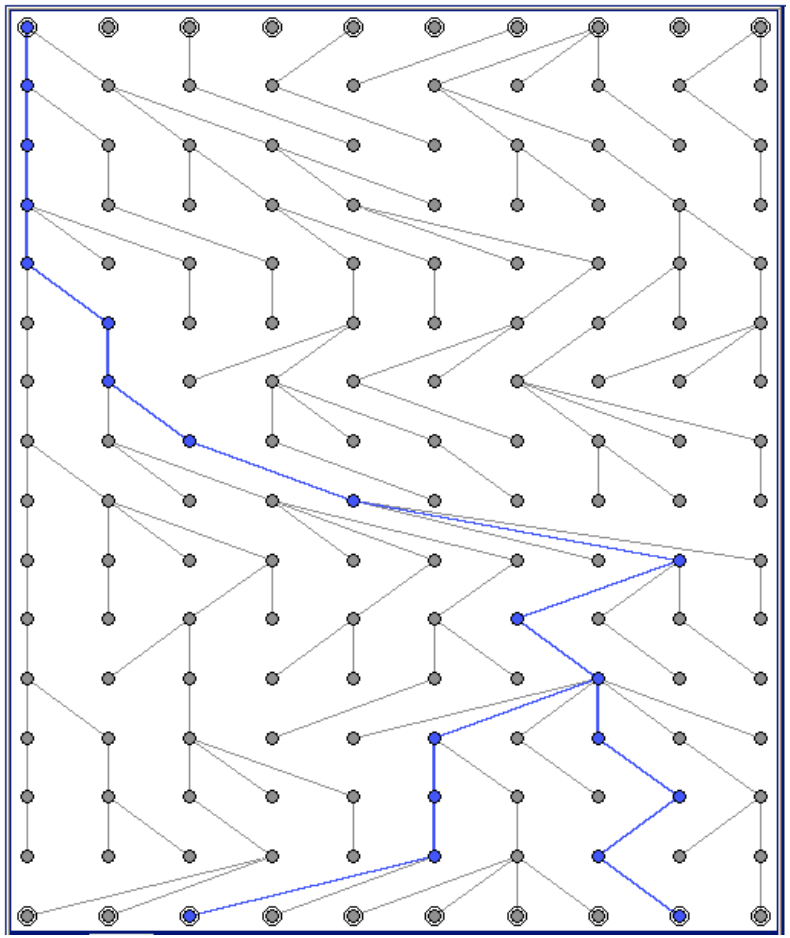
www.coalescent.dk

- prob. of coalescence in previous generation $1/(2N)$
- prob. of coalescence in 2 generations $(1 - 1/(2N))(1/(2N))$
- prob. of coalescence in t generations $(1 - 1/(2N))^{t-1}(1/(2N))$
- t has a geometric distribution

Exponential distribution



$$u = t/2N, \quad p(u) = e^{-u}$$

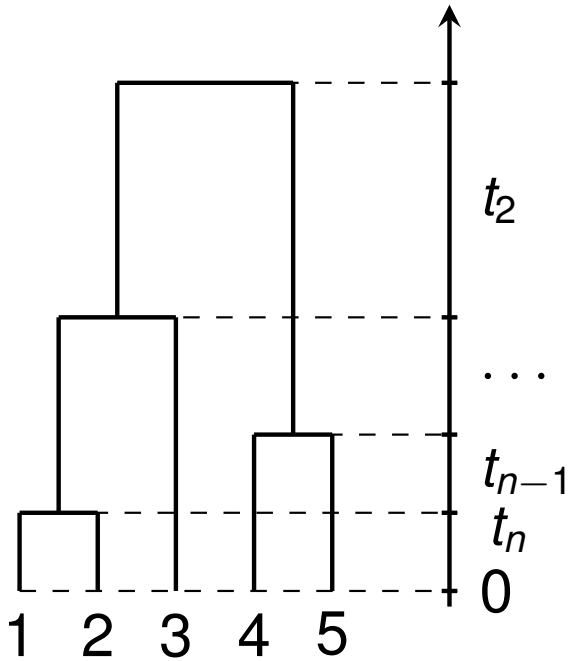


www.coalescent.dk

- age of ancestor of 2 individuals has geometric distribution
- for $n << N$, approx. an exponential distribution
- mean of t_2 is $2N$, (std dev of t_2 is $2N$)
- rescaling: $u_2 = t_2/(2N)$ has mean 1 and stdev 1

Generalization for $n > 2$

- make Wright-Fisher simulations (pop. size $2N$)
- for each simulation, take n chromosomes at final time (present)
- trace back their genealogy
- measure t_j (in generations) and set $u_j = t_j/2N$ (rescaling)
- distribution of t_j and u_j over simulations?



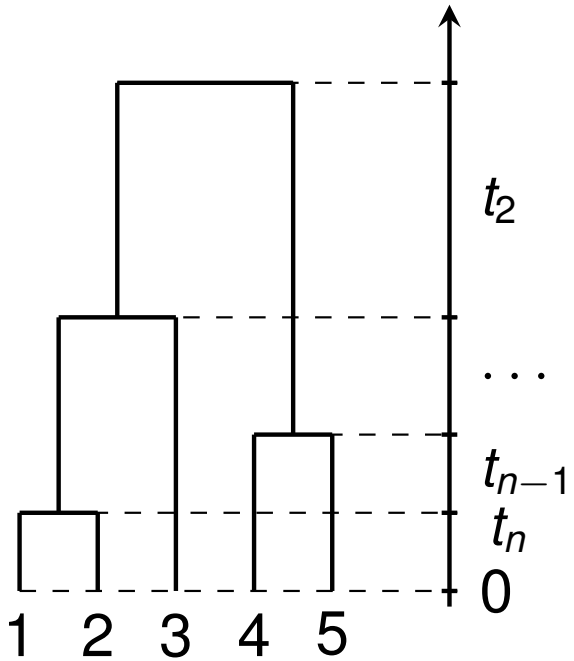
rate of coalescence

$$r_2 = 1/2N$$

$$r_j = \binom{j}{2}\frac{1}{2N} = \frac{j(j-1)}{4N}$$

Generalization for $n > 2$

- make Wright-Fisher simulations (pop. size $2N$)
- for each simulation, take n chromosomes at final time (present)
- trace back their genealogy
- measure t_j (in generations) and set $u_j = t_j/2N$ (rescaling)
- distribution of t_j and u_j over simulations?



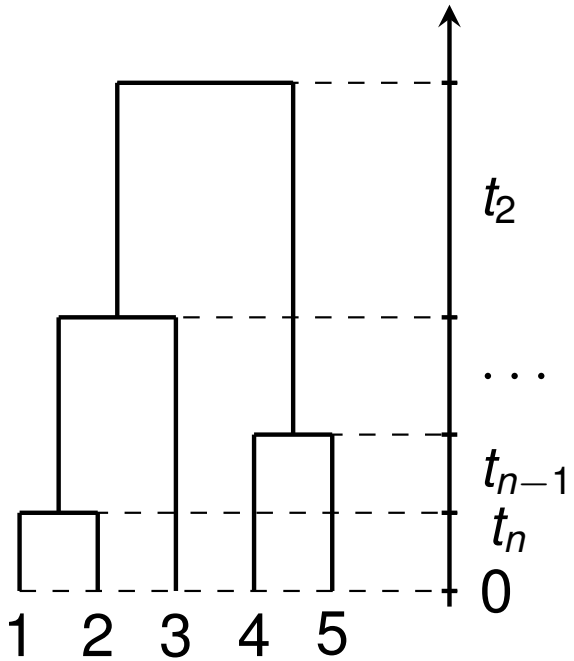
mean coalescence times

$$\bar{t}_2 \simeq 2N$$

$$\bar{t}_j \simeq \frac{4N}{j(j-1)}, \; j = 2..n$$

Generalization for $n > 2$

- make Wright-Fisher simulations (pop. size $2N$)
- for each simulation, take n chromosomes at final time (present)
- trace back their genealogy
- measure t_j (in generations) and set $u_j = t_j/2N$ (rescaling)
- distribution of t_j and u_j over simulations?



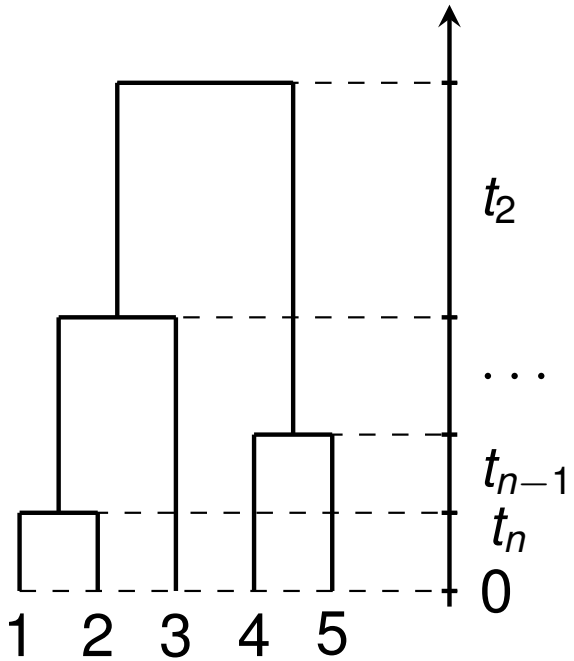
mean coalescence times

$$\overline{u}_2 \simeq 1$$

$$\overline{u}_j \simeq \frac{2}{j(j-1)}, \; j = 2..n$$

Generalization for $n > 2$

- make Wright-Fisher simulations (pop. size $2N$)
- for each simulation, take n chromosomes at final time (present)
- trace back their genealogy
- measure t_j (in generations) and set $u_j = t_j/2N$ (rescaling)
- distribution of t_j and u_j over simulations?

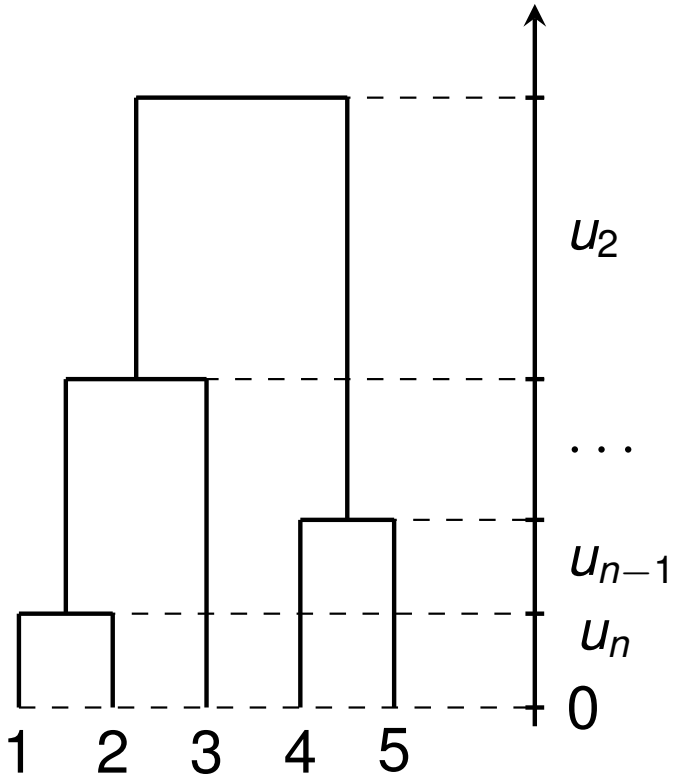


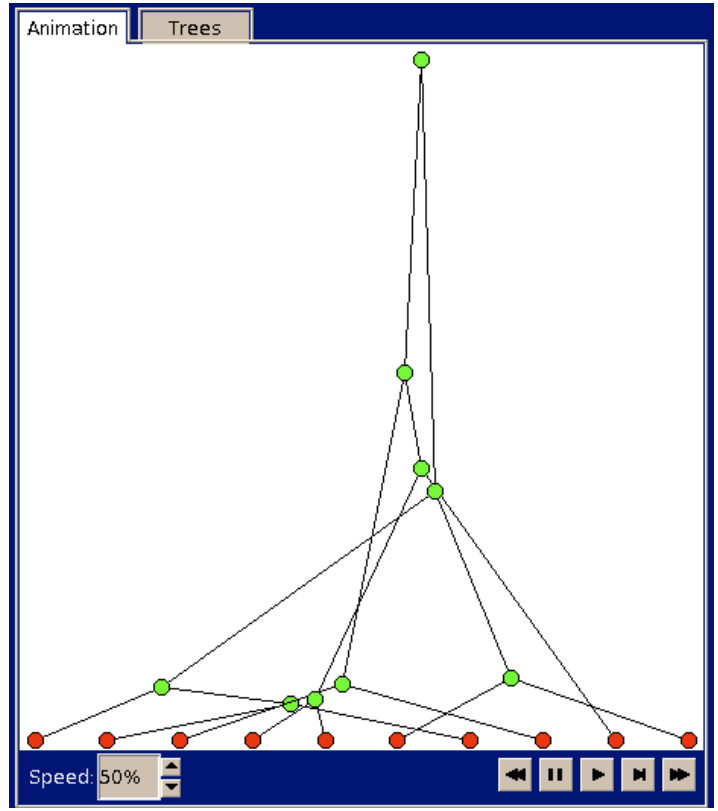
distribution of coal. times

$$t_j \sim Exp\left(mean = \frac{4N}{j(j-1)}\right)$$

$$u_j \sim Exp\left(mean = \frac{2}{j(j-1)}\right)$$

Drawing from the coalescent





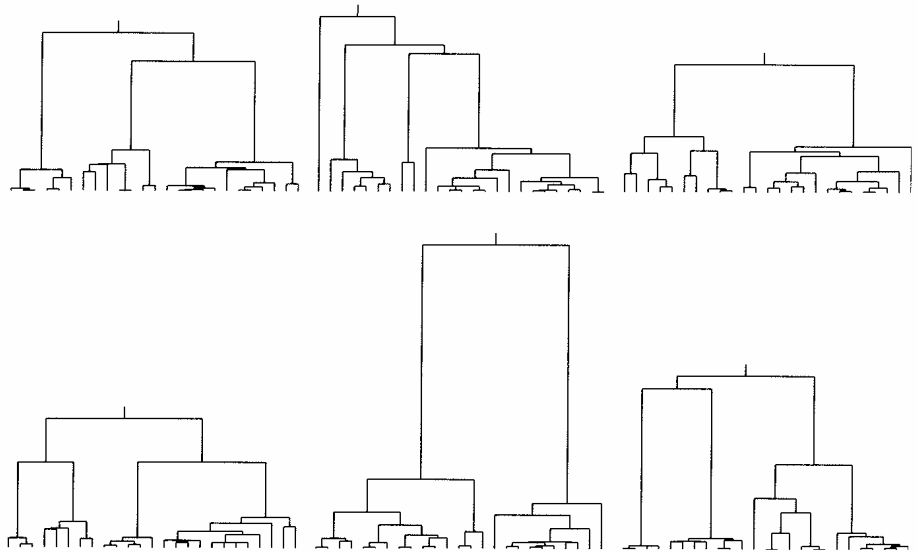
www.coalescent.dk

Algorithm

for $j = n..2$:

- draw $u_j \sim Exp\left(mean = \frac{j(j-1)}{2}\right)$
- join 2 of the j remaining lineages taken at random

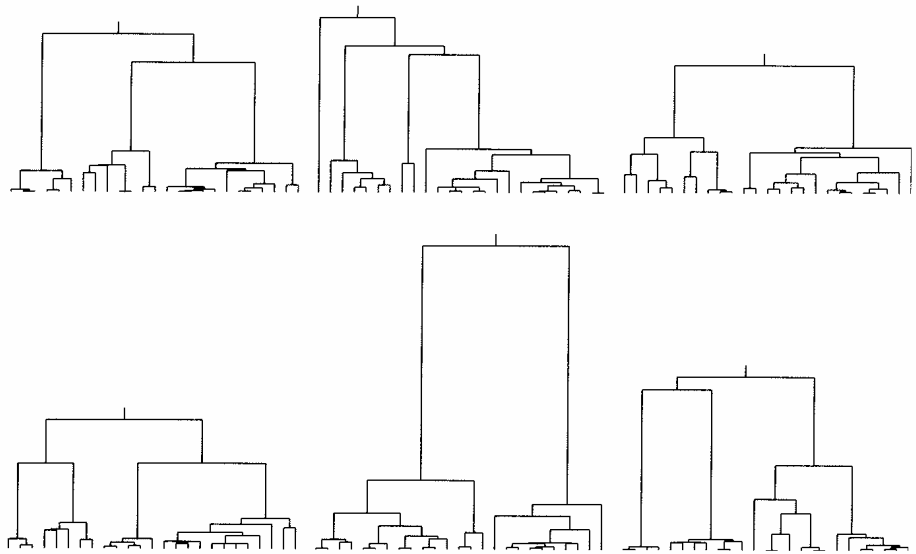
Drawing from the coalescent



Forward versus backward simulation

- forward: Wright Fisher simulation + backtracking of ancestors
- backward: Kingman's coalescent: drawing exponential variables
- equivalence ($n << N$), but
- Kingman's approach more efficient (in n instead of N^2)

Drawing from the coalescent



- large variability of deep branches
- high uncertainty on population size estimate based on one locus
- suggests approaches averaging over several independent loci

What is coalescent theory useful for?

Theory

- obtaining insights about patterns in sequence variation
- deriving theoretical expectations
  (e.g. age of sample's last common ancestor)

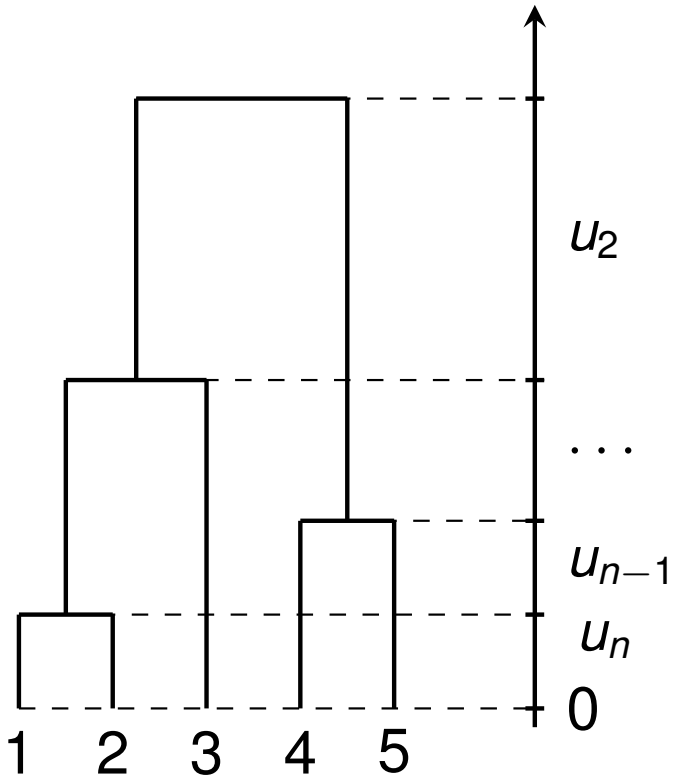
Simulations

- null distribution for hypothesis testing
- detecting departures from neutrality (selection)

Parameter estimation

- estimating $\theta = 4Nu$ based on observed polymorphism
- estimating demographic scenarios (see course 2)

Mean age of most recent common ancestor (MRCA)



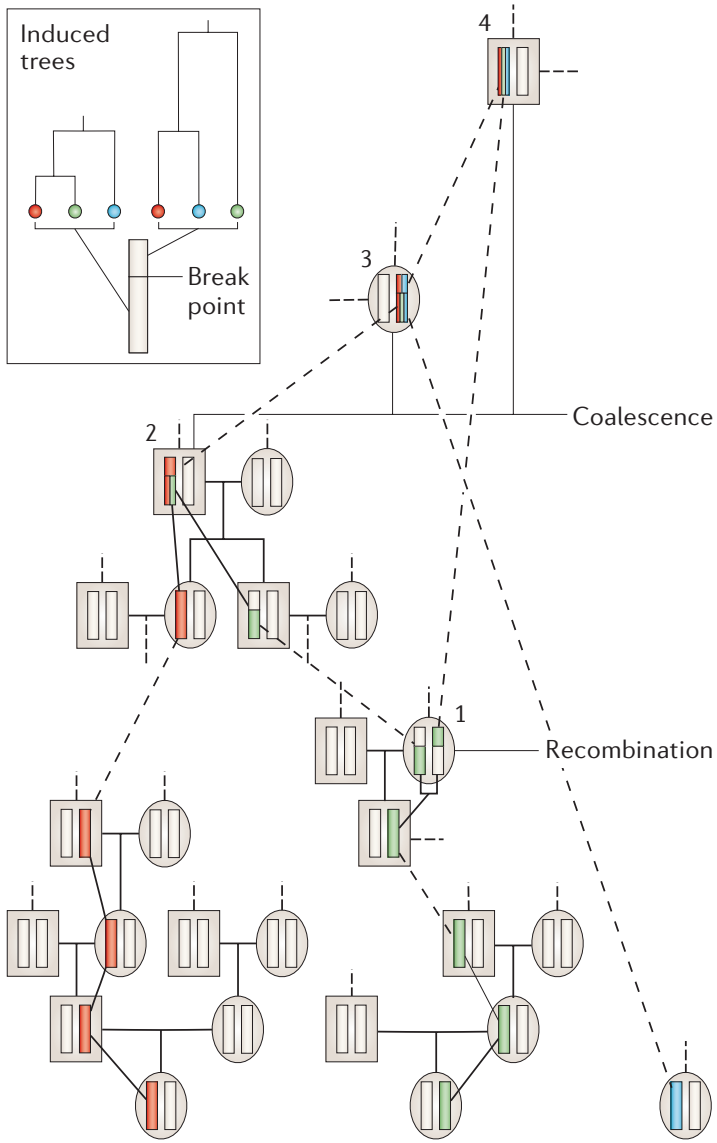
$$\begin{aligned} T_n &= u_n + u_{n-1} + \ldots + u_2 \\ E[T_n] &= 2(1 - 1/n) \end{aligned}$$

- expected MRCA age reaches a limit (4N generations) for large n
- intra-specific variation gives access to relatively shallow past
- in contrast to interspecific divergence (human chimp: 6 Myrs)

Age of most recent common ancestor

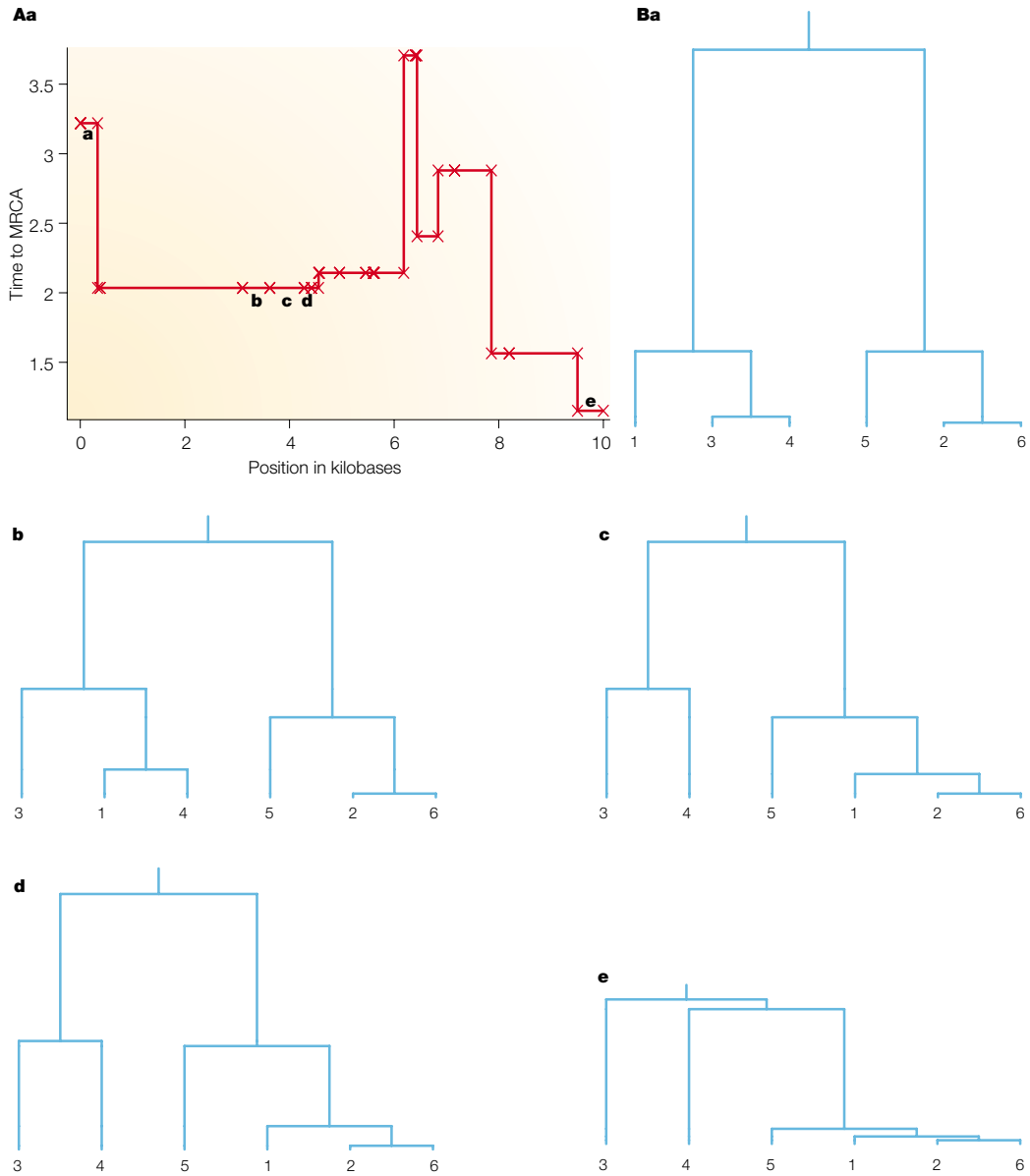
- mitochondrial: 200 000 years (Soares et al, 2009, Am J Human Genet 84:740)
- Y chromosome: 55 000 years (Thomson et al, 2000, PNAS, 97:7360)
- nuclear genome: variation along genome

Genealogies and recombination



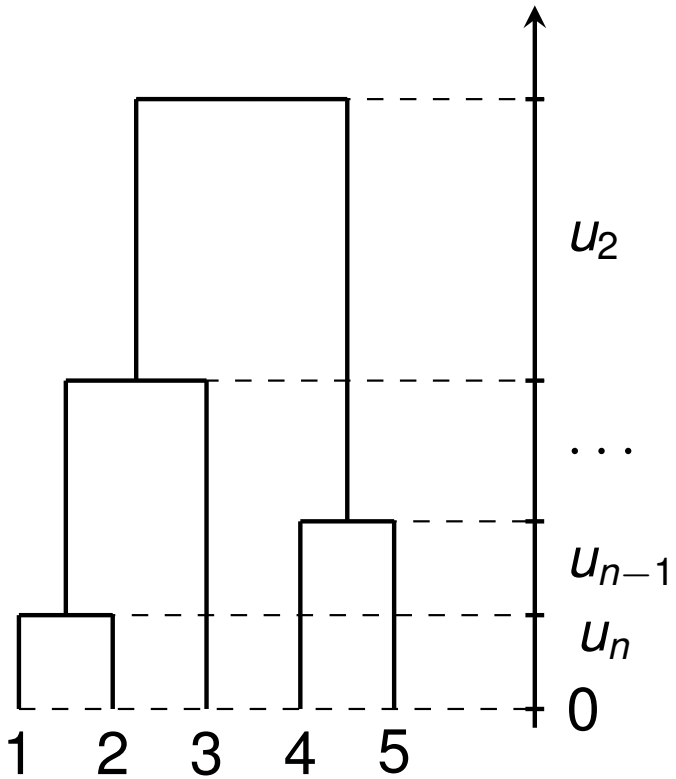
Marjoram and Tavaré, 2006, Nat Rev Genet, 7:759

Genealogies and recombination



Rosenberg and Nordborg, 2002, Nat Rev Genet, 3:380

Total length of the genealogy



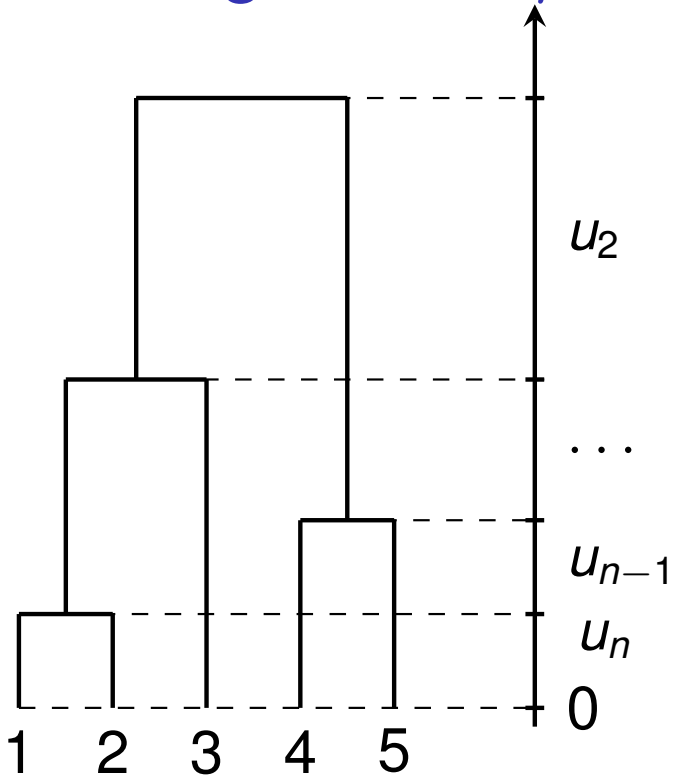
$$
\begin{aligned}
L_n &= \sum_{j=2}^{n} j\, u_j \\
E[L_n] &= \sum_{j=2}^{n} j \frac{2}{j(j-1)} \\
&= 2 \sum_{j=2}^{n} \frac{1}{(j-1)}
\end{aligned}
$$

for large n

$$E[L_n] \sim 2 \ln n$$

(slow increase)

Estimating $\theta = 4N\mu$: Watterson's estimator



$$L_n = \sum_{j=2}^{n} j\, u_j$$

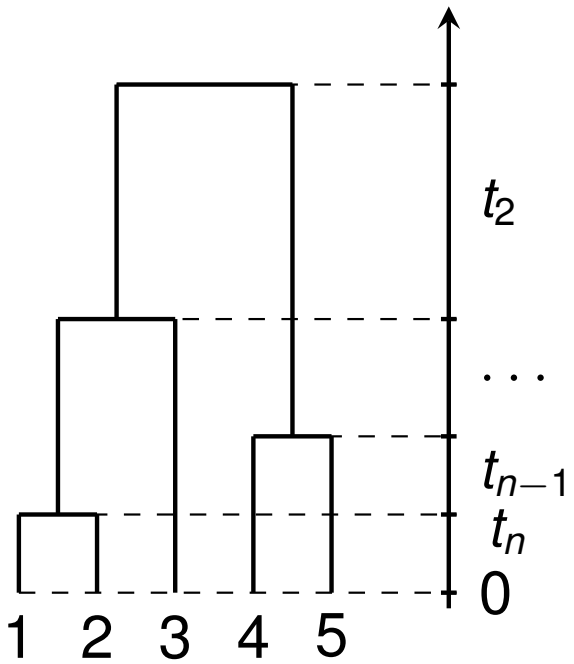
$$E[L_n] = 2 \sum_{j=2}^{n} \frac{1}{(j-1)}$$

- S_n: number of sites segregating in the sample
- low mutation rate: S_n = total # mutations along genealogy

$$E[S_n] = 2N\mu\, E[L_n] = \theta\, E[L_n]/2$$

$$\hat{\theta} = \frac{2S_n}{E[L_n]}$$

Estimating $\theta = 4N\mu$: Tajima versus Watterson



Tajima's estimator of scaled mutation rate

- π_{ij}: fraction of polymorphic sites between haplotypes i and j

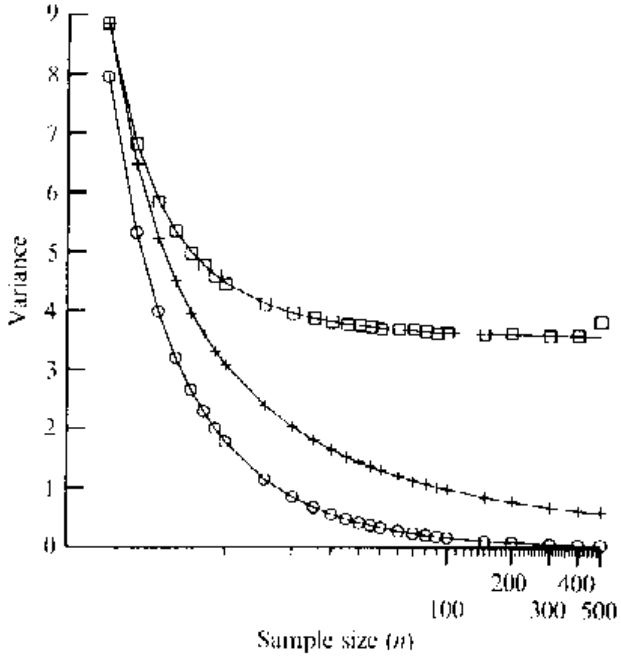
$$\hat{\pi} = \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij}$$

Watterson's estimator

- S_n: number of sites segregating in the sample
- $E[L_n]$: mean total length of genealogy

$$\hat{\theta} = \frac{2S_n}{E[L_n]}$$

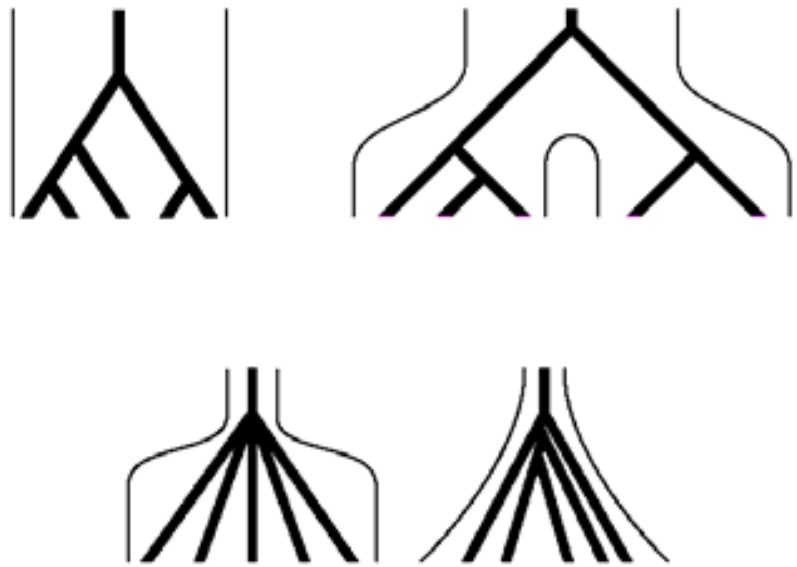
Variance of the two estimators



Felsenstein 1992

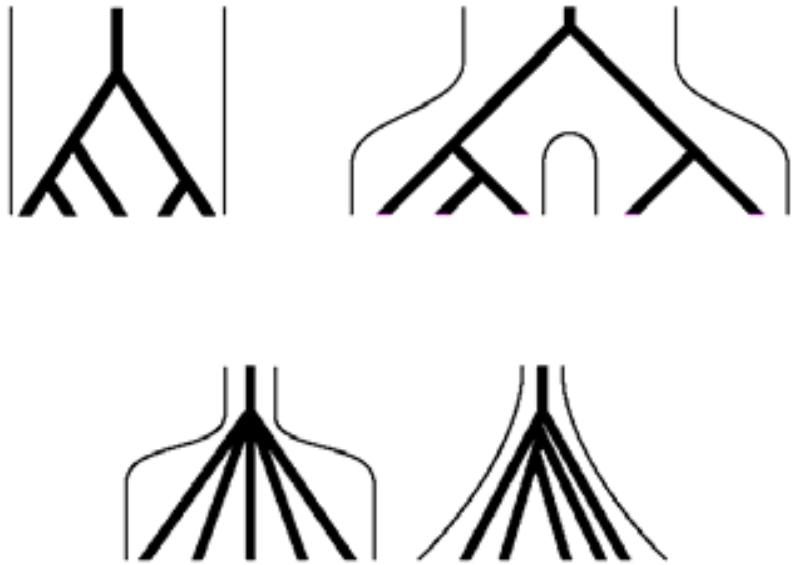
- Tajima's estimator is not consistent
- Watterson's estimator consistent but not optimal
- maximum likelihood (see later) optimal and more general

Demography and population structure



- changes in population size induce changes in rate of coalescence
- at time t, rate of coalescence of j lineages is $j(j-1)/4N(t)$
- increasing population: comparatively higher rates in distant past
- decreasing population: comparatively higher rates near present

Demography and population structure



- Tajima's and Watterson's estimates respond differently to changes in N
- increasing population: $d = \hat{\pi} - \hat{\theta} < 0$
- decreasing population: $d = \hat{\pi} - \hat{\theta} > 0$
- Tajima's $D = d/\hat{V}(d)$

Hypothesis testing using Tajima's D

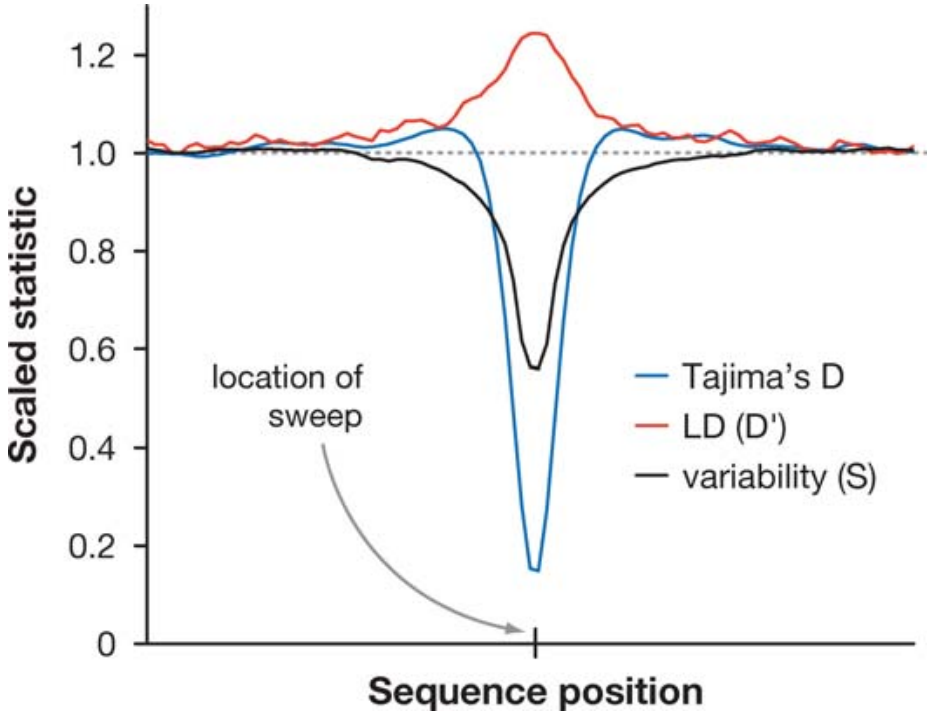
Principle

- estimate $\hat{\pi}$ and $\hat{\theta}$, compute D
- simulate genealogies and distribute mutations over it with rate $\hat{\theta}$
- on each replicate, estimate $\hat{\pi}$ and $\hat{\theta}$, compute D: null distribution

Scope and limits

- significant deviation: departure from any assumption
- demography ($D < 0$: population increase)
- selection ($D < 0$: directional selection, $D > 0$ balancing selection)
- panmixia (but D is more robust to this)

Tajma's D and selection



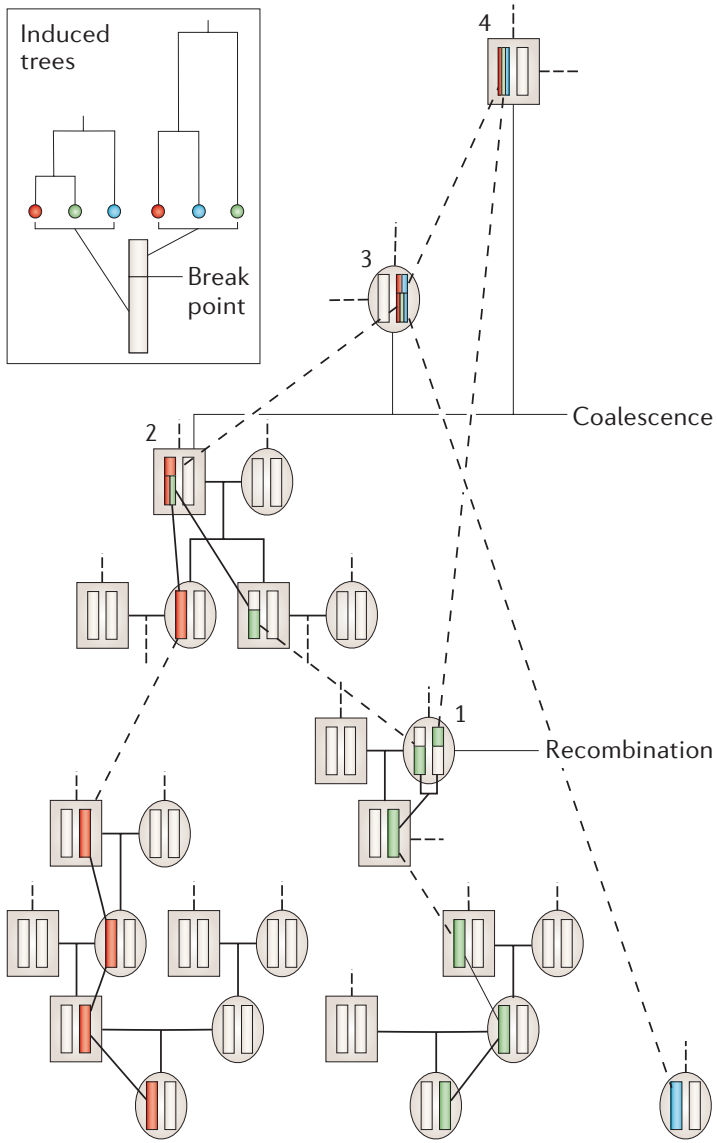
Nielsen, 2005, Ann Rev Genet 39:197

- directional selection like population increase (at selected locus)
- locally in genome, looks like demographic expansion
- recombination progressively dissipates linkage with nearby neutral polymorphisms

Extensions to Kingman's coalescent

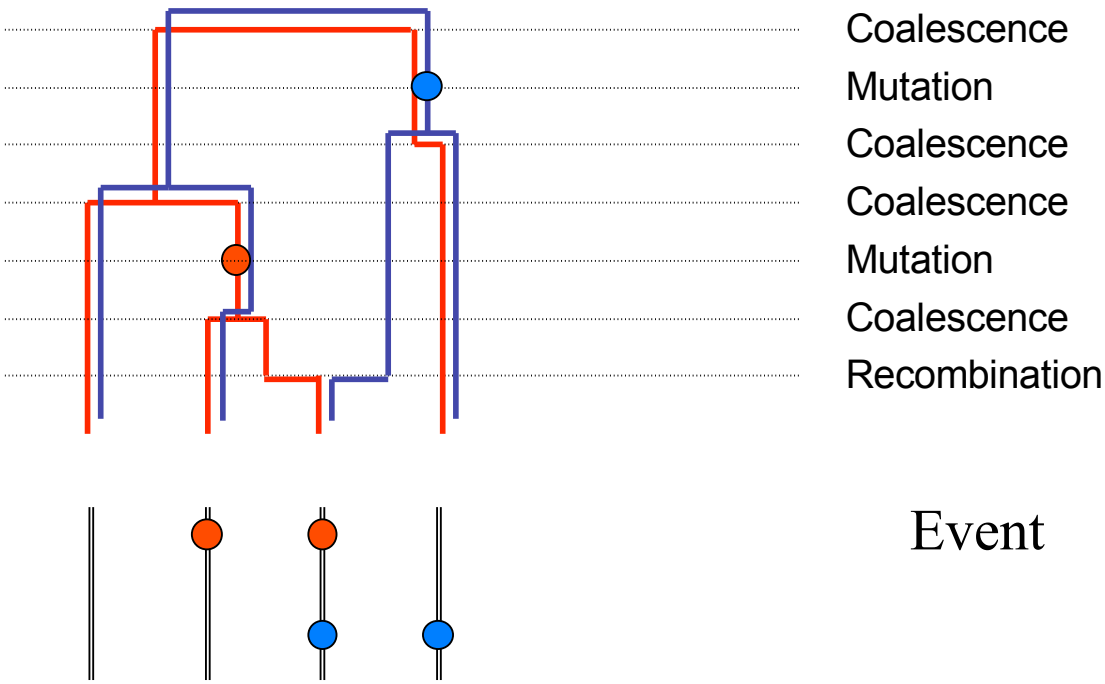
- with demographic variation (time-dependent $N(t)$)
- with population structure (demes with migration between demes)
- with recombination (ancestral recombination graphs)
  - Hudson 1983, Theor Popul Biol 23:183.
  - important tool for estimating recombination rates along genomes
- with selection (ancestral selection graphs)
  - Krone and Neuhauser, 1997, Theor Popul Biol 51:210.

Genealogies and recombination



Marjoram and Tavaré, 2006, Nat Rev Genet, 7:759

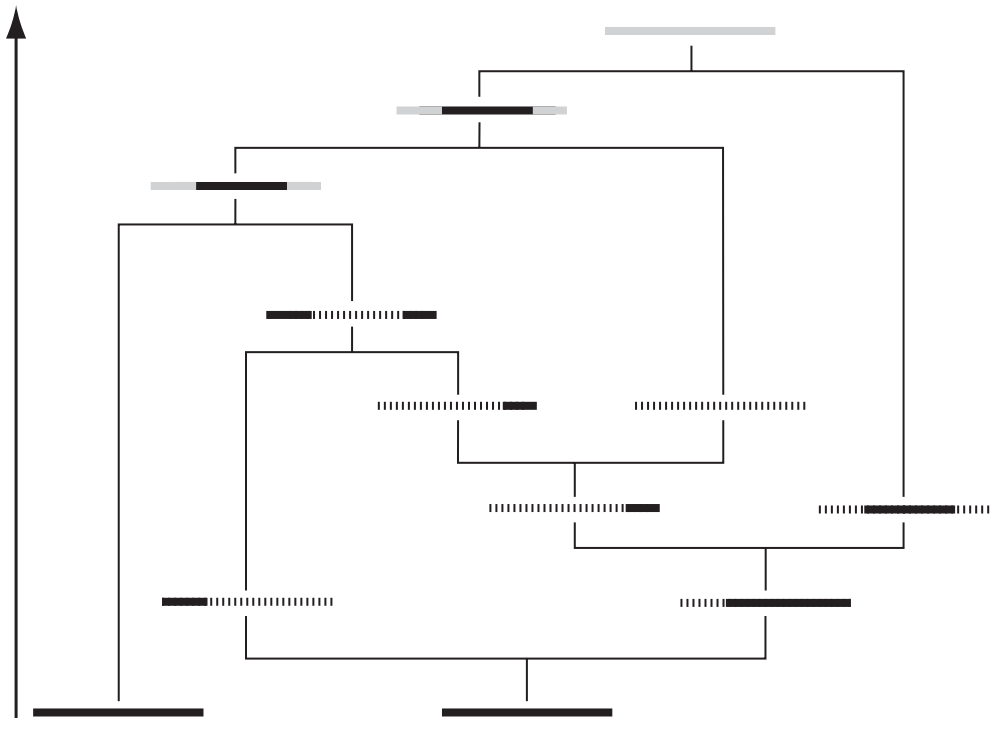
Ancestral recombination graph: 2 loci



from Awadalla (McVean, Awadalla and Fearnhead, Genetics, 160:1231)

- scaled recombination rate $\rho = 4Nr$
- coalescence at rate $j(j-1)/2$
- recombination at rate $j\rho/2$

Ancestral recombination graph: continuous segment of loci

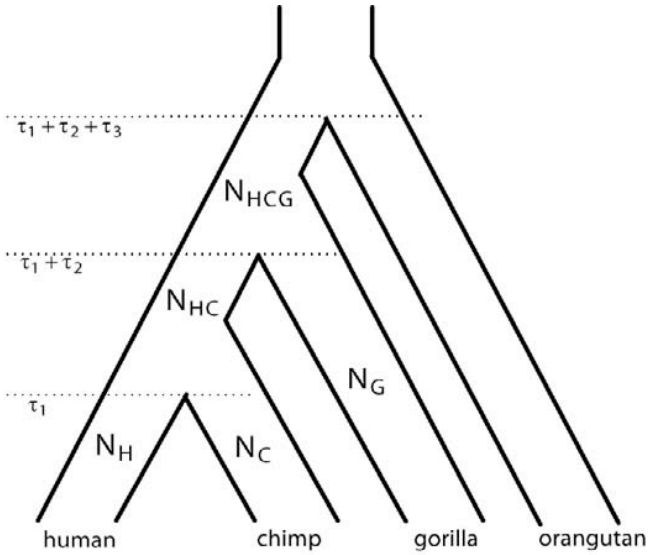


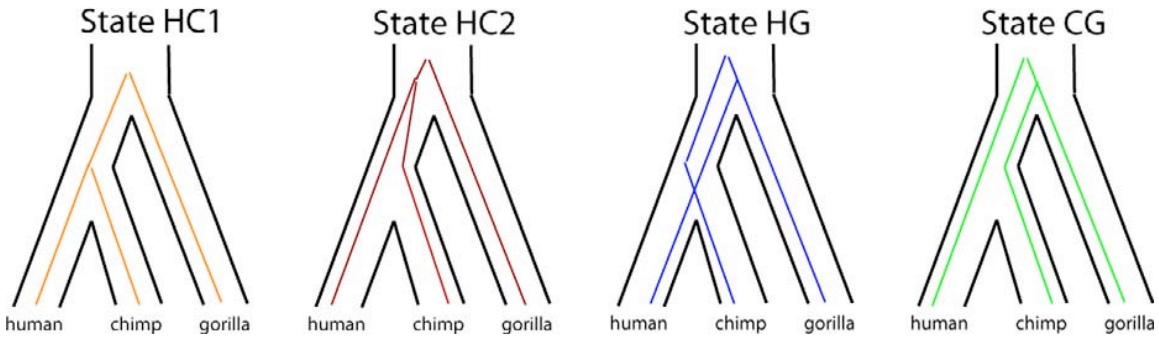
Coal.	Recomb.	Anc. mat.
1	ρ	$\frac{1}{2}\rho+\lambda$
3	$\frac{3}{2}\rho$	$\frac{1}{2}\rho+\lambda$
6	$2\,\rho$	$\rho+\lambda$
10	$\frac{5}{2}\rho$	ρ
6	$2\,\rho$	ρ
3	$\frac{3}{2}\rho$	ρ
1	ρ	ρ

Hein, Shierup and Wiuf, 2005

- scaled recombination rate (for whole segment) $\rho = 4Nr$
- coalescence at rate $j(j-1)/2$
- recombination at rate $j\rho/2$

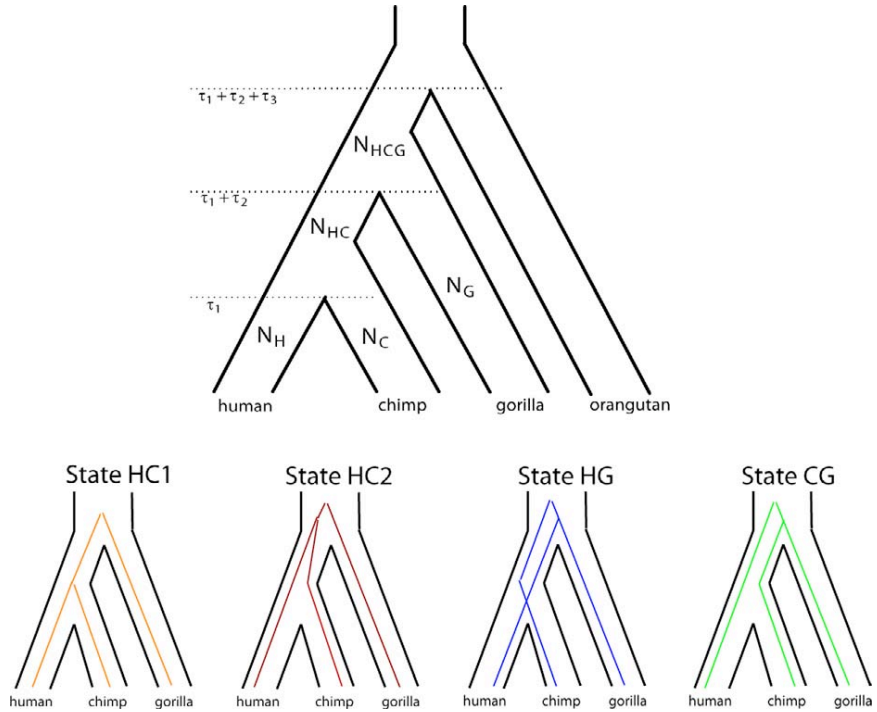
Lineage sorting





Hobolth et al, PLoS Genetics, 2007, 3 p.e7

Lineage sorting: structured coalescent



Probability of locus genealogy

$$
\begin{aligned}
p(HC1) &= 1 - e^{-2\tau_2/N_{HC}} \\
p(HC2) = p(HG) = p(CG) &= 1/3\, e^{-2\tau_2/N_{HC}} \\
p(HC) &= p(HC1) + p(HC2)
\end{aligned}
$$

Estimating ancestral population size

Tree mismatch approach (Nei 1987)

- for each locus, reconstruct most likely tree
- count proportions of trees = HC, HG or CG
- solve equation (last slide) for τ_2/N_{HC}
- assuming $\tau_2 = 1.6$ Myrs, this yields $N_{HC} = 100,000 \pm 50,000$.

Problems

- bias due to stochastic tree reconstruction errors
- even under no lineage sorting, trees might differ due to finite alignment size
- results in an inflated estimate for N_{HC}
- need to use probabilistic models to improve on this estimate

Summary and conclusions

Summary

- rate of coalescence of j lineages is $j(j-1)/4N$
- depth of genealogy reflects population size
- shape of genealogy reflects demographic history
- Kingman's coalescent: simple and powerful model for
  - understanding population genetics
  - estimating parameters
  - testing models

From there

- coalescent at the core of probabilistic models for statistical inference
- represents the natural law for integrating over unknown genealogies