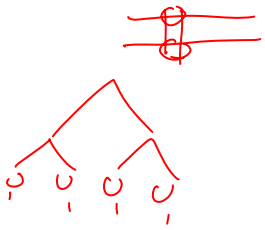
Computational Genomics

The Coalescent Process

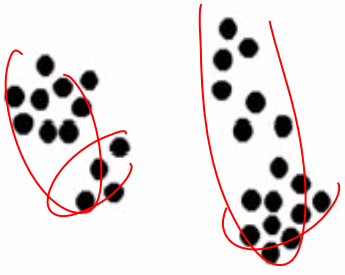


Eric Xing

Lecture 12, February 22, 2007

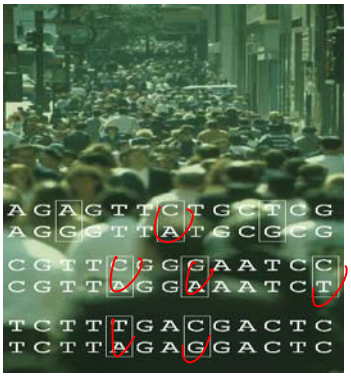
DTW Chap 13

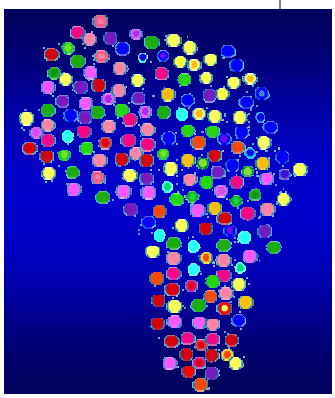
Clustering



- How to label them ?
  - inference
- How many clusters ???
  - model selection ?
  - or inference ?

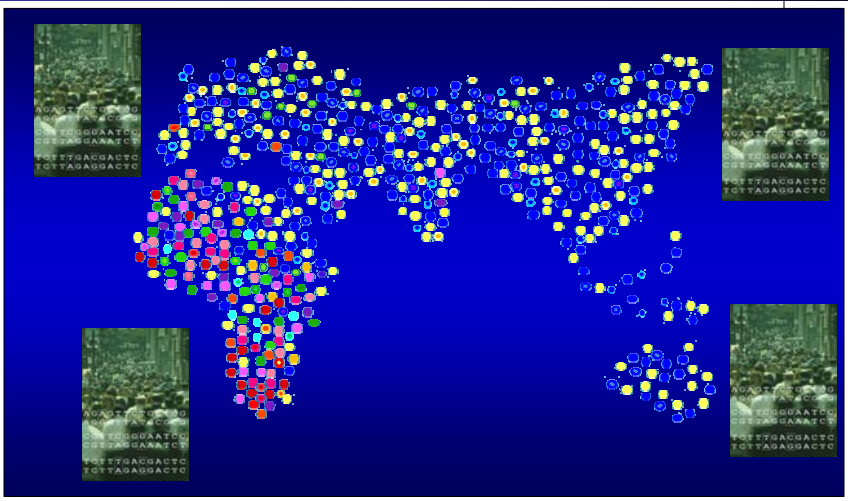
Genetic Demography





- Are there genetic prototypes among them ?
- What are they ?
- How many ? (how many ancestors do we have ?)

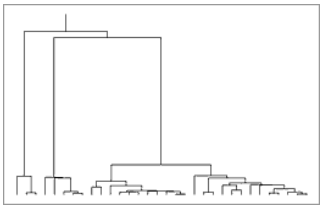
Multi-population Genetic Demography



- Inference done separately, or jointly?

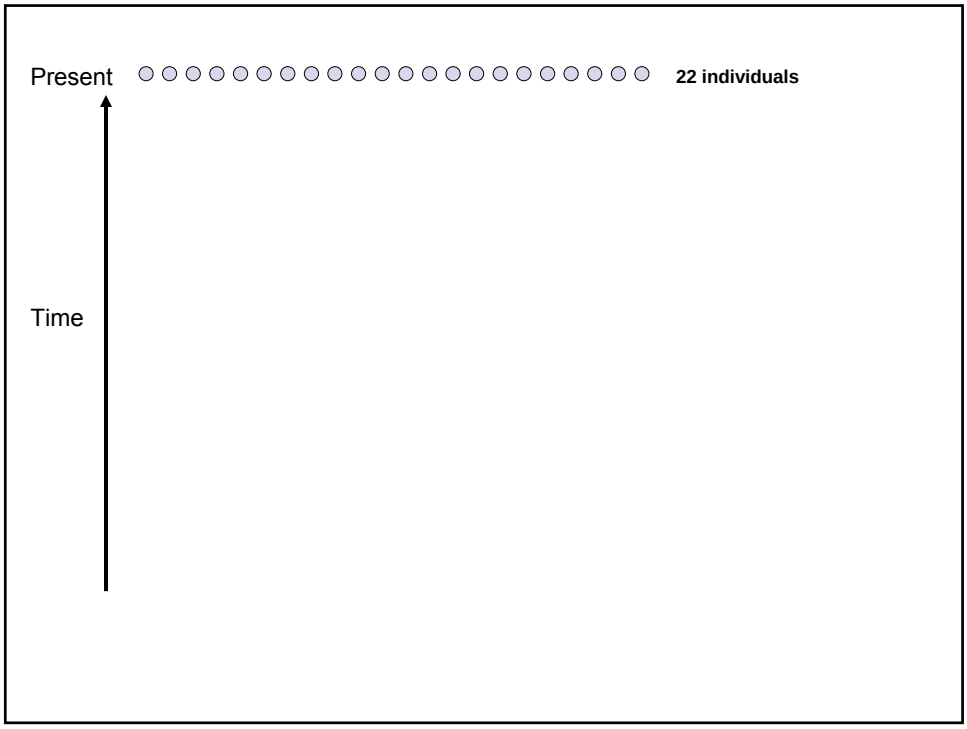
Sir John Kingman,
Head of the Isaac Newton Institute of Mathematical Sciences

The coalescent

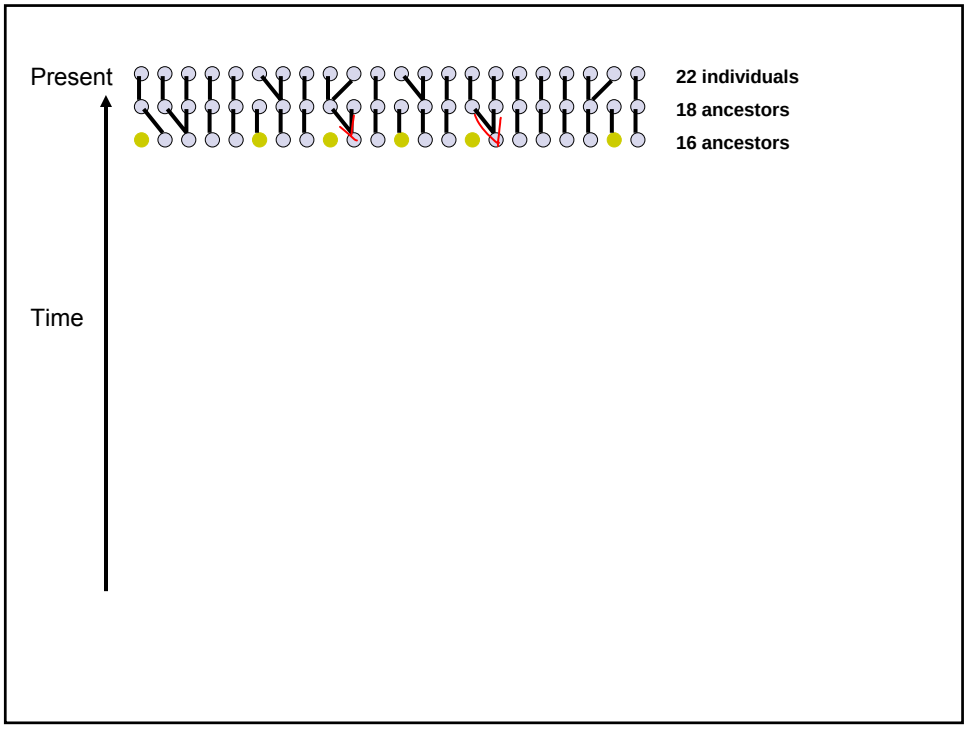


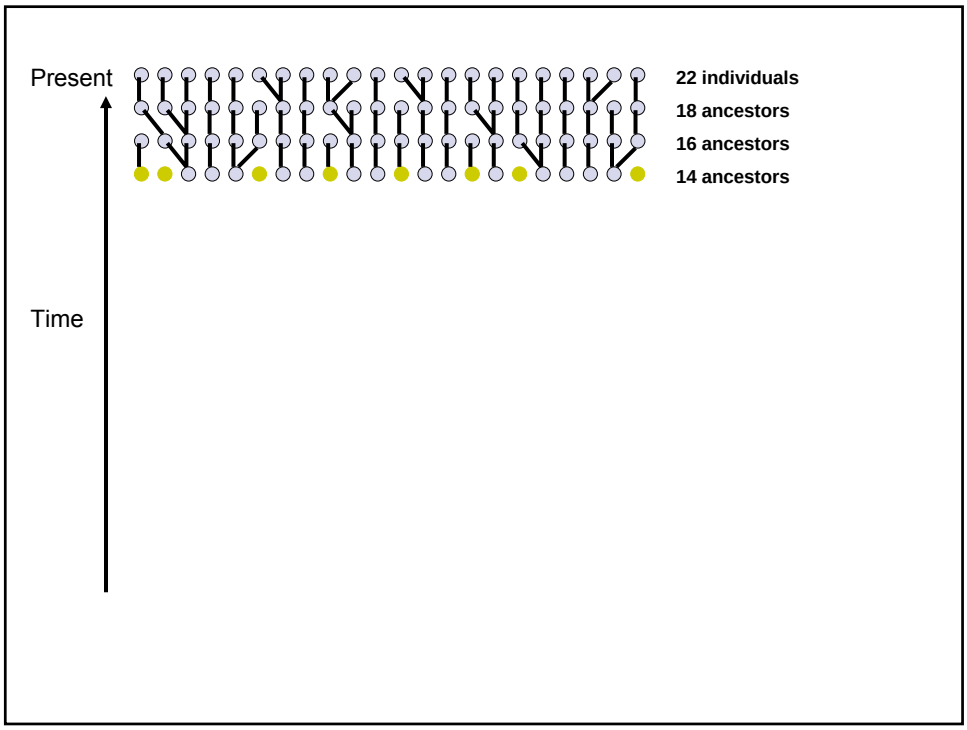
Coalescent Theory

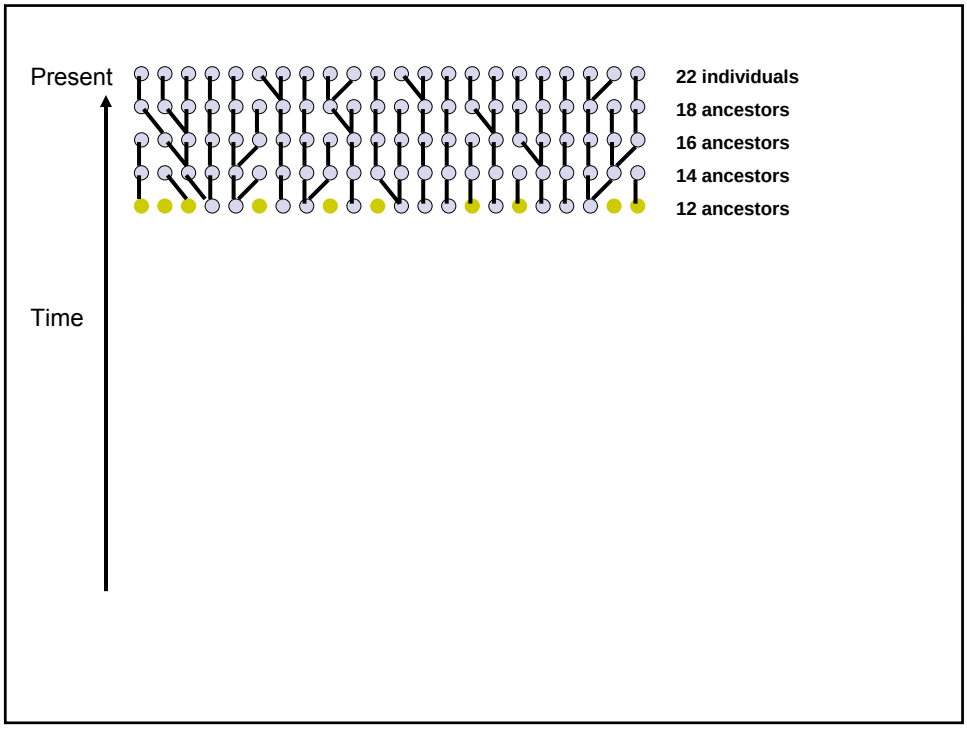
- how we can build up a genealogical tree to relate a sample of *n* haploid individuals, collected in the present day?
  - The following series of slides shows how you can build up a genealogical tree to relate a sample of 22 individuals, collected in the present day, at a single haplotype locus (e.g. the non-recombining Y chromosome).
  - Because (for the Y chromosome) one son has only one father, but one father can have more than one son, coalescent events occur in the genealogy which inevitably result in a reduction of ancestors. Eventually, one ancestor remains – the Most Recent Common Ancestor (MRCA).

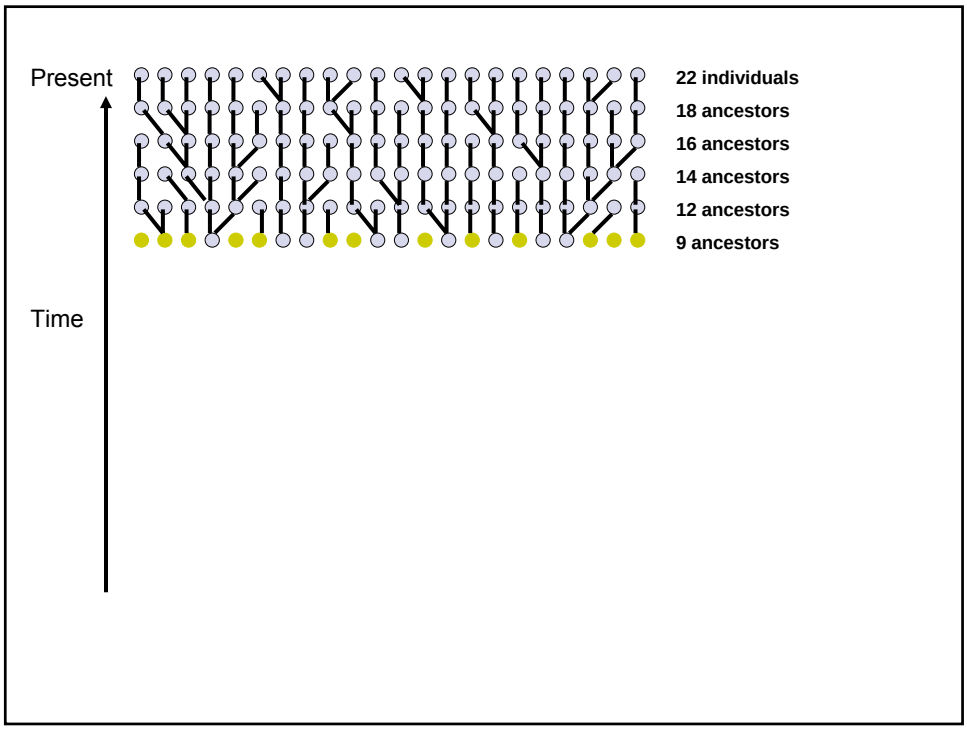

Present
22 individuals
Time

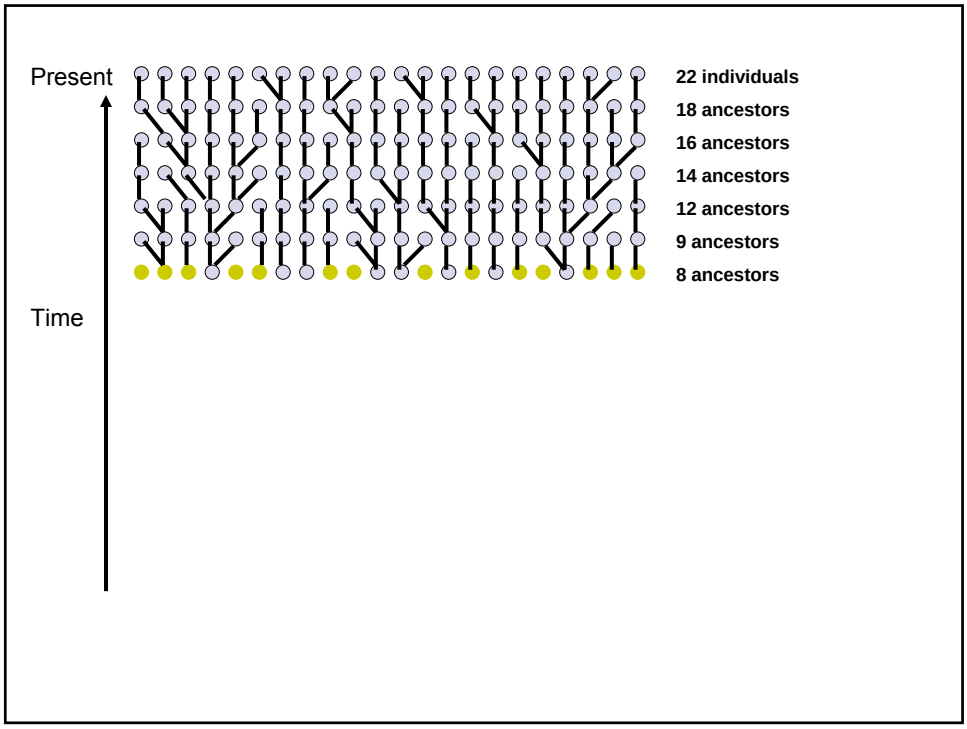

Present
22 individuals
18 ancestors
Time

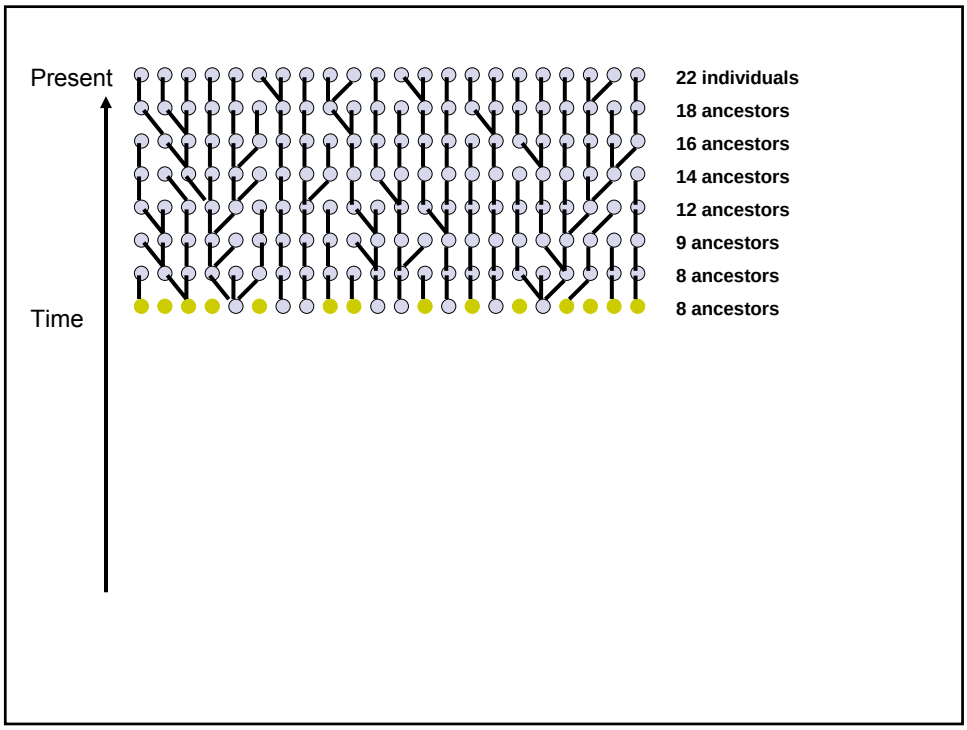

Present
Time
22 individuals
18 ancestors
16 ancestors

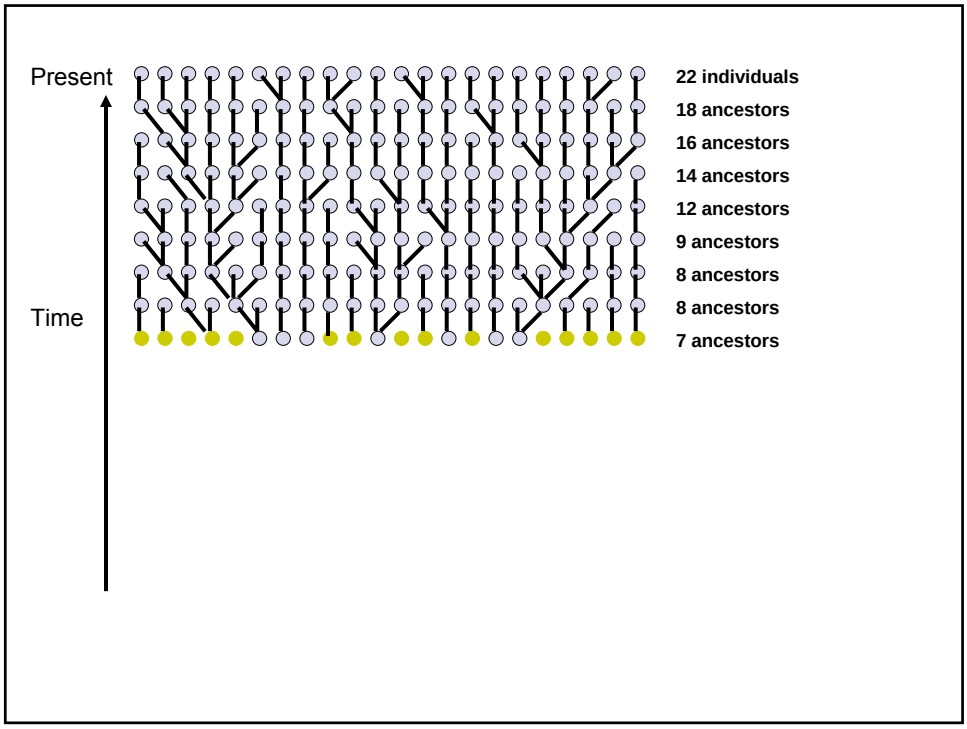

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors

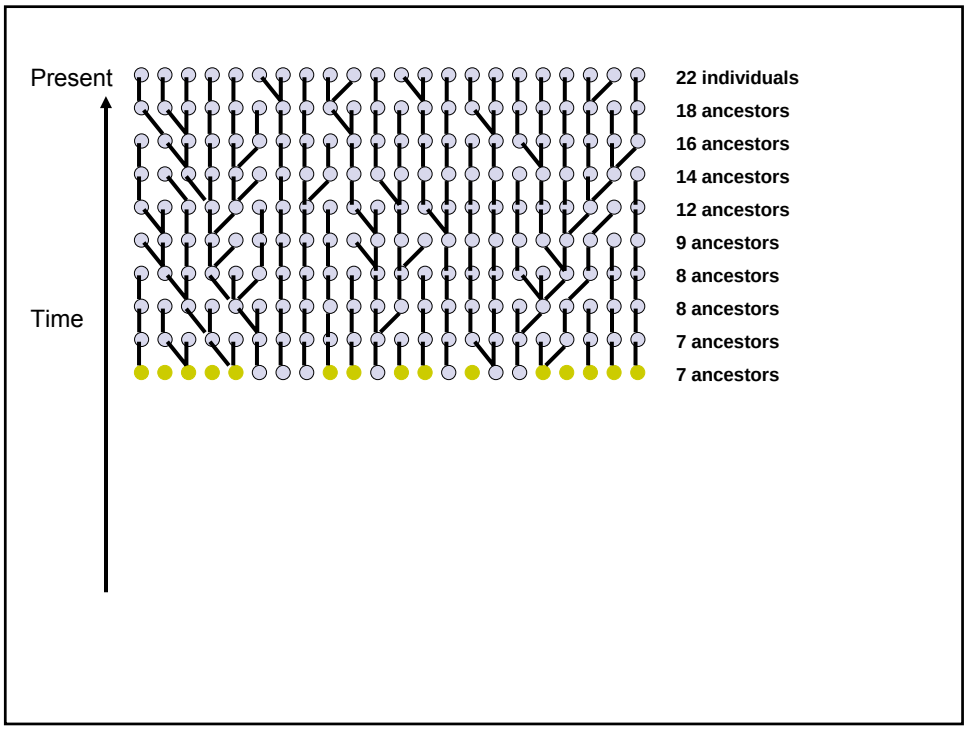

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors

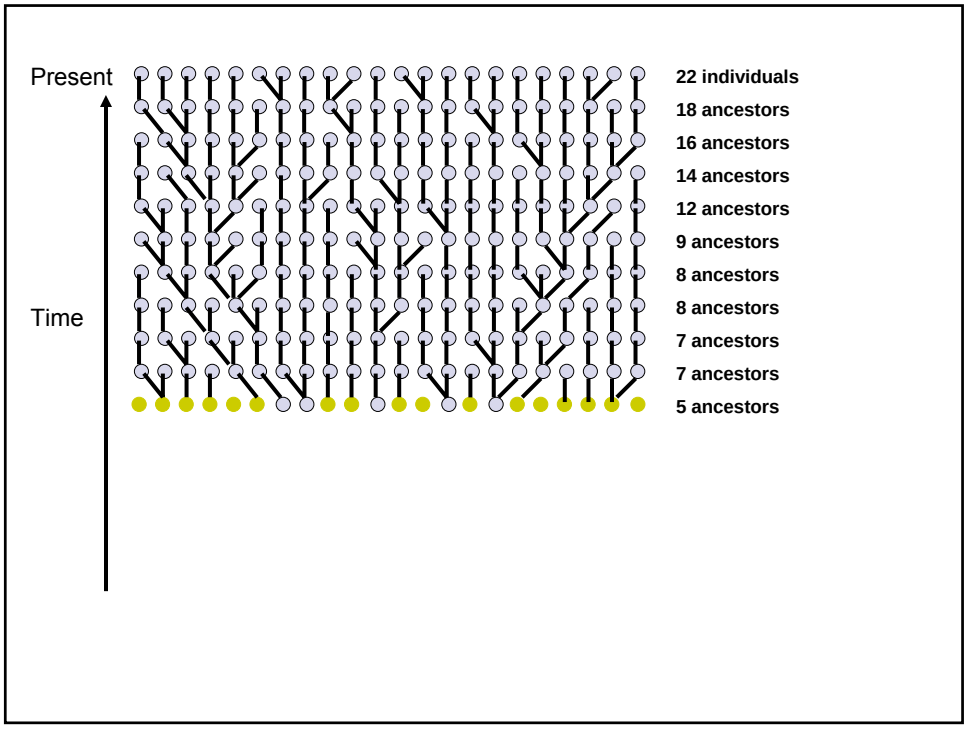

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors

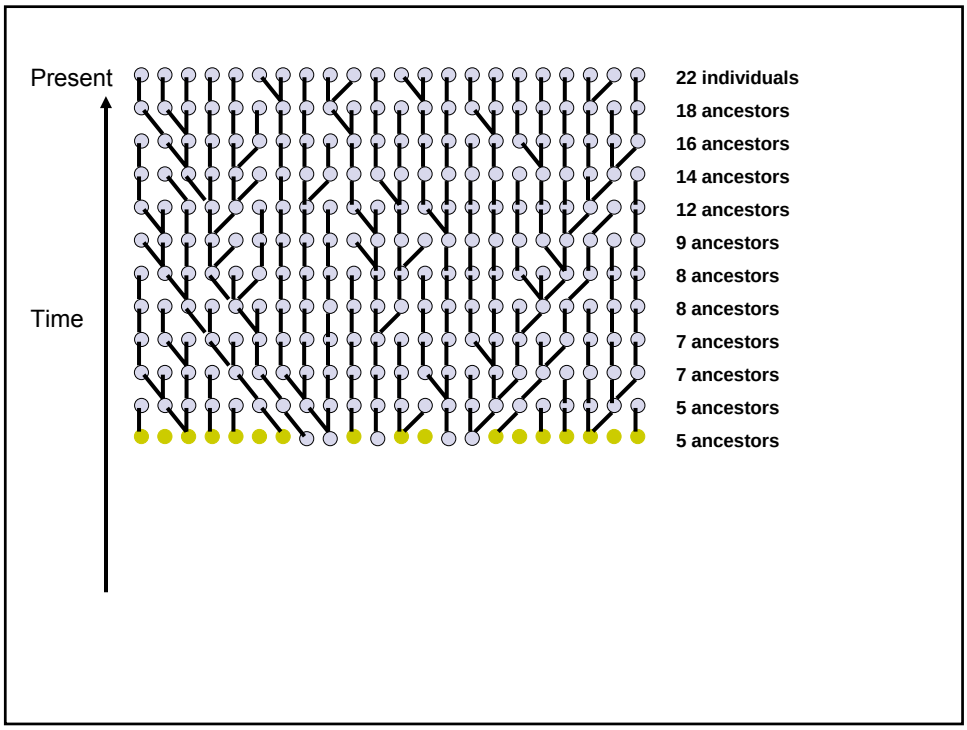

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors

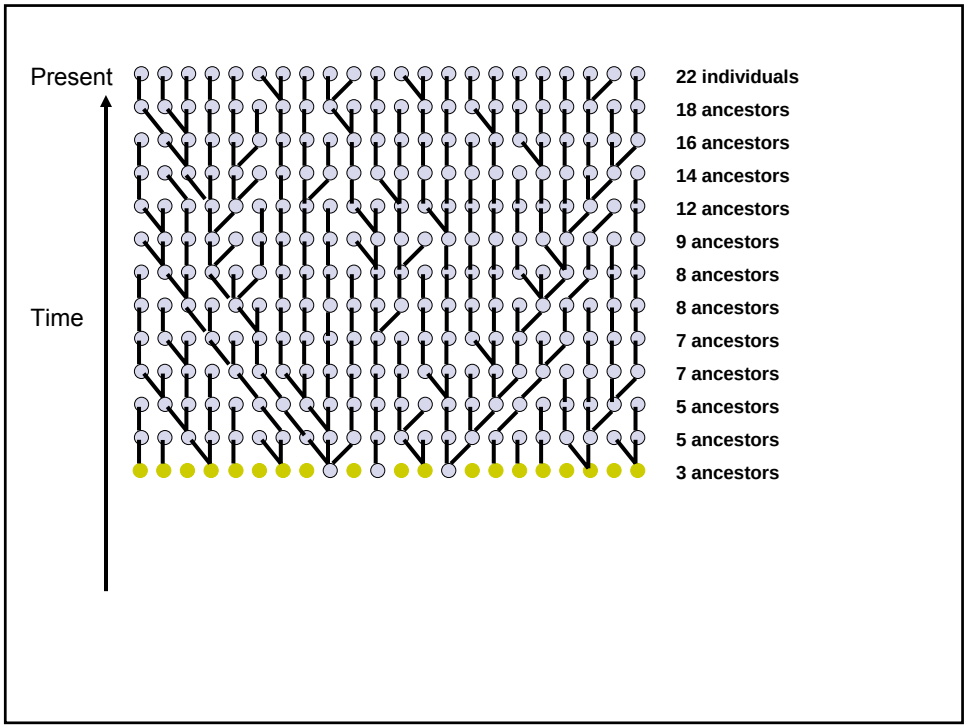

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors

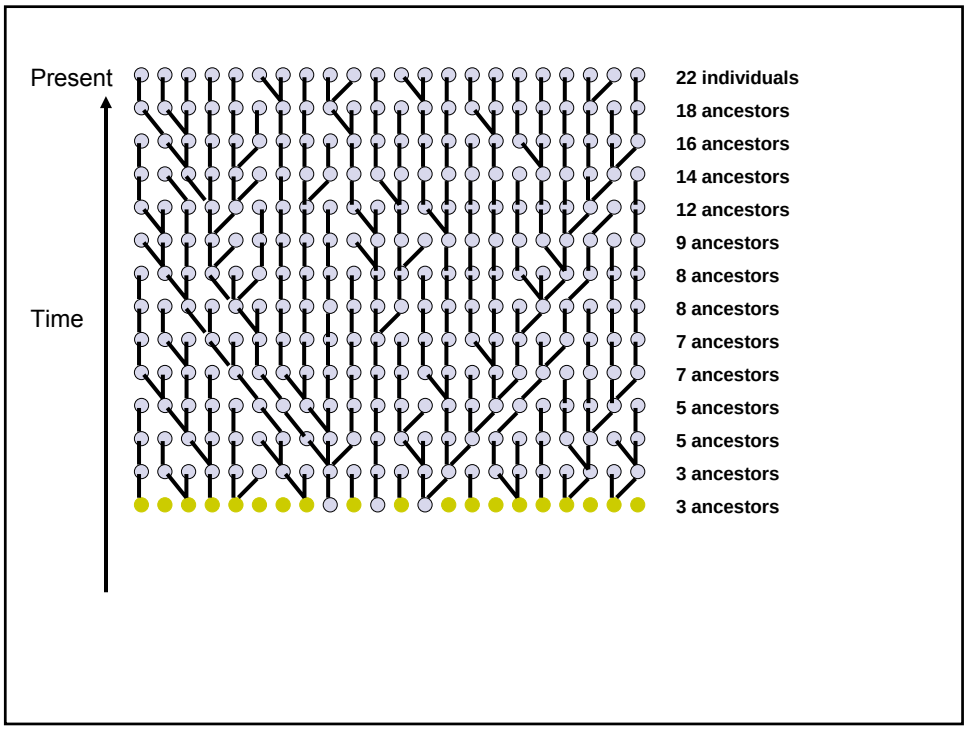

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors

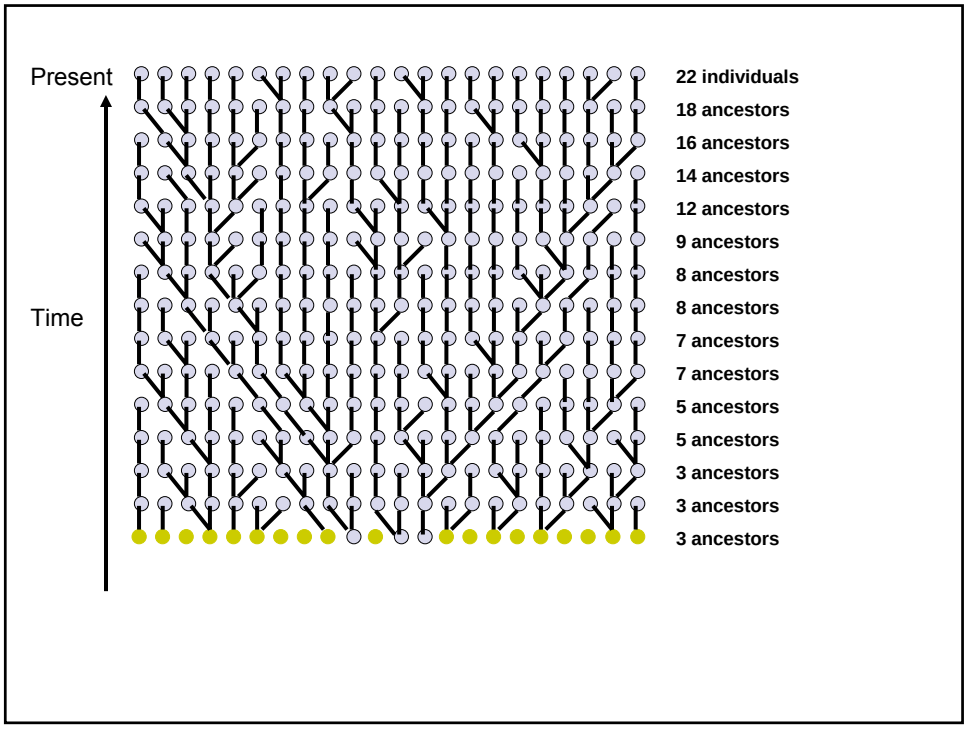

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors

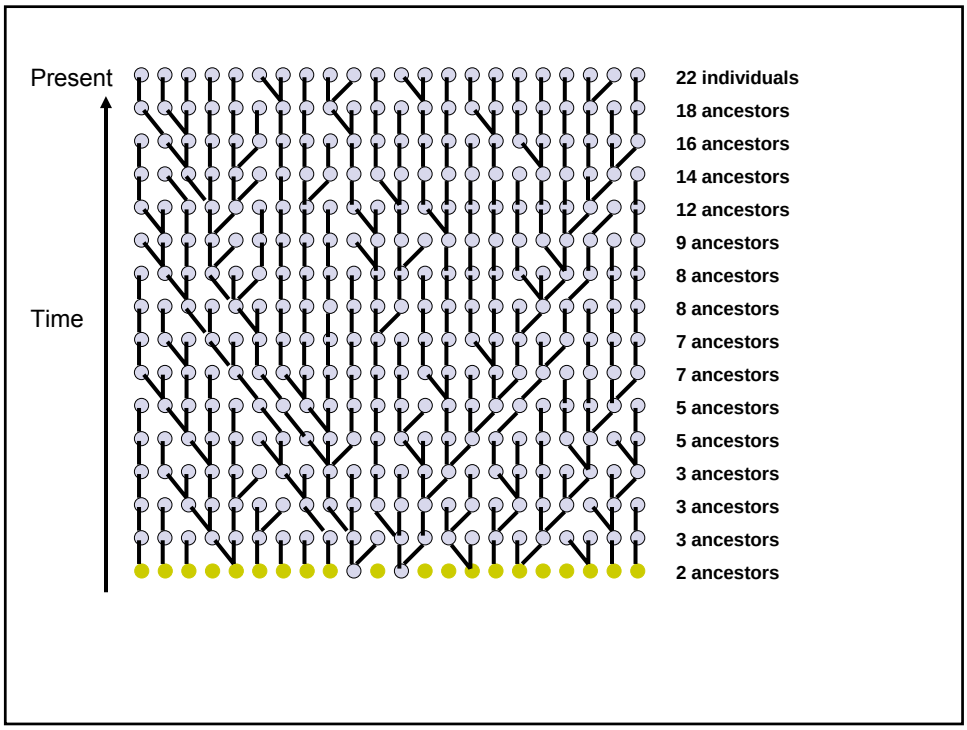

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors

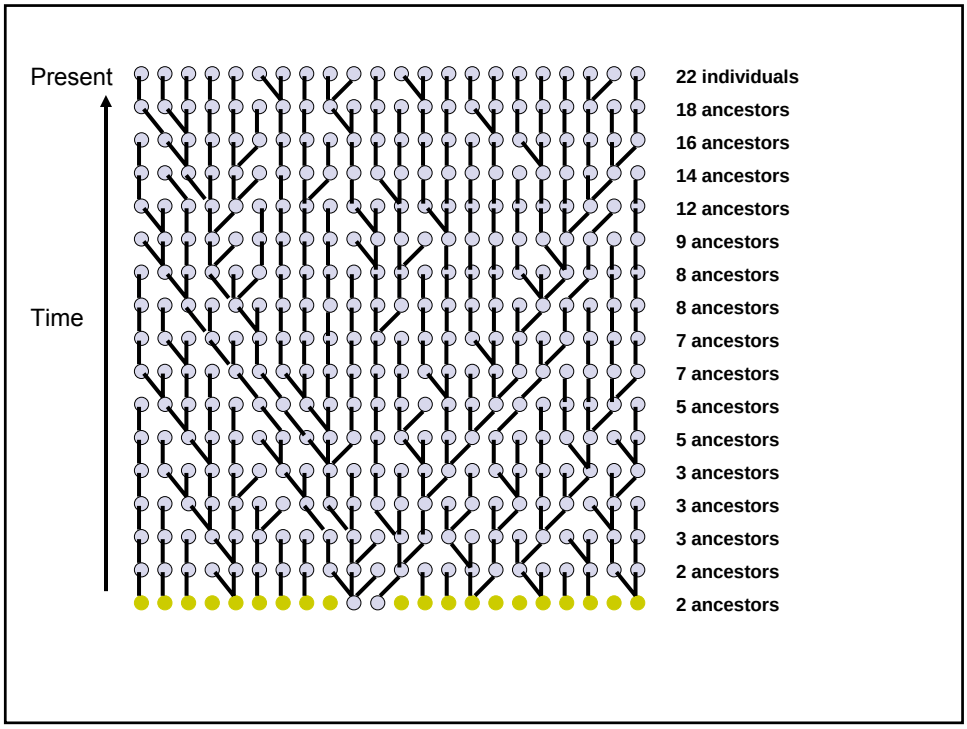

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors

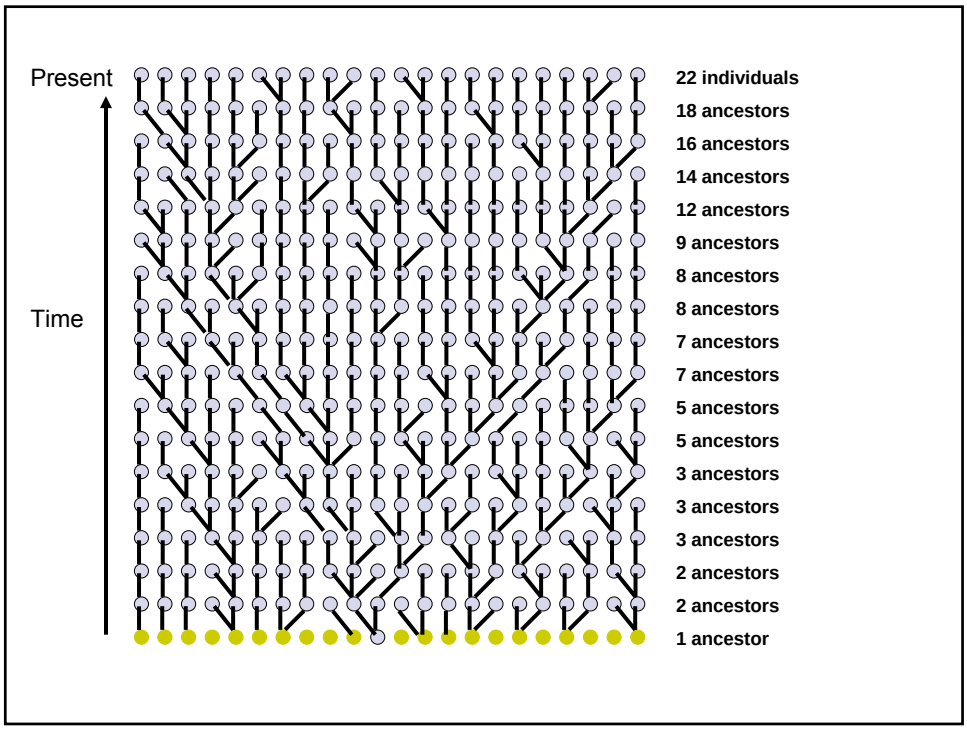

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors

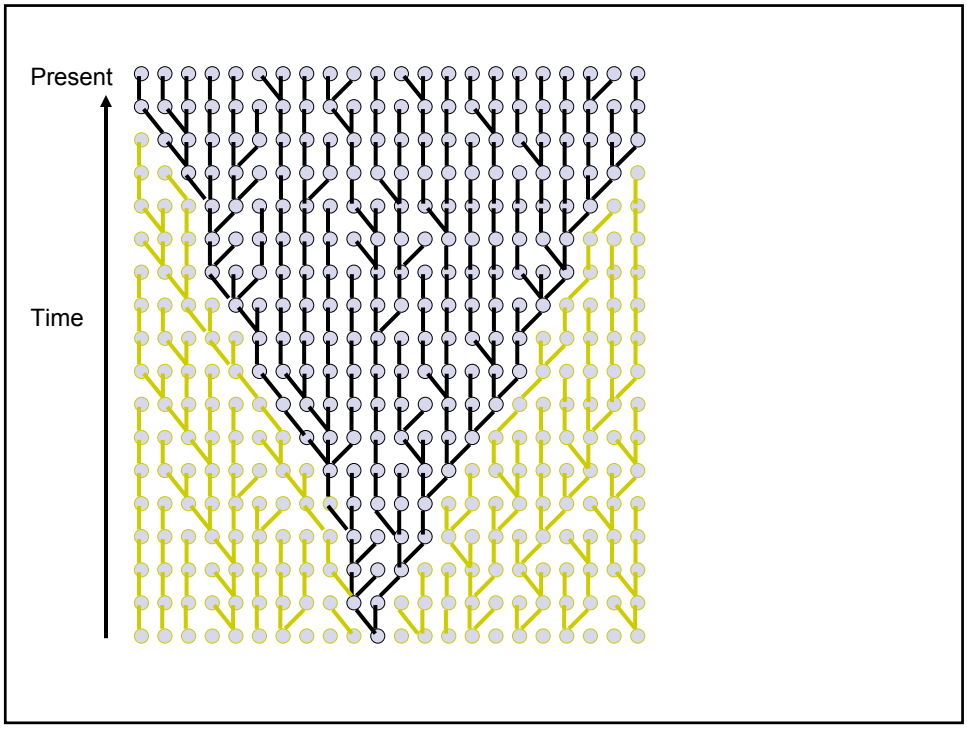

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors
3 ancestors

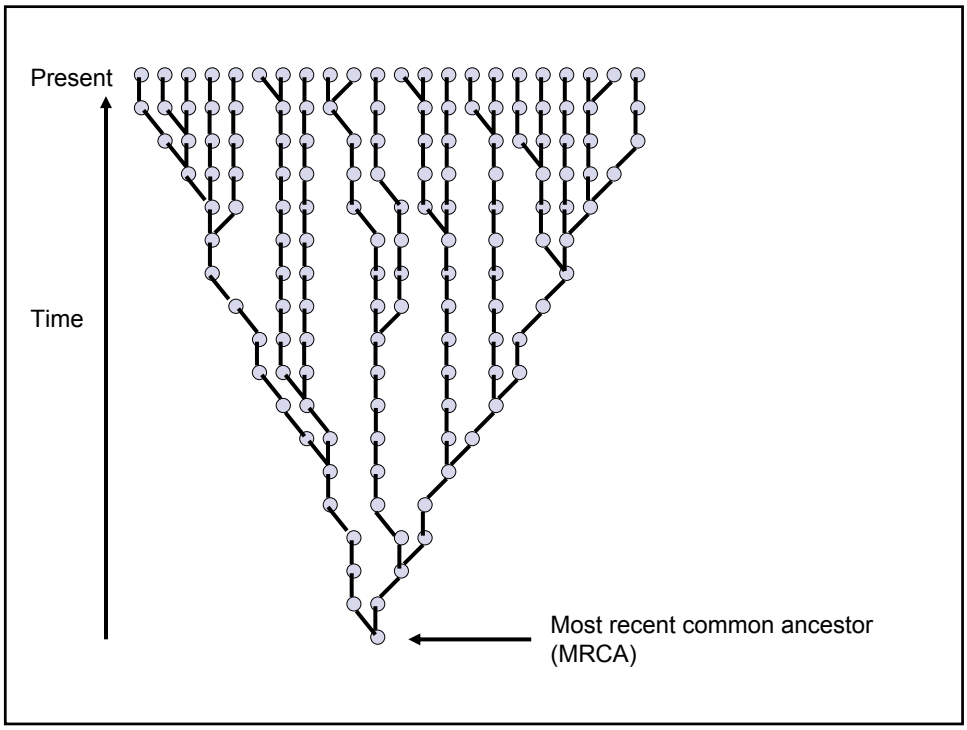

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors
3 ancestors
3 ancestors


Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors
3 ancestors
3 ancestors
2 ancestors

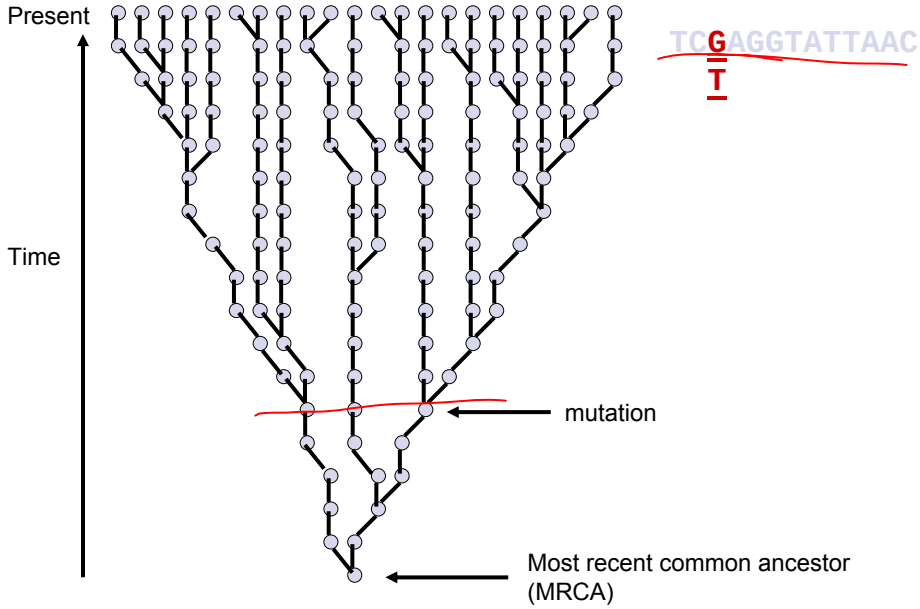

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors
3 ancestors
3 ancestors
2 ancestors
2 ancestors

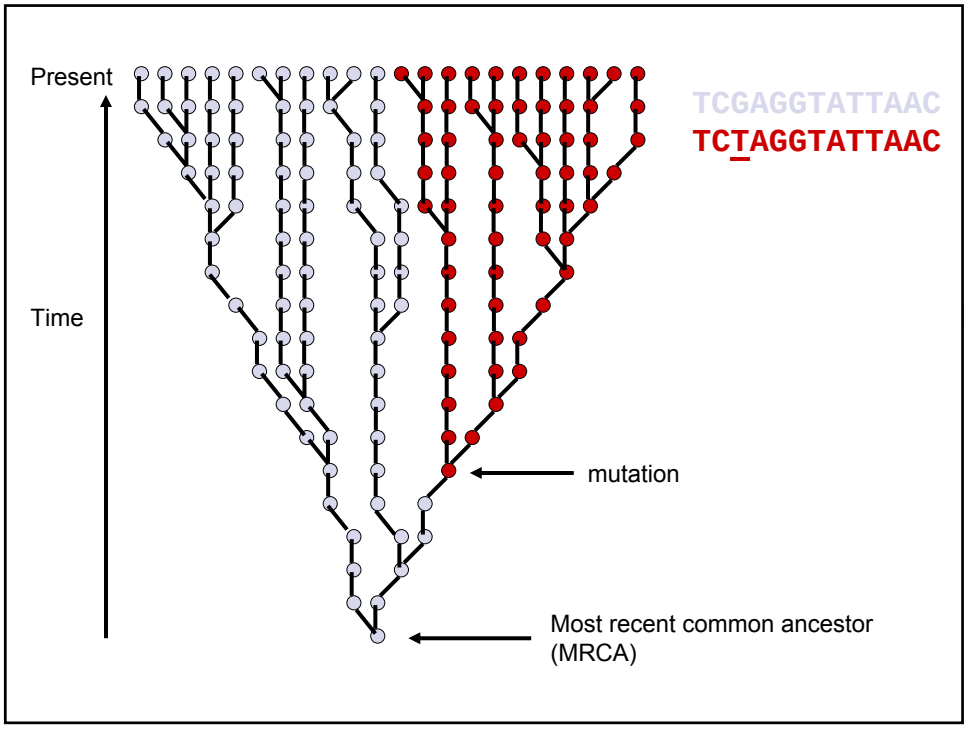

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors
3 ancestors
3 ancestors
2 ancestors
2 ancestors
1 ancestor

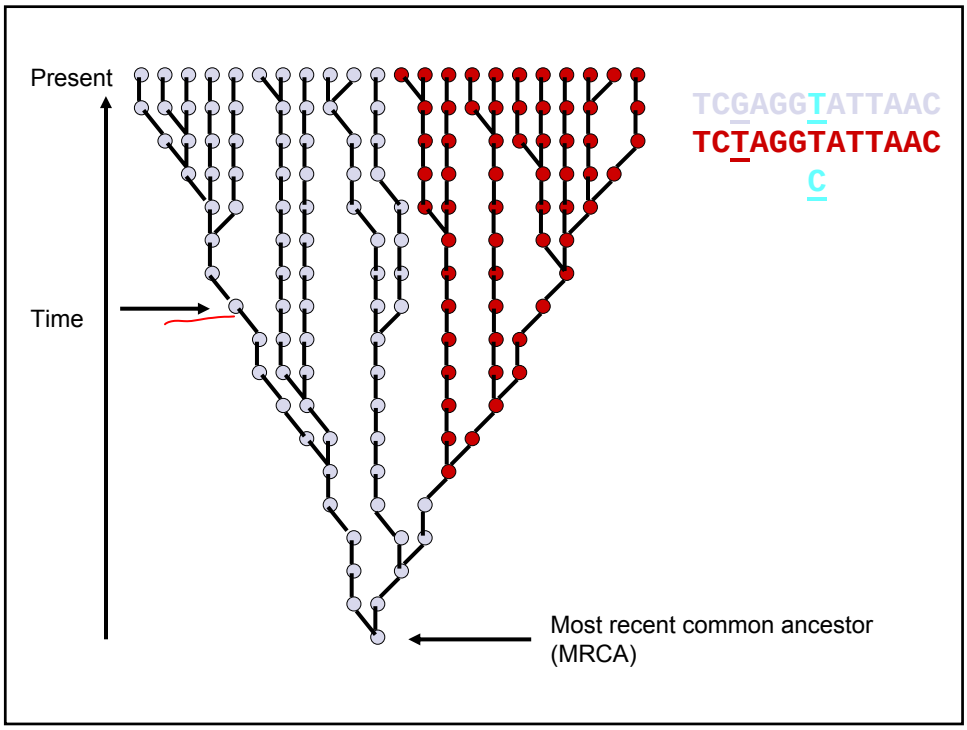

Present
Time

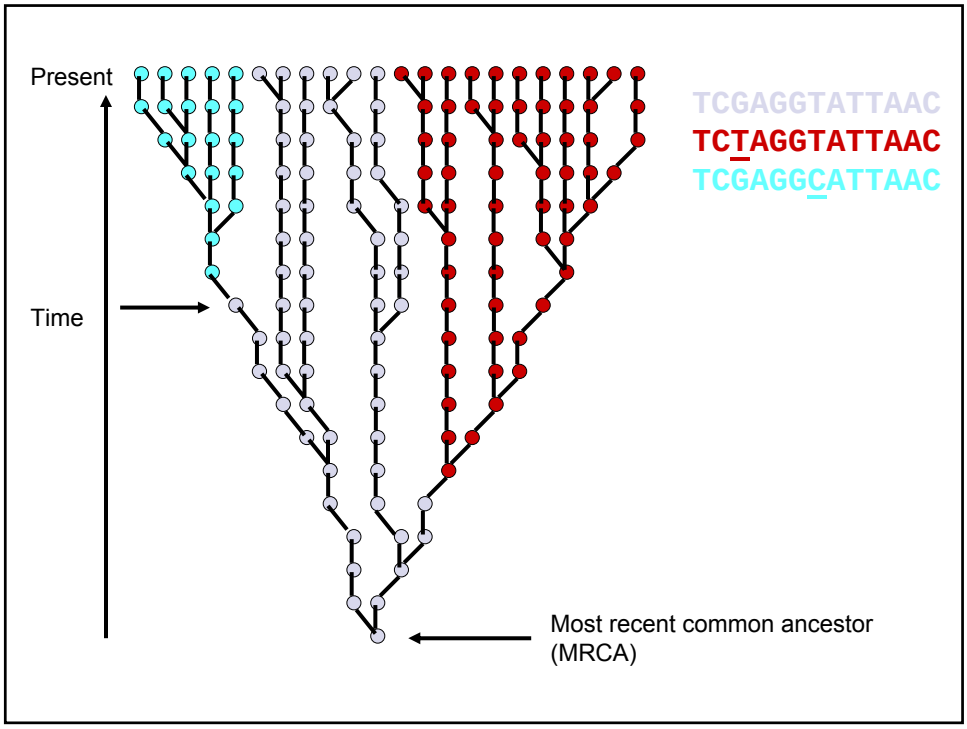

Present
Time
Most recent common ancestor
(MRCA)

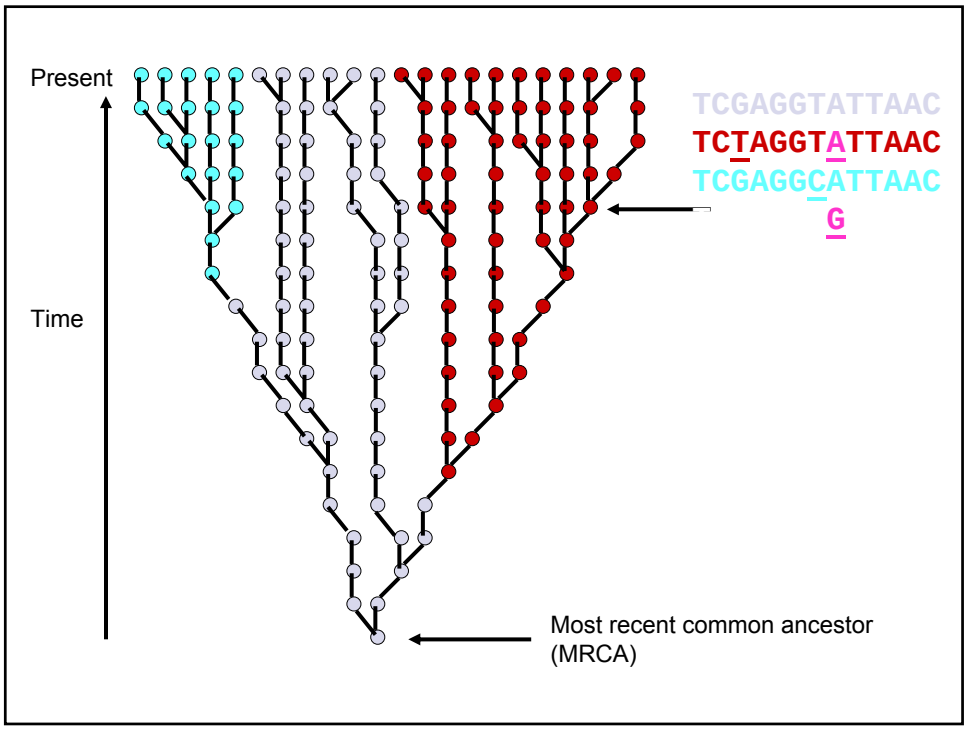
- Mutational events can now be added to the genealogical tree, resulting in **polymorphic** sites. If these sites are typed in the modern sample, they can be used to split the sample into **sub-clades** (represented by different colors)

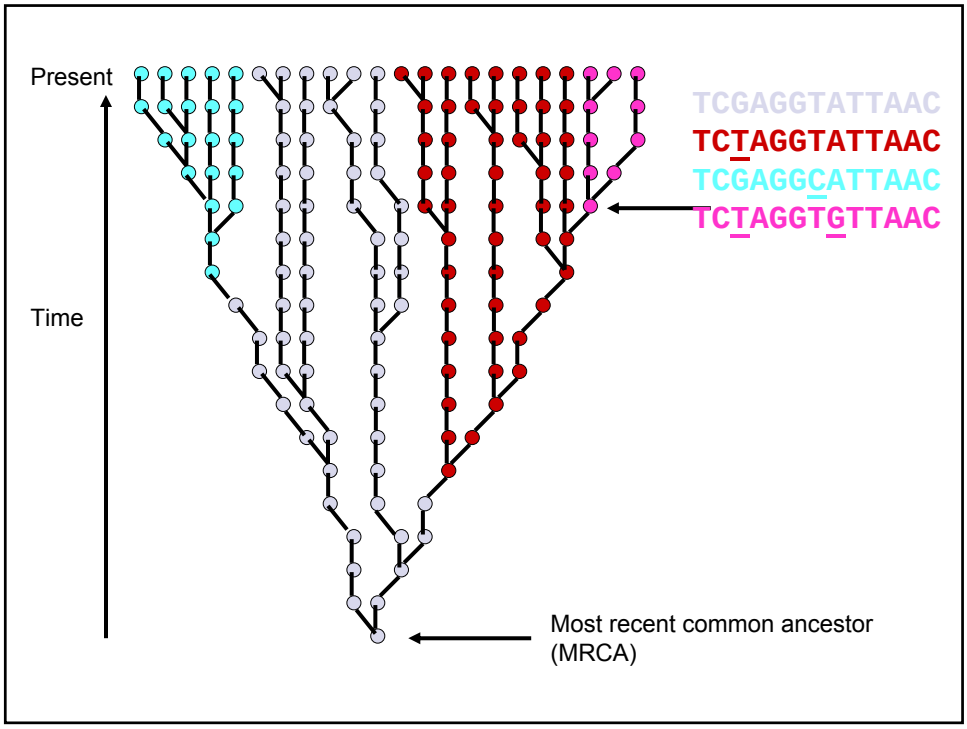


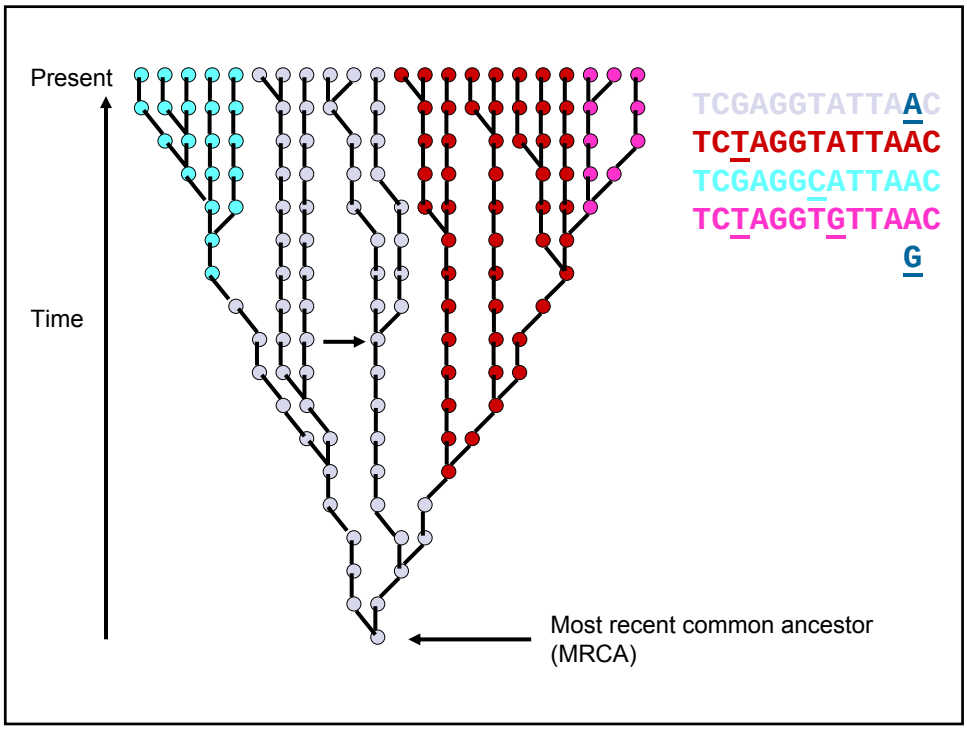

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
mutation
Most recent common ancestor
(MRCA)

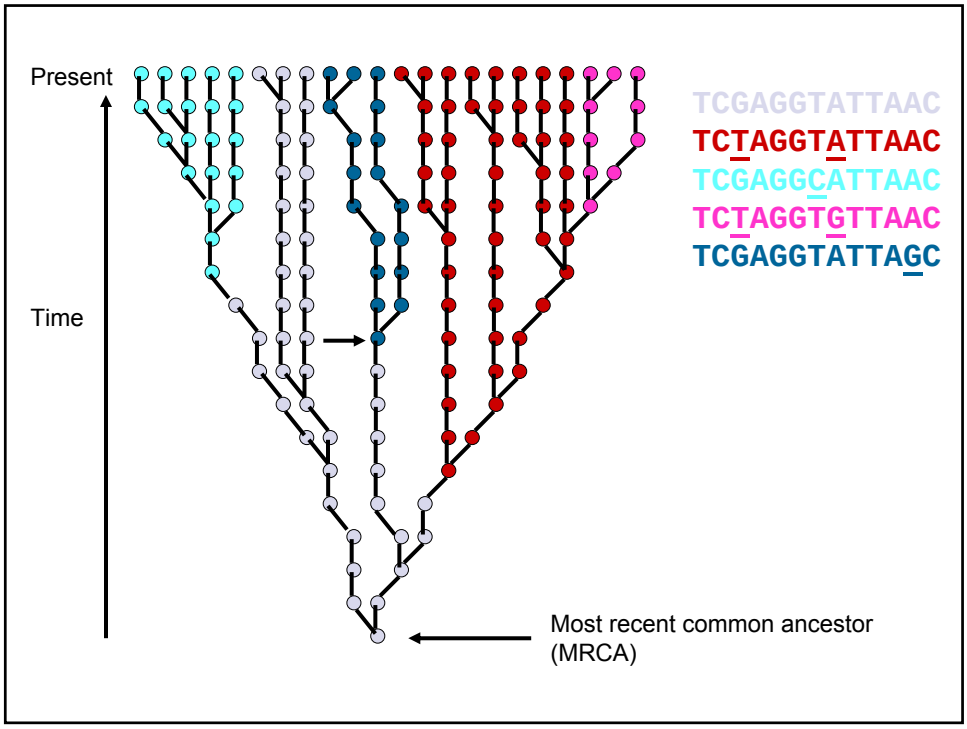

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
C
Most recent common ancestor
(MRCA)

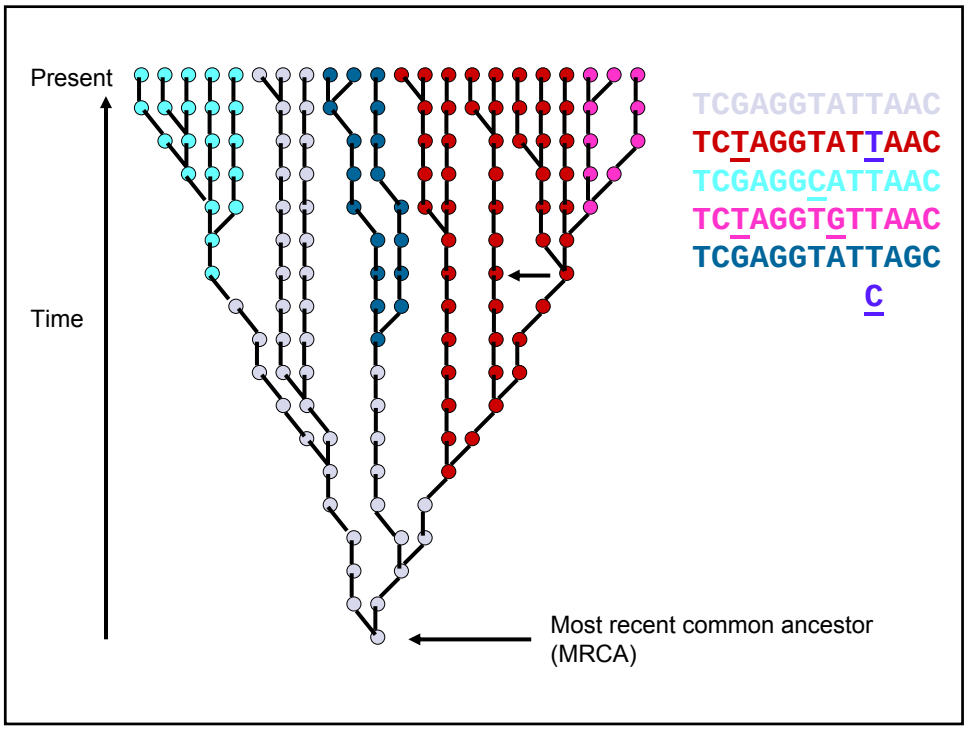

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
Most recent common ancestor
(MRCA)

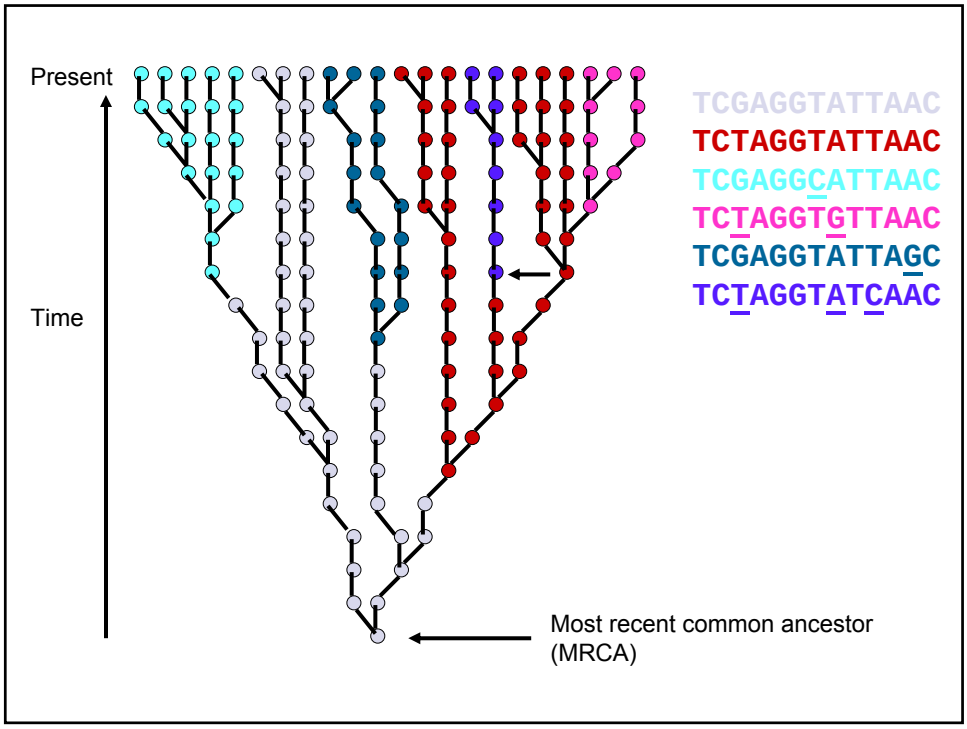

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
G
Most recent common ancestor
(MRCA)

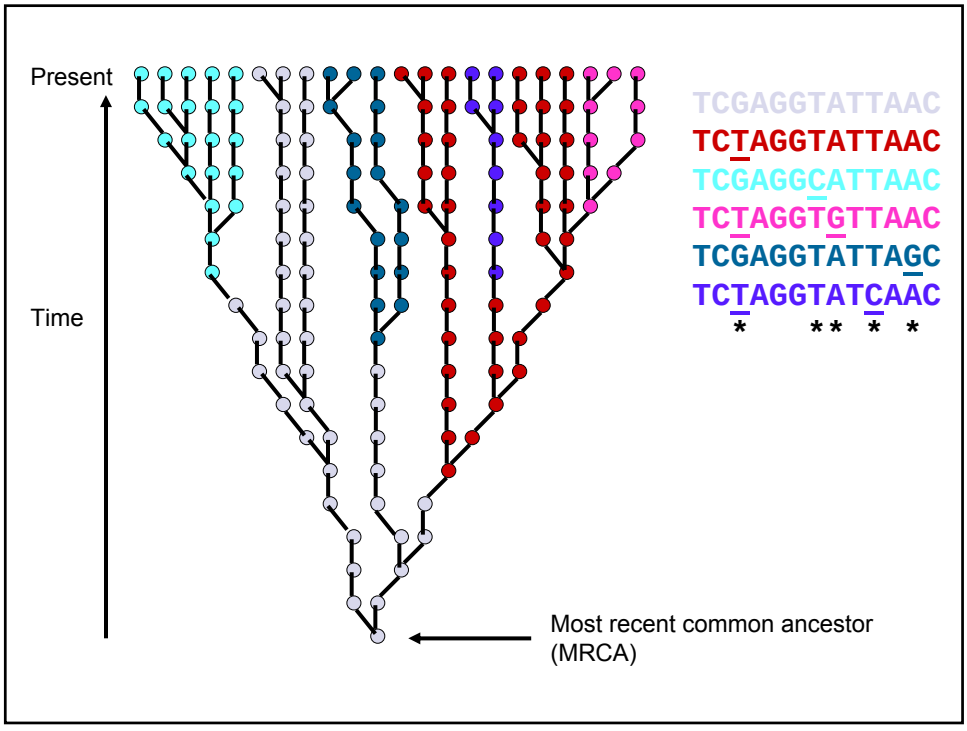

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
Most recent common ancestor
(MRCA)


Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
A
G
Most recent common ancestor
(MRCA)


Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
Most recent common ancestor
(MRCA)


Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
C
Most recent common ancestor
(MRCA)


Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
TCTAGGTATCAAC
Most recent common ancestor
(MRCA)


Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
TCTAGGTATCAAC
* ** * *
Most recent common ancestor
(MRCA)

The Statistical Models

- To move beyond mere description, and to attempt such things as estimating the TMRCA (Time to Most Recent Common Ancestor) of the tree, it is necessary to adopt certain modeling assumptions.
- For now lets forget about mutations, but just concern ourselves with the coalescence

Kingman's coalescent process

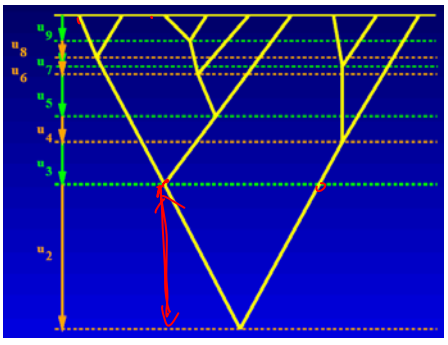
- Random collision of lineages as go back in time
- Collision is faster the smaller the effective population size
  - In a haplotype population of effective population size N,

Average time for k copies to coalesce to k-1 copies:

$= \frac{2N}{k(k-1)}$ **generations**

Average time for two copies to coalesce:

= N generations



Average time for n copies to coalesce:

$= 2N\left(1 - \frac{1}{n}\right)$ **generations**

Derivation? ---- Hw!

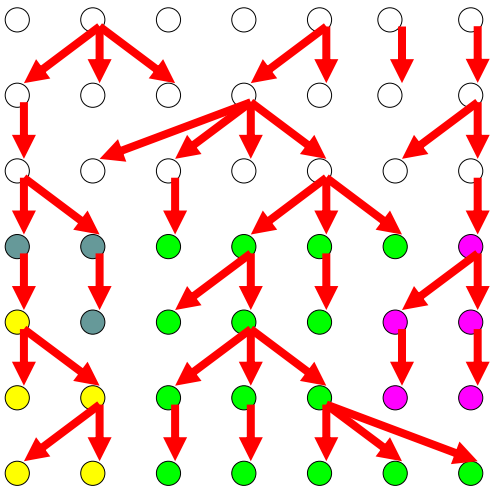
Hint of the derivation

The Wright-Fisher (WF) model

- The coalescent is descriptive, but not generative!
- A classic generative model is the **Wright-Fisher model**. This is the canonical model of genetic drift in populations. It was invented in 1932 and 1930 by Sewall Wright and R. A. Fisher.
- It starts with the following assumptions:
  - random mating and a random number of offspring (strictly, following a Poisson distribution)
  - no recombination (i.e. a single locus),
  - constant population size,
  - no selection,

The Wright-Fisher (WF) model

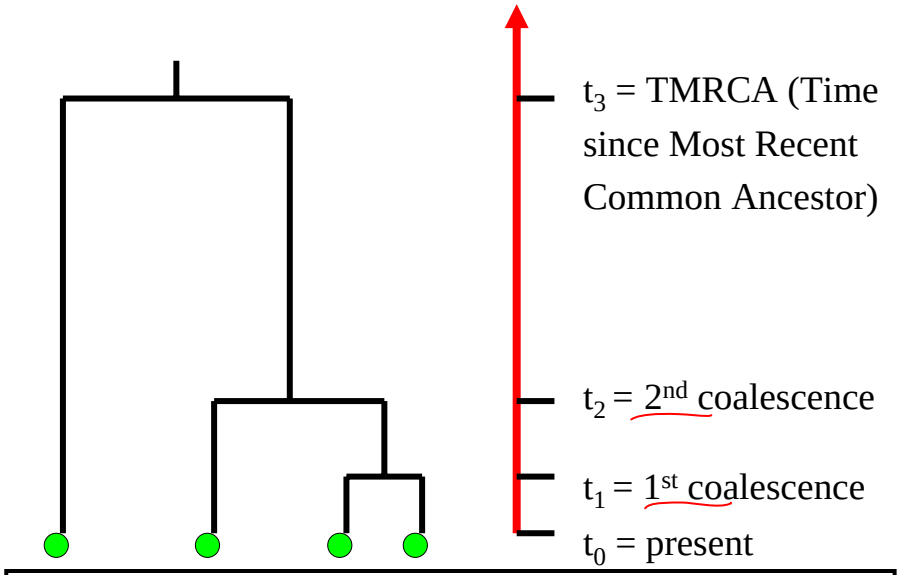
- It is a forwards-in-time model of a **neutral** locus in a constant-size, random-mating, haploid population evolving in **discrete** generations.
- Each individual in generation *t* has a random number (possibly 0) of offspring in generation *t+1*. Each is:
  - identical to the parent with probability 1-μ;
  - otherwise a mutation occurs.
- With WF, one can attempt such things as estimating the TMRCA (Time to Most Recent Common Ancestor) of the tree, etc.



Coalescent theory

When we consider the same set of assumptions but now simulate going "backwards in time", we arrive at the standard coalescent model with *infinite-allele-mutations*.

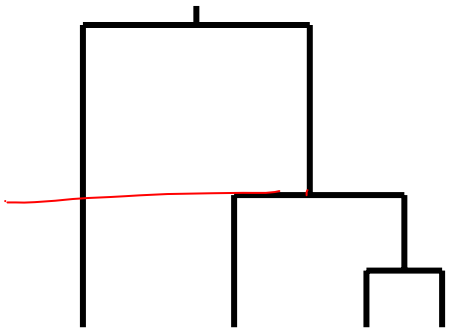
- A coalescent is the backwards-in-time "cousin" of the WF model: similar assumptions, but traces the ancestry of n observed alleles.
- Ancestry is represented via a genealogical tree: leaves are observed alleles, root is the most recent common ancestor (MRCA).



Time is measured in units of N generations: 1 coalescent time unit = NG years, where G is generation time in years.

Time back to the next coalescence when there are k lineages has the exponential distribution with mean and standard deviation both $2/k(k\text{-}1)$;

e.g. $k = 4$:

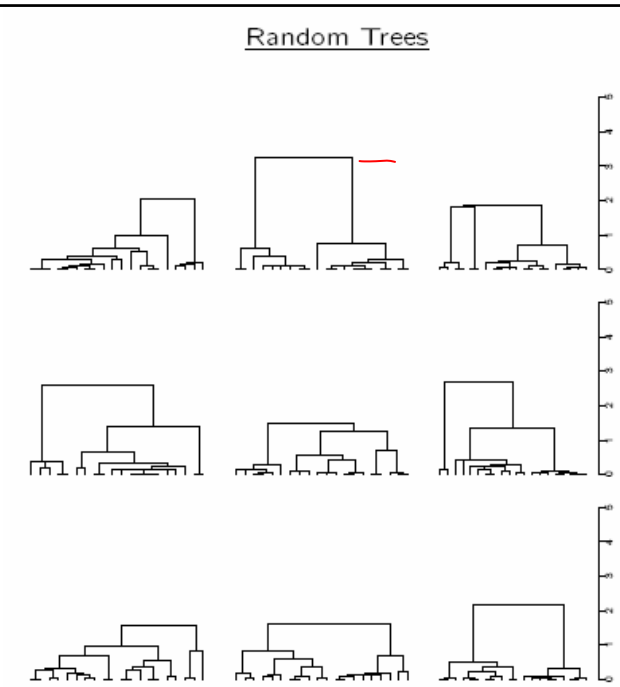


Height of tree: mean = 3/2 sd = 1·07

Total branch length: mean =11/3 sd = 2·33

The TMRCA under the coalescent

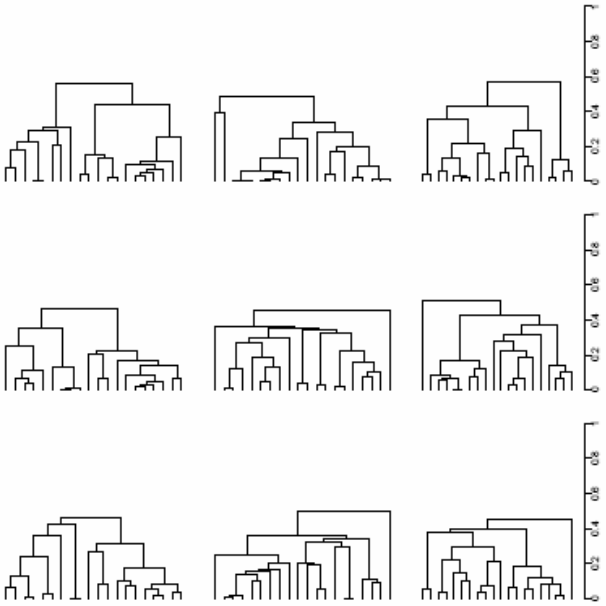
- The TMRCA (height of the genealogical tree) is on average $2(n\text{-}1)/n$; the average time in which there are just two ancestral lineages is 1.
  - the number of ancestors of a sample drops rapidly (backwards in time);
  - for more than half its history, on average, a sample has only two ancestors;
  - data often clustered.
- When we simulate from the standard coalescent, we find that there is considerable variation in the TMRCA from one simulation to the next.
  - Most coalescent event occur in the recent past (at the tips of the tree)



Coalescent with variable population size

- The situation changes if we expand the coalescent model to incorporate a factor of **exponential population growth**.
- Now there is less variation in the TMRCA between simulations, and more coalescent events occur in the more distant past (near the root of the tree).

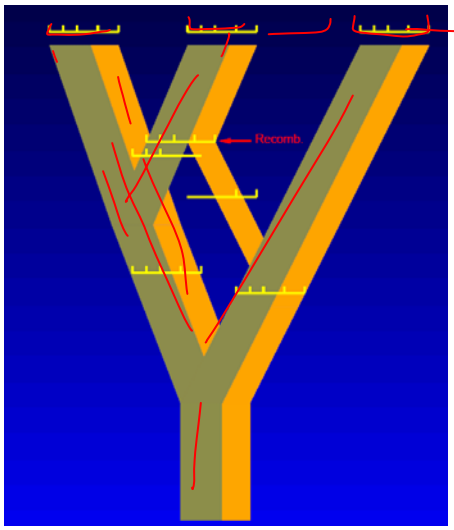
Exponentially Growing Populations



Generalisations of the standard coalescent model

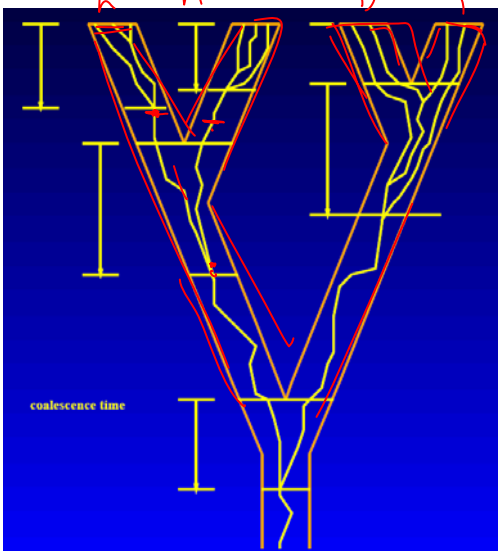
- Variable population size: coalescences occur more rapidly when the population size is small.
- Population subdivision with migration.
- Some forms of selection.
- Recombination: the ancestral recombination graph (ARG)

A recombining coalescent



Different markers have slightly different coalescent trees

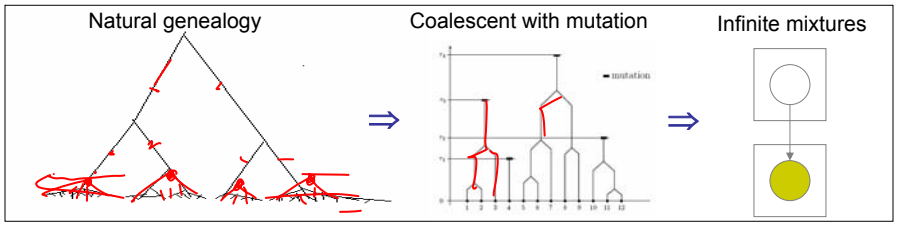
Coalescents in related species

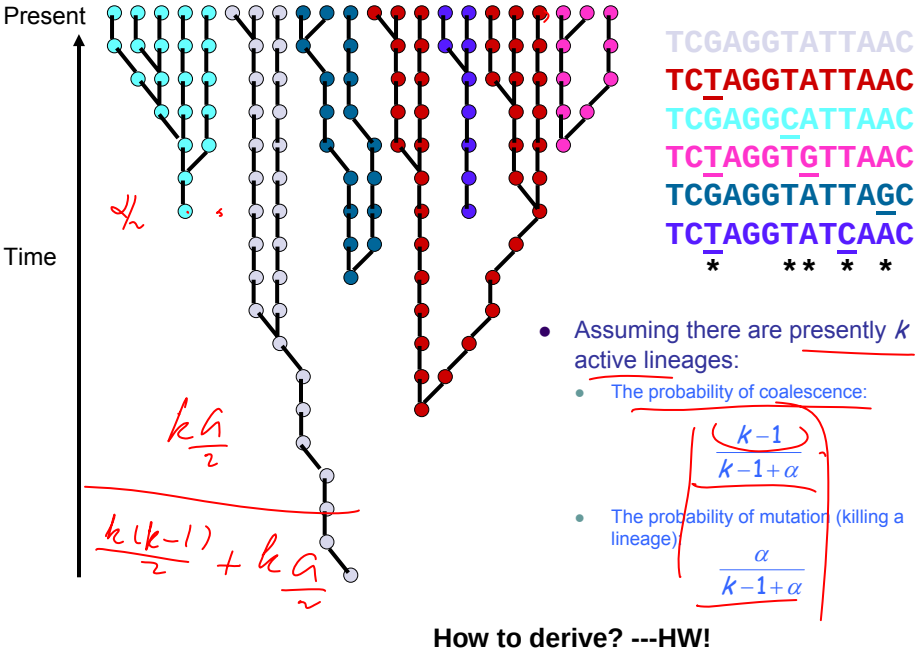


Consistency of gene tree with species tree

How to approximate a coalescent?

- Kingman coalescent process with binary lineage merging
- New population haplotype alleles emerge along all branches of the coalescence tree at rate a/2 per unit length
- This can be approximated by an infinite mixture model (aks, Dirichlet process mixture)

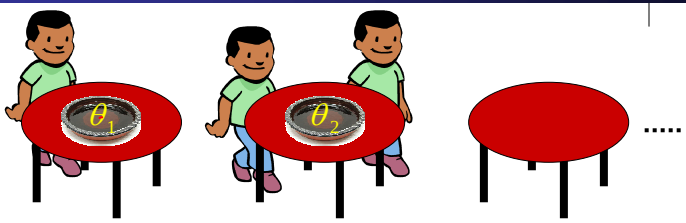




- Assuming there are presently k active lineages:
  - The probability of coalescence: $\frac{k-1}{k-1+\alpha}$
  - The probability of mutation (killing a lineage) $\frac{\alpha}{k-1+\alpha}$

How to derive? ---HW!

Chinese Restaurant Process

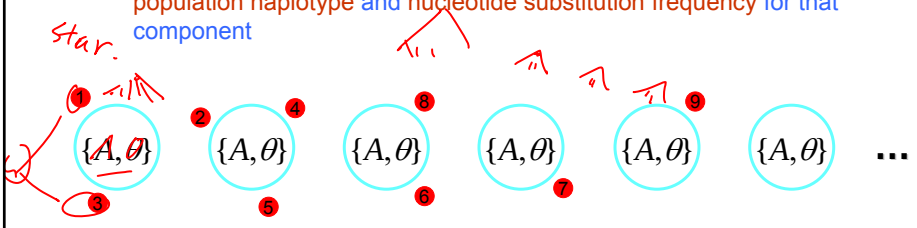


$$P(c_i = k \mid \mathbf{c}_{-i}) = \begin{matrix} 1 \\ \frac{1}{1+\alpha} \\ \frac{1}{2+\alpha} \\ \frac{1}{3+\alpha} \\ \frac{m_1}{i+\alpha-1} \end{matrix} \quad \begin{matrix} 0 \\ \frac{\alpha}{1+\alpha} \\ \frac{1}{2+\alpha} \\ \frac{2}{3+\alpha} \\ \frac{m_2}{i+\alpha-1} \end{matrix} \quad \cdots \quad \begin{matrix} 0 \\ 0 \\ \frac{\alpha}{2+\alpha} \\ \frac{\alpha}{3+\alpha} \\ \frac{\alpha}{i+\alpha-1} \end{matrix}$$

CRP defines an exchangeable distribution on partitions over an (infinite) sequence of samples, such a distribution is formally known as the Dirichlet Process (DP)

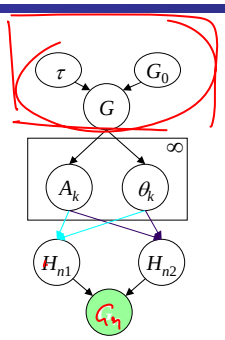
The DP Mixture of Ancestral Haplotypes

- The customers around a table form a cluster
  - associate a mixture component (*i.e.*, a population haplotype) with a table
  - sample $\{a, \theta\}$ at each table from a base measure G_0 to obtain the population haplotype and nucleotide substitution frequency for that component



  - With $p(h|\{A, \theta\})$ and $p(g|h_1,h_2)$, the CRP yields a posterior distribution on the number of population haplotypes (and on the haplotype configurations and the nucleotide substitution frequencies)

A Hierarchical Bayesian Infinite Allele model



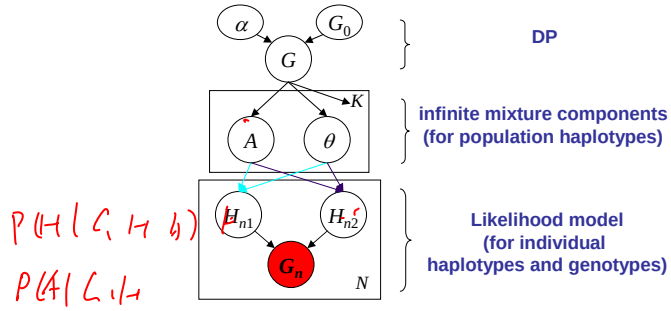
- Assume an individual haplotype h is stochastically derived from a population haplotype a_k with nucleotide-substitution frequency θ_k:

$$h \sim p(h \mid \{a, \theta\}_k).$$

- Not knowing the correspondences between individual and population haplotypes, each individual haplotype is a mixture of population haplotypes.

- The number and identity of the population haplotypes are unknown
  - use a Dirichlet Process to construct a prior distribution G on $\mathcal{H} \times \mathcal{R}^J$.

DP-haplotyper



- Inference: Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling
  - Metropolis Hasting

Model components

- Choice of base measure:

$$G_0 \sim \text{Unif}(a) \cdot \prod_j \text{Beta}(\theta_j)$$

- Nucleotide-substitution model:

$$p(h_i \mid \{a,\theta\}_k) = \prod_j p(h_{i,j} \mid a_{k,j}, \theta_{k,j})$$

$$\text{where} \quad p(h_{i,j} \mid a_{k,j}, \theta_{k,j}) = \begin{cases} \theta_{k,j} & \text{if} \quad h_{i,j} = a_{k,j} \\ 1-\theta_{k,j} & \text{if} \quad h_{i,j} = a_{k,j} \end{cases}$$

- Noisy genotyping model:

$$p(g_i \mid h_{i_1}, h_{i_2}) = \prod_j p(g_{i,j} \mid h_{i_1,j}, h_{i_2,j})$$

$$\text{where} \quad p(g_{i,j} \mid h_{i_1,j}, h_{i_2,j}) = \begin{cases} \gamma & \text{if} \quad h_{i_1,j} \oplus h_{i_2,j} = g_{i,j} \\ \dfrac{1-\gamma}{2} & \text{if} \quad h_{i_1,j} \oplus h_{i_2,j} \neq g_{i,j} \end{cases}$$

Gibbs sampling

Starting from some initial haplotype reconstruction $H^{(0)}$, pick a first table with an arbitrary $a_1^{(0)}$, and form initial population-hap pool $\mathbf{A}^{(0)} = \{a_1^{(0)}\}$:

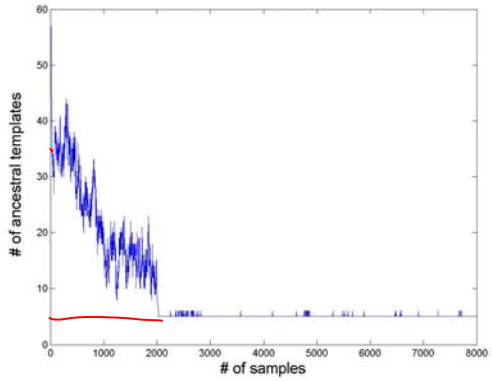
i) Choose an individual *i* and one of his/her two haplytopes *t*, uniformly and at random, from all ambiguous individuals;

ii) Sample $c_{i_t}^{(t+1)}$ from $p(c_{i_t}^{(t+1)} \mid c_{-i_t}^{(t)}, H^{(t)}, \mathbf{A}^{(t)})$, update $c^{(t+1)}$;

iii) Sample $a_k^{(t+1)}$, where $k = c_{i_t}^{(t+1)}$, from $p(a_k^{(t+1)} \mid \forall h_{-i'_{t'}}^{(t)} \text{ s.t. } c_{i'_{t'}}^{(t+1)} = k)$; update $\mathbf{A}^{(t+1)}$;

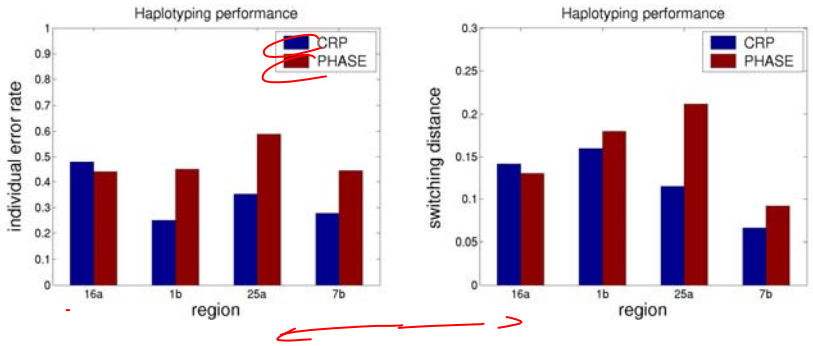
iii) Sample $h_{i_t}^{(t+1)}$ from $p(h_{i_t}^{(t+1)} \mid c_{i_t}^{(t+1)}, H_{-i_t}^{(t)}, \mathbf{A}^{(t+1)})$, update $H^{(t+1)}$.

Convergence of Ancestral Inference

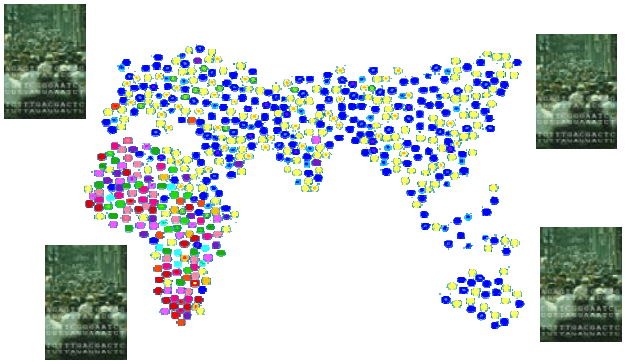


Haplotyping Error

The Gabriel data

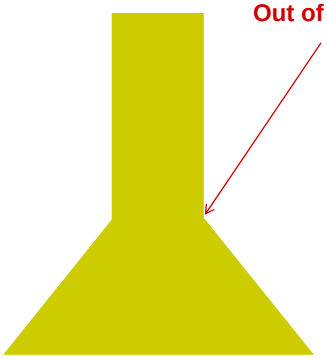


Multi-population Genetic Gemography

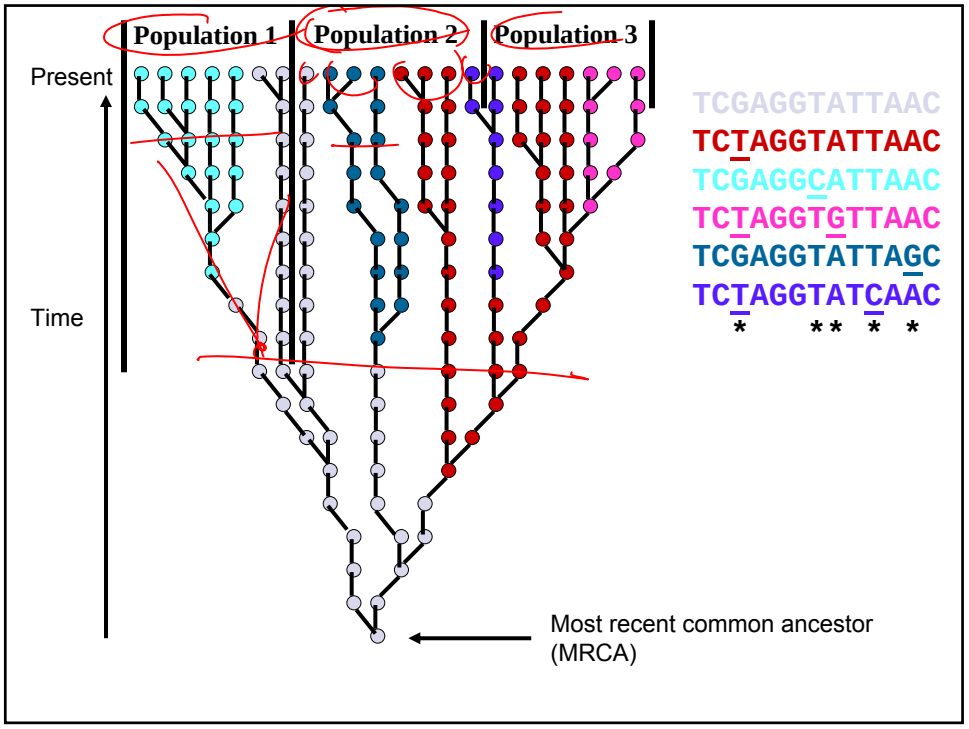


- Pool everything together and solve 1 hap problem?
  - --- ignore population structures
- Solve 4 hap problems separately?
  - --- data fragmentation
- Co-clustering … solve 4 *coupled* hap problems jointly

Why humans are so similar and polymorphic patterns are regional

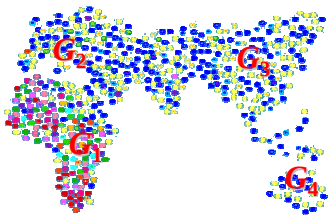


- Population bottleneck: a small population that interbred reduced the genetic variation
- Out of Africa ~ 100,000 years ago

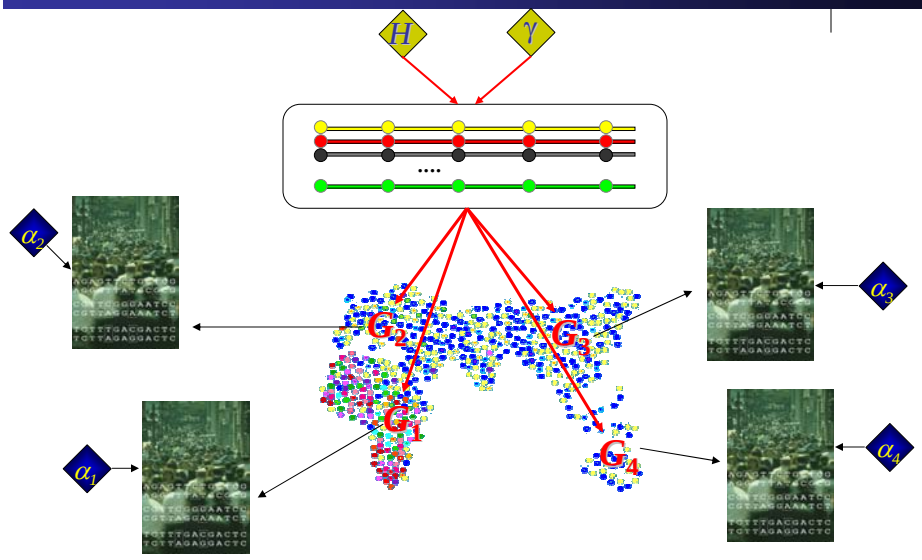


Population Specific DPs

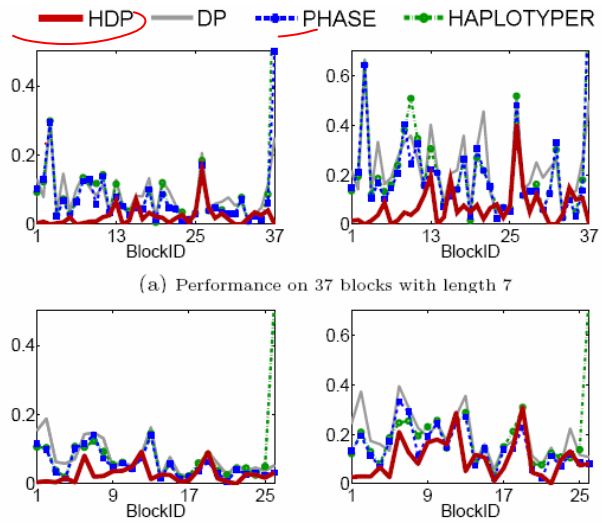
- Each population can be associated with a unique DP capturing population-specific genetic demography
- Different population may have unique haplotypes
- Different population may share common haplotypes
- Thus Population specific DPs are marginally dependent



Hierarchical DP Mixture



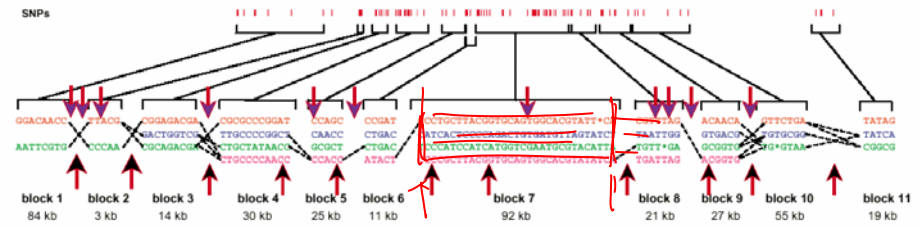
HapMap Data



(a) Performance on 37 blocks with length 7

The block structure of haplotypes

- The Daly et al (2001) data set
  - This consists of 103 **common SNPs** (>5% minor allele frequency) in a 500 kb region implicated in Crohn disease, genotyped in 129 trios (mom, pop, kid) from a European derived population, giving 258 transmitted and 258 untransmitted chromosomes.



  - The haplotype blocks span up to 100kb and contain 5 or more common SNPs. For example, one 84 kb block of 8 SNPs shows just two distinct haplotypes accounting for 95% of the observed chromosomes.

Another study: the Patil et al data

- The **haplotype patterns** for 20 independent globally diverse chromosomes defined by 147 common human chr 21 SNPs spanning 106 kb of genomic sequence.
  - Each **row** represents an **SNP**.
  - Blue box = major, yellow = minor allele.
  - Each **column** represents a single **chromosome**.
- The 147 SNPs are divided into 18 **blocks** defined by **black** lines.
  - The *expanded box on the right* is an SNP block of **26** SNPs over 19kb of genomic DNA.
  - The **4** most common of **7** different haplotypes include **80%** of the chromosomes, and can be distinguished with **2** SNPs.

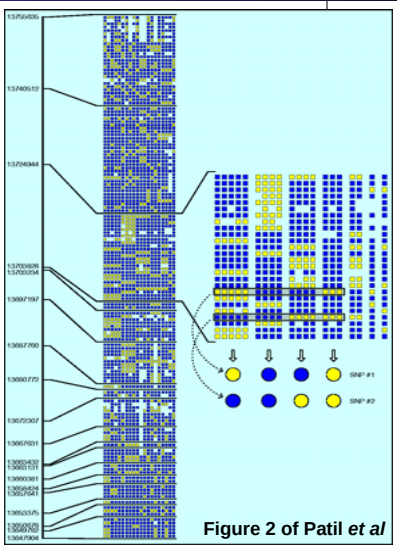
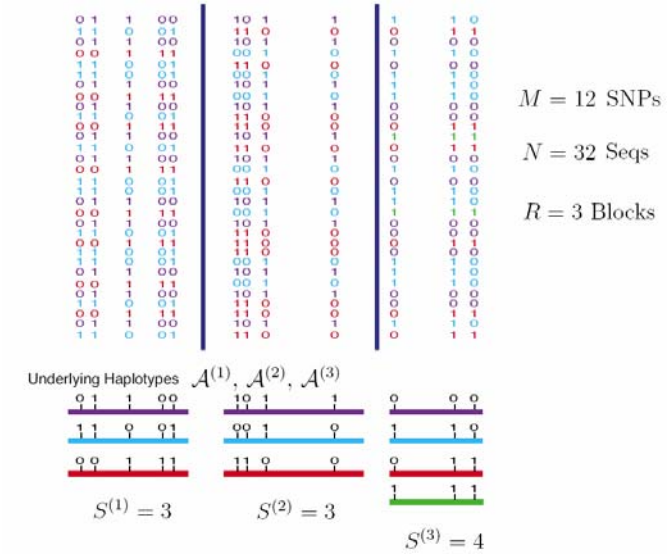
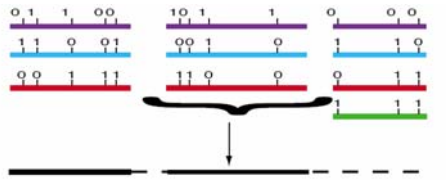


Figure 2 of Patil *et al*

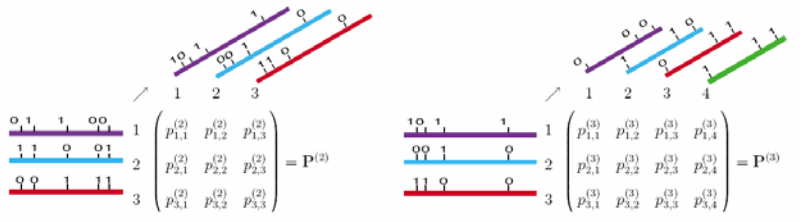
Haplotype Blocks



Markov Dependence Between Blocks



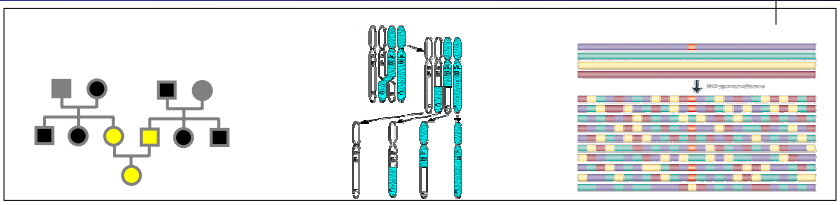
- On any chromosome, the haplotype carried at the *k*th block is drawn from $A^{(k)}$ according to unknown probabilities that depend on the haplotype at block $k - 1$.

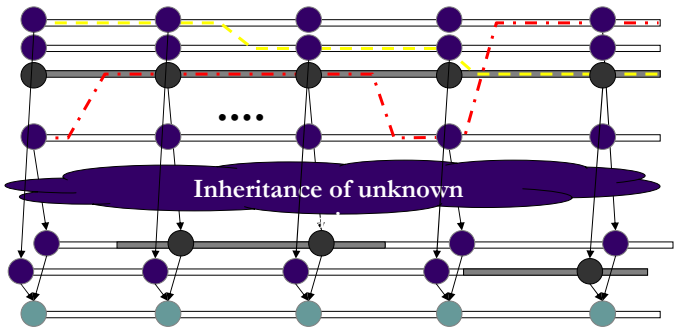


Block Partitioning Algorithms

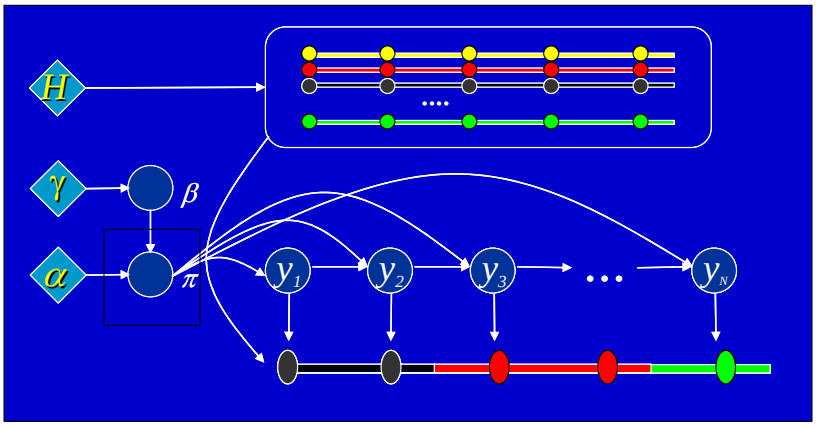
- **Greedy** algorithm (Patil *et al* 2001).
  - Begin by considering all possible blocks of ≥1 consecutive SNPs.
  - Next, exclude all blocks in which < 80% of the chromosomes in the data are defined by haplotypes represented more than once in the block (80% coverage).
  - Considering the remaining overlapping blocks simultaneously, select the one which maximizes the **ratio** of **total SNPs in the block** to the **number required to uniquely discriminate** haplotypes represented more than once in the block. Any of the remaining blocks that physically overlap with the selected block are discarded, and the process repeated until we have selected a set of contiguous, non-overlapping blocks that cover the 32.4 Mb of chr 21 2ith no gaps and with every SNP assigned to a block.
- Hidden Markov Models
  - Maximum a posterior inference (i.e., viterbi) (Daly et al. 2001)
  - Minimum description length (Anderson et al.2003)
- Dynamic programming (Sun et al. 2002)

Haplotypes are Shaped by Recombination





Hidden Markov DP for Recombination



Reference

- E.P. Xing, R. Sharan and M.I Jordan, Bayesian Haplotype Inference via the Dirichlet Process. Proceedings of the 21st International Conference on Machine Learning (ICML2004),
- E. P. Xing and K. Sohn, Hidden Markov Dirichlet Process: Modeling Genetic Recombination in Open Ancestral Space, Journal of Bayesian Analysis, 2007
- E.P. Xing, K. Sohn, M.I. Jordan and Y.W. Teh, Bayesian Multi-Population Haplotype Inference via a Hierarchical Dirichlet Process Mixture, Proceedings of the 23st International Conference on Machine Learning (ICML 2006).
- N Patil *et al* . Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21 *Science* 294 2001:1719-1723.
- M J Daly *et al* . High-resolution haplotype structure in the human genome *Nat. Genet*. 29 2001: 229-232
- Anderson, E.C., Novembre, J. (2003) "Finding haplotype block boundaries using the minimum description length principle." American Journal of Human Genetics 73(2):336-354.