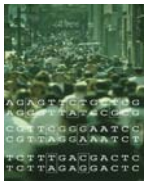
Computational Genomics

10-810/02-710, Spring 2009

Genome variation and coalescent theory

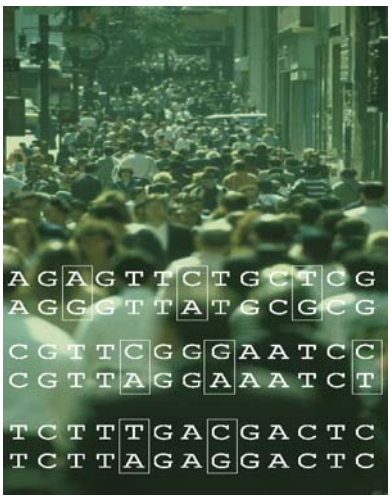


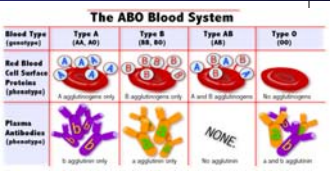
Eric Xing

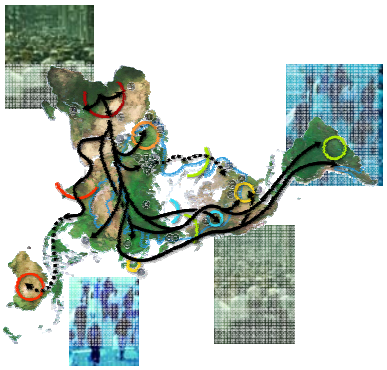
Lecture 19, March 30, 2009

Reading: DTW Chap 13, & assignment

Genome Polymorphisms





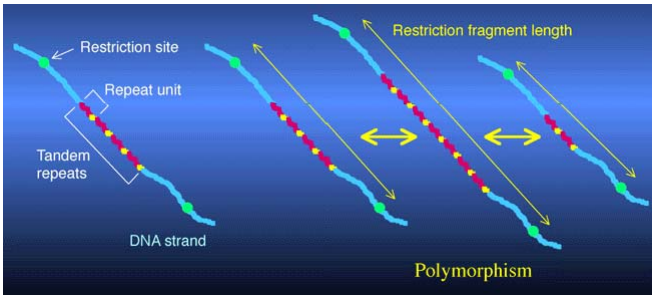


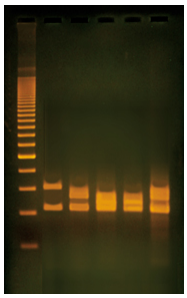
Type of polymorphisms

- Insertion/deletion of a section of DNA
  - Minisatellites: repeated base patterns (several hundred base pairs)
  - Microsatellites: 2-4 nucleotides repeated
  - Presence or absence of Alu segments
- Single base mutation (SNP)
  - Restriction fragment length (RFLP)
  - Creating restriction sites via PCR primer
  - Direct sequencing

Frequency of SNPs greater than that of any other type of polymorphism

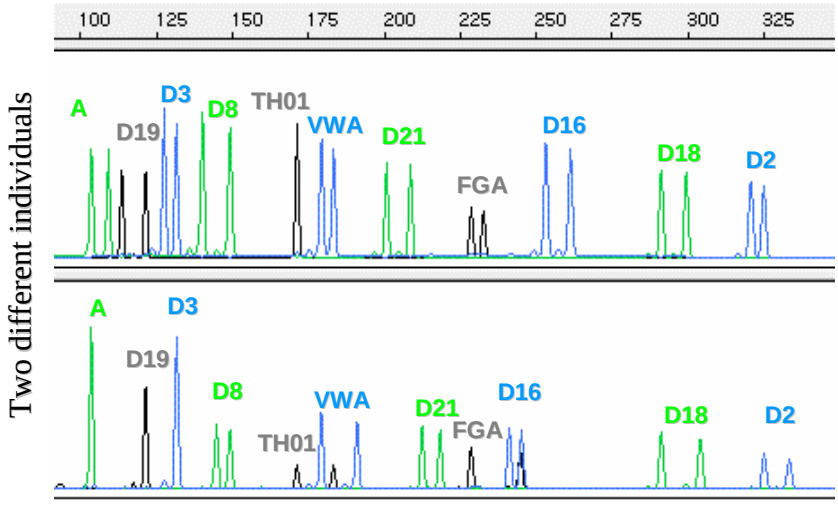
Variable Number of Tandem Repeats (VNTR) Polymorphism



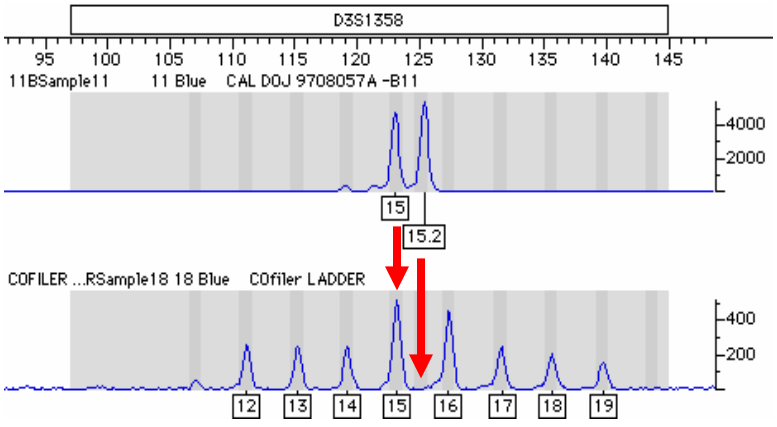


Multiplex STR Analysis

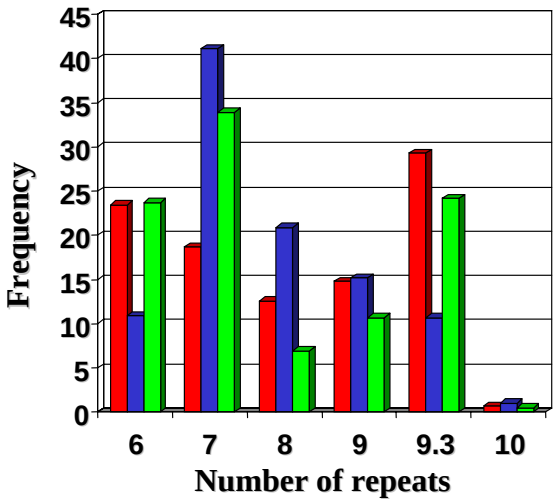
AmpFlSTR® SGM Plus™ kit



Measure against Reference Ladder



STR Allele Frequencies

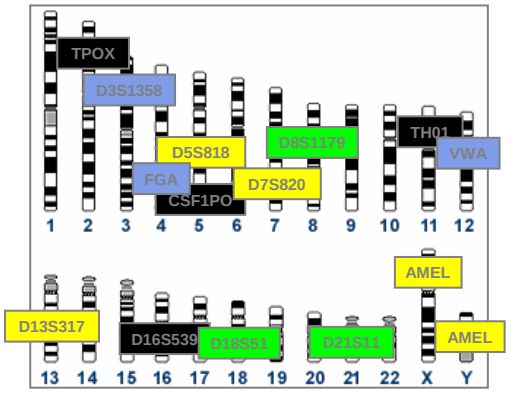


TH01 Marker

- Caucasians (N=427)
- Blacks (N=414)
- Hispanics (N=414)

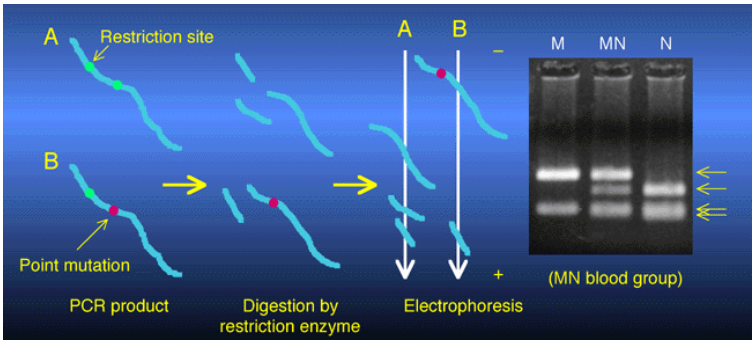
Fingerprints: Nuclear DNA standard

- FBI CODIS (Combined DNA Index System) standard for nuclear DNA utilizes 13 highly-variable tetramer STR sites.



Multiplexed groupings:
Blue, Green II,
Yellow, & Cofiler

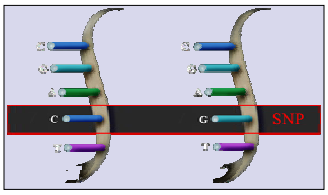
Restriction Fragment Length Polymorphism (RFLP)

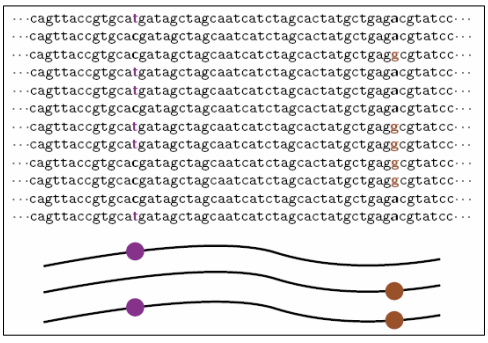


Single Nucleotide Polymorphism (SNP)

- DNA sequence variation due to differences of a single nucleotide - A, T, C, or G - among members of the species

– Each variant is called an "allele"
– Almost always bi-allelic
– Account for most of the genetic diversity among different (normal) individuals, e.g. drug response, disease susceptibility

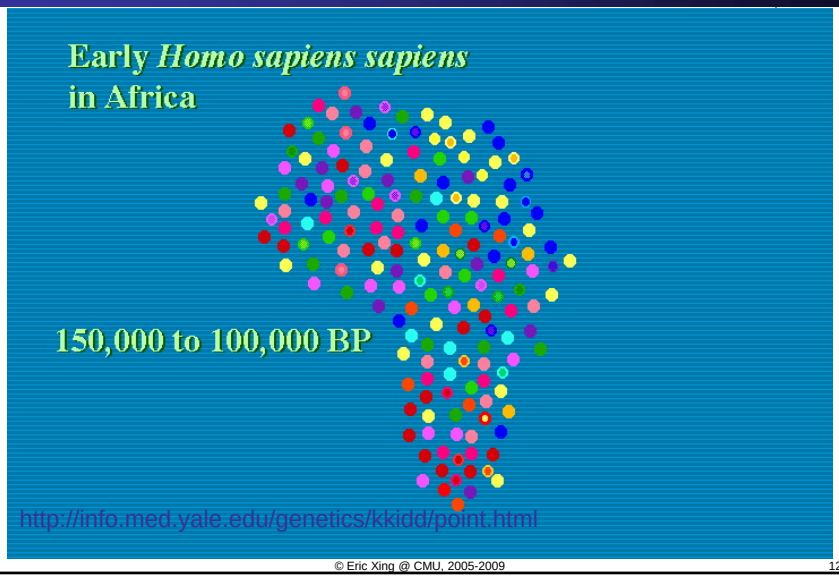




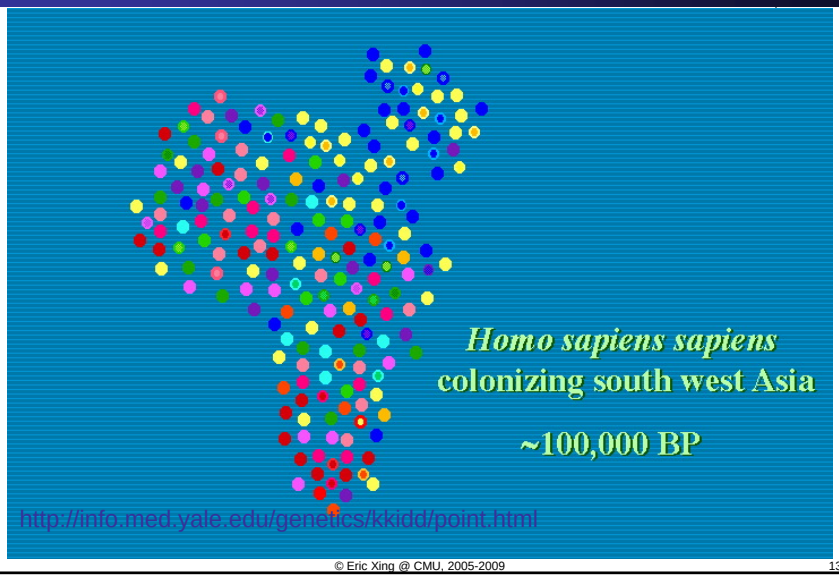
Exploiting Genetic Variations

- **Population Evolution**: the majority of human sequence variation is due to substitutions that have occurred **once** in the history of mankind at individual base pairs
  - There can be big differences between populations!
- **Markers for pinpointing a disease**: certain polymorphisms are in "Linkage Disequilibrium" with disease phenotypes
  - Association study: check for differences in SNP patterns between cases and controls
- **Forensic analysis**: the polymorphisms provide individual and familiar signatures

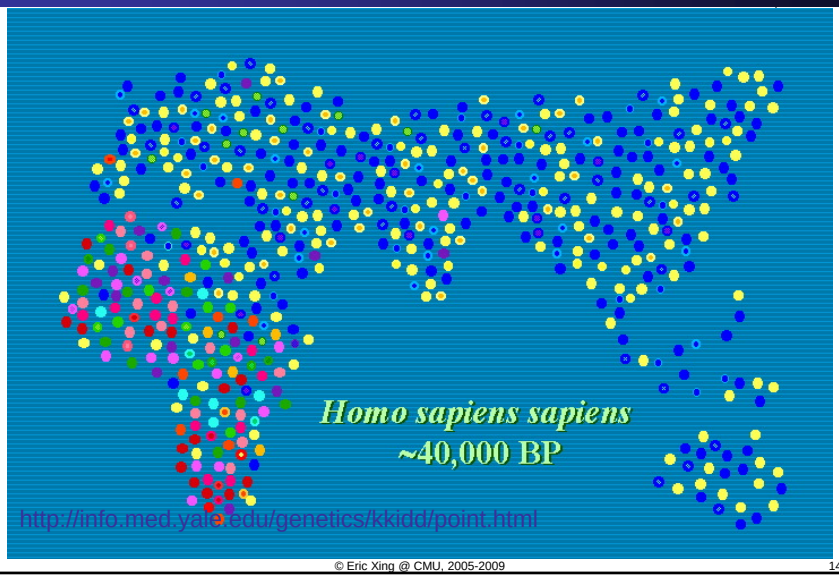
Migration of human variation



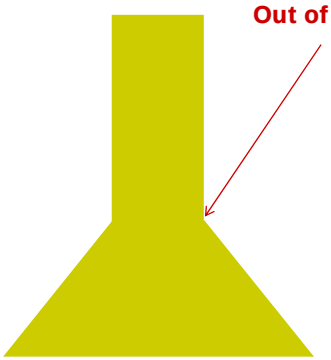
Migration of human variation



Migration of human variation

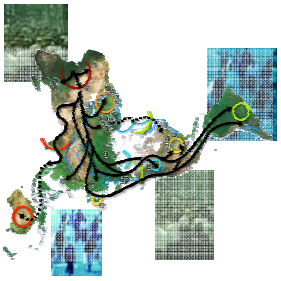


Why humans are so similar and polymorphic patterns are regional

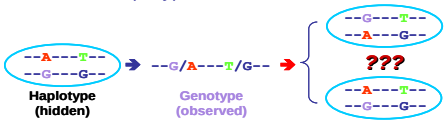


- Population bottleneck: a small population that interbred reduced the genetic variation
- Out of Africa ~ 100,000 years ago

Genetic Inference

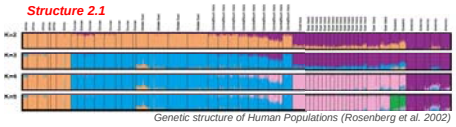


- Determine genetic markers
  - Haplotype inference

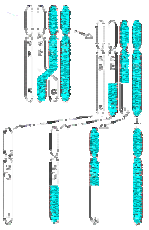


- Reveal genome inheritance events
  - Recombination hotspot identification

- Deconvolve population structure
  - Ancestral spectrum analysis



Genetic structure of Human Populations (Rosenberg et al. 2002)

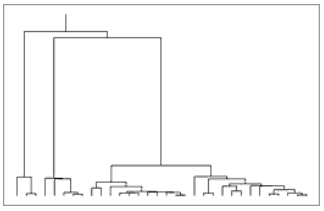


Inferring population genetics

- Size of ancestral pool
- The timings of bottlenecks and migrations
- Selections?

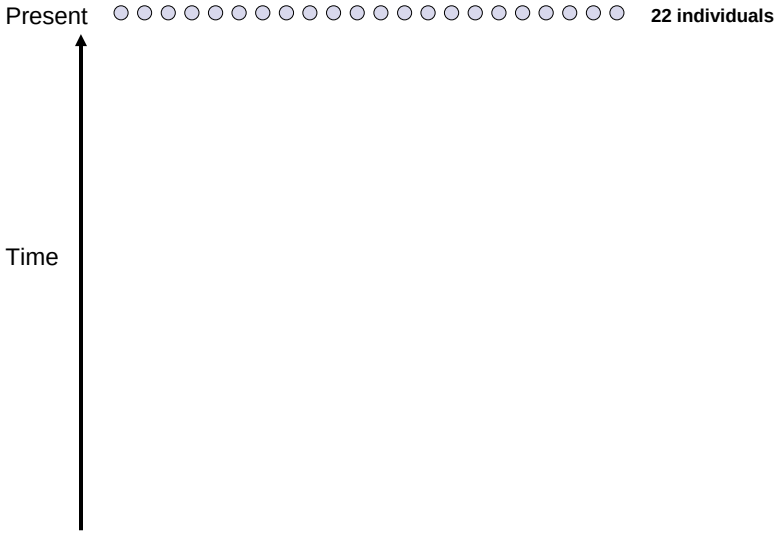
Sir John Kingman,
Head of the Isaac Newton Institute of Mathematical Sciences

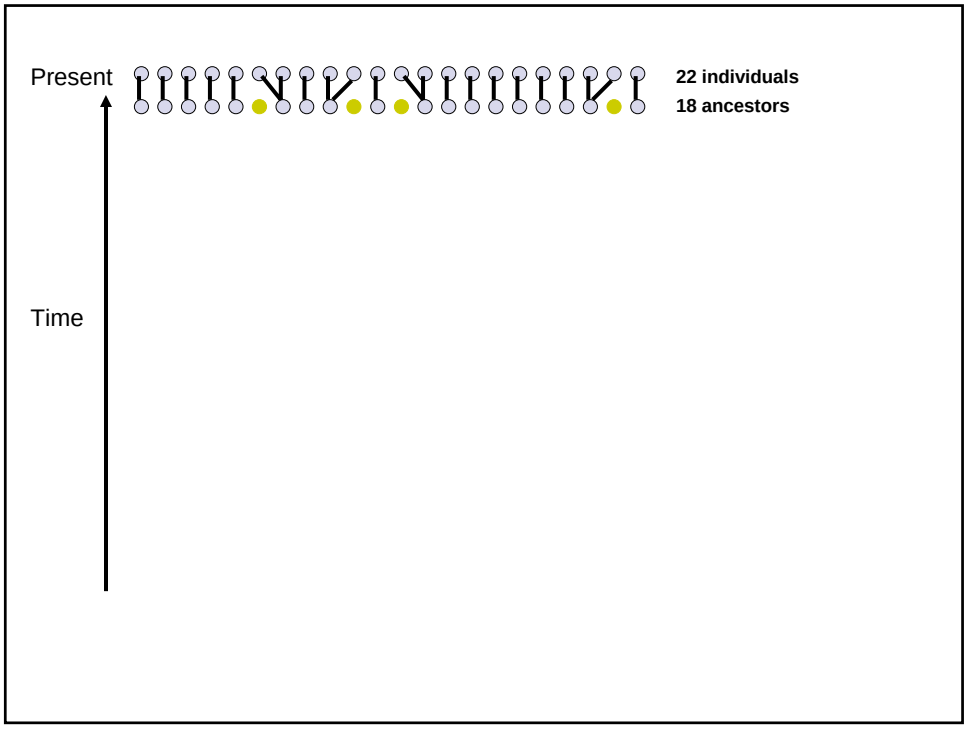
The coalescent

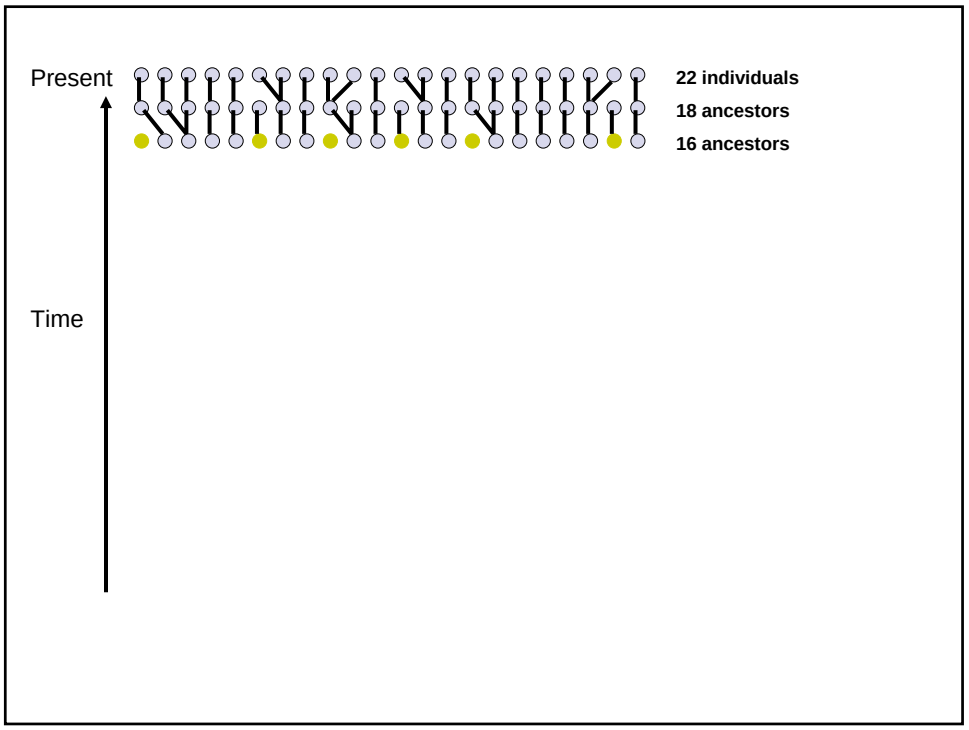


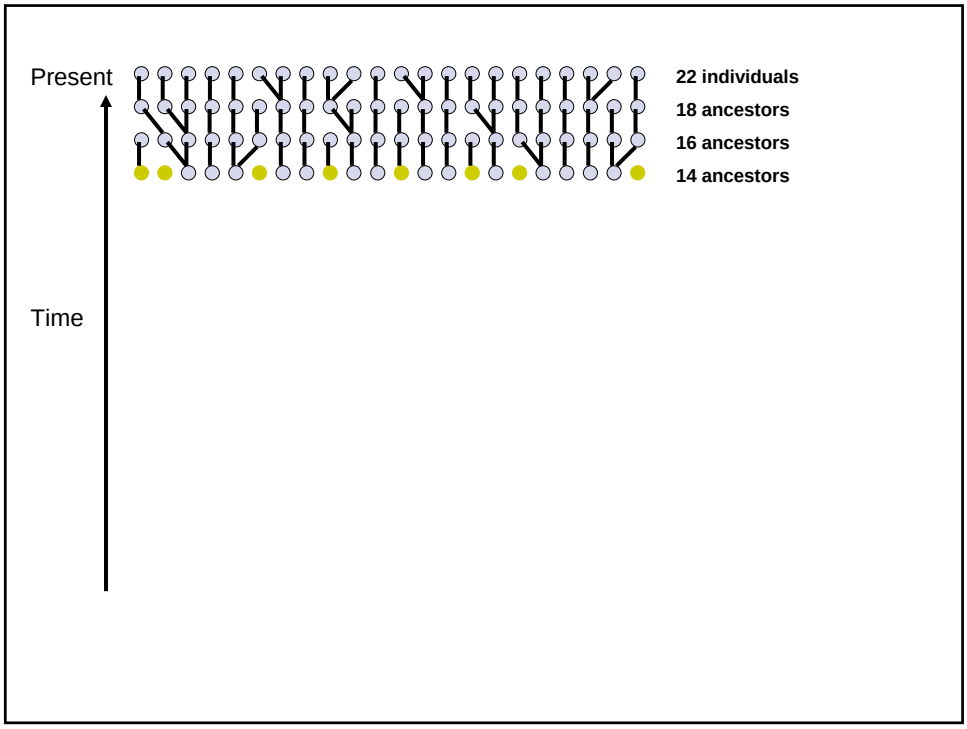
Coalescent Theory

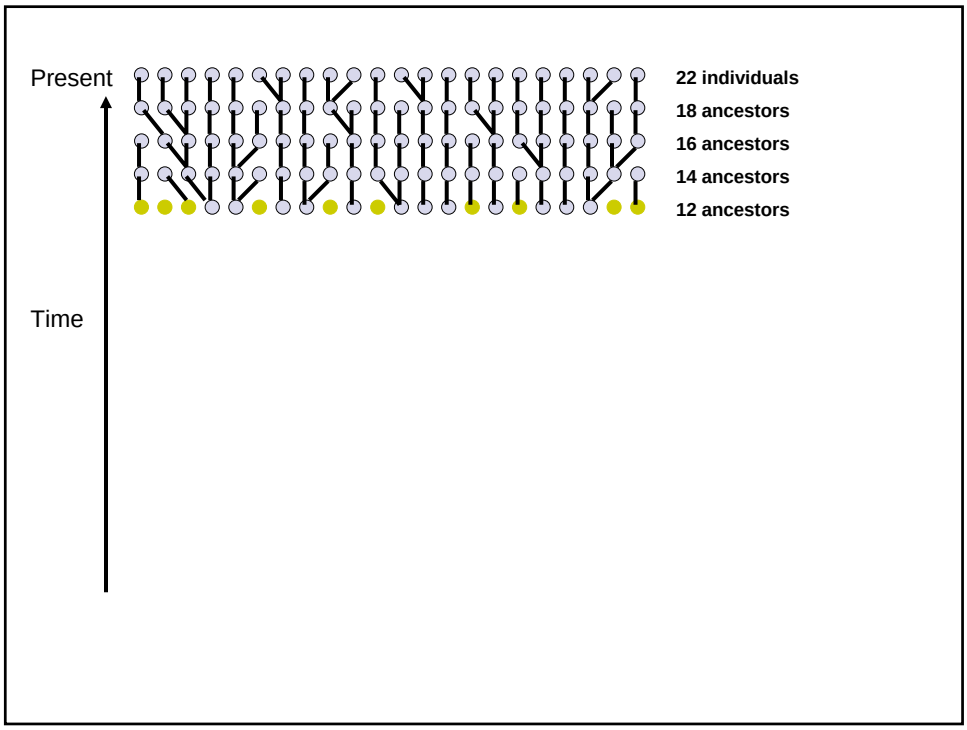
- how we can build up a genealogical tree to relate a sample of ***n*** haploid individuals, collected in the present day?
  - The following series of slides shows how you can build up a genealogical tree to relate a sample of 22 individuals, collected in the present day, at a single haplotype locus (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).

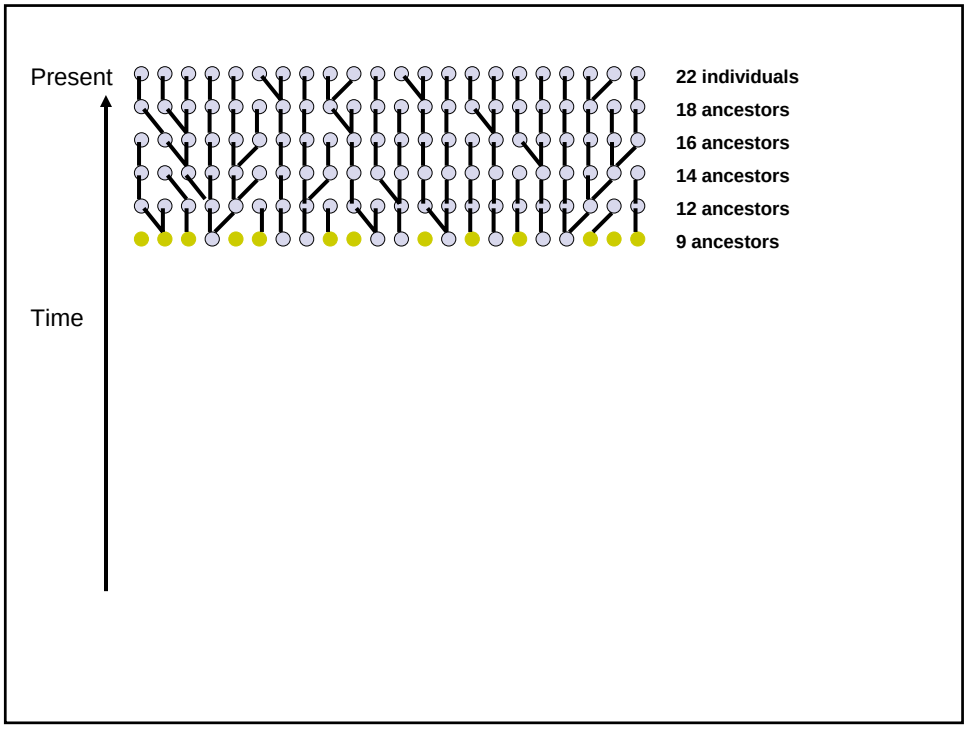


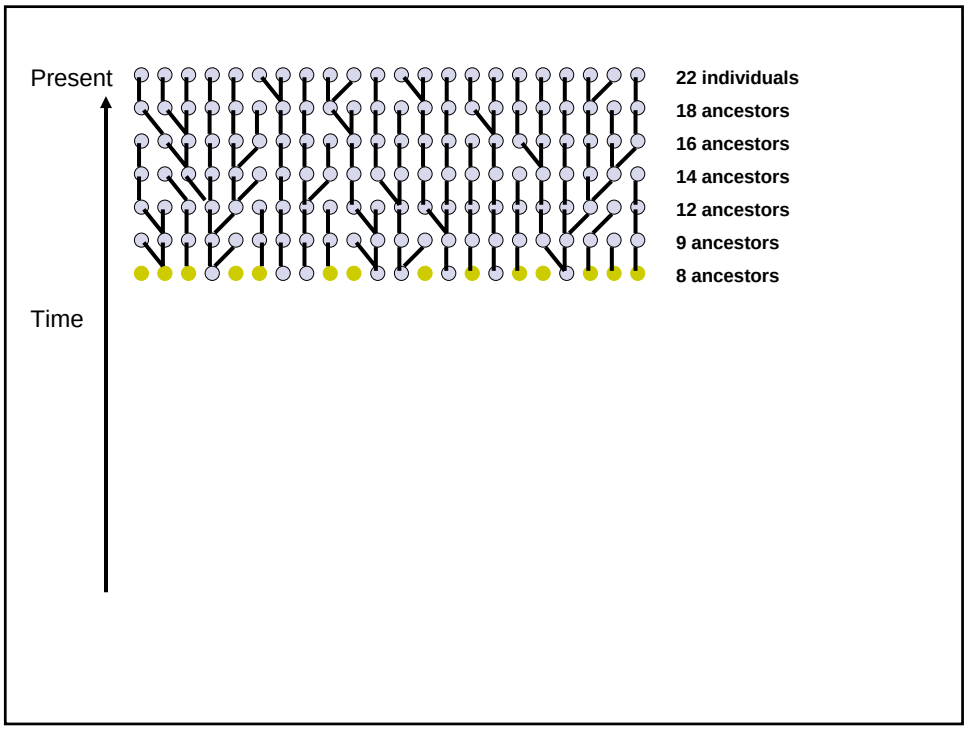

Present
Time
22 individuals
18 ancestors

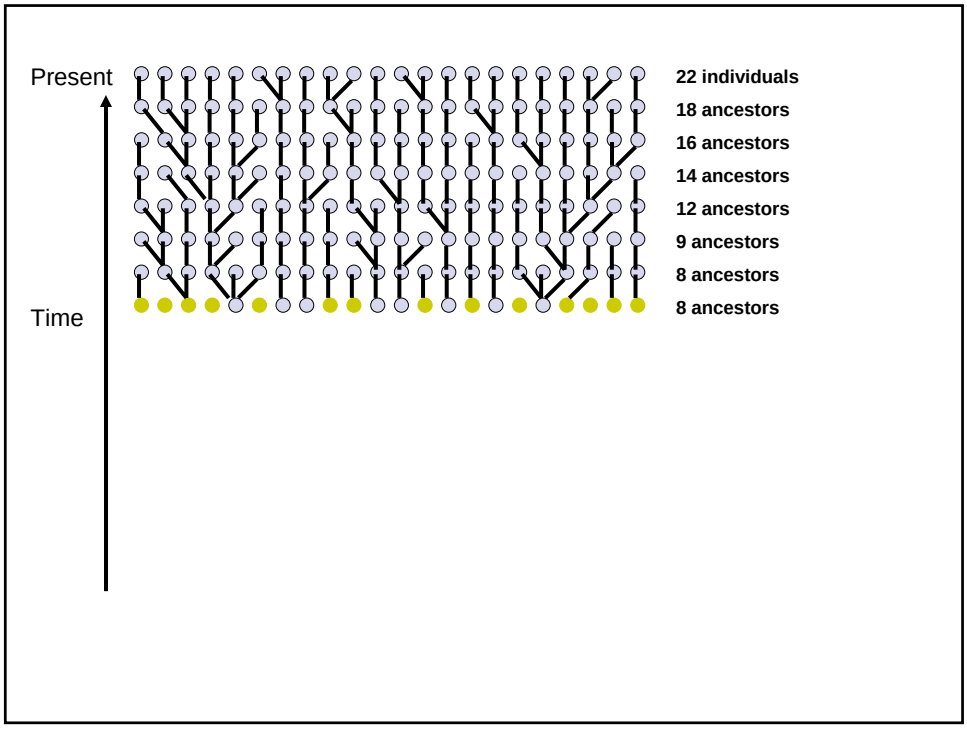

Present
Time
22 individuals
18 ancestors
16 ancestors

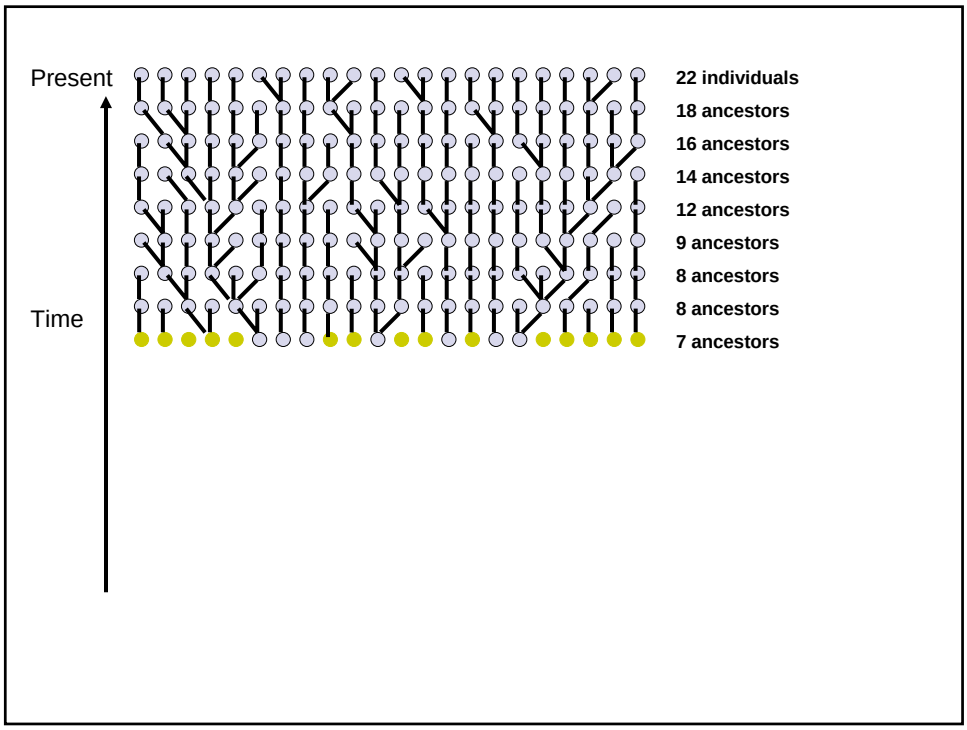

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors

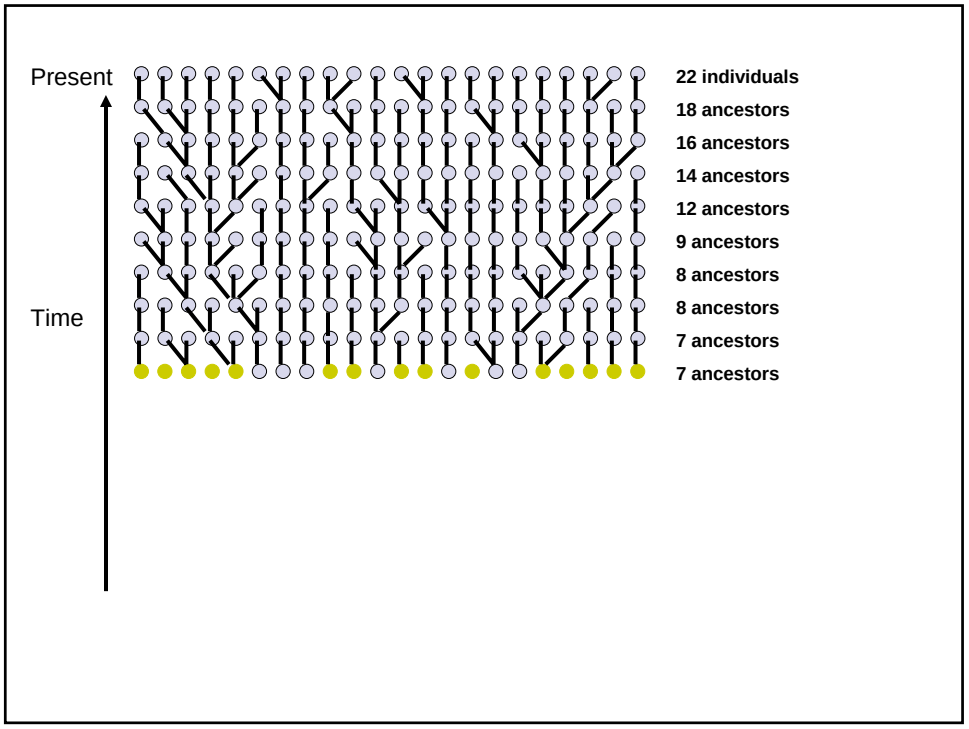

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors

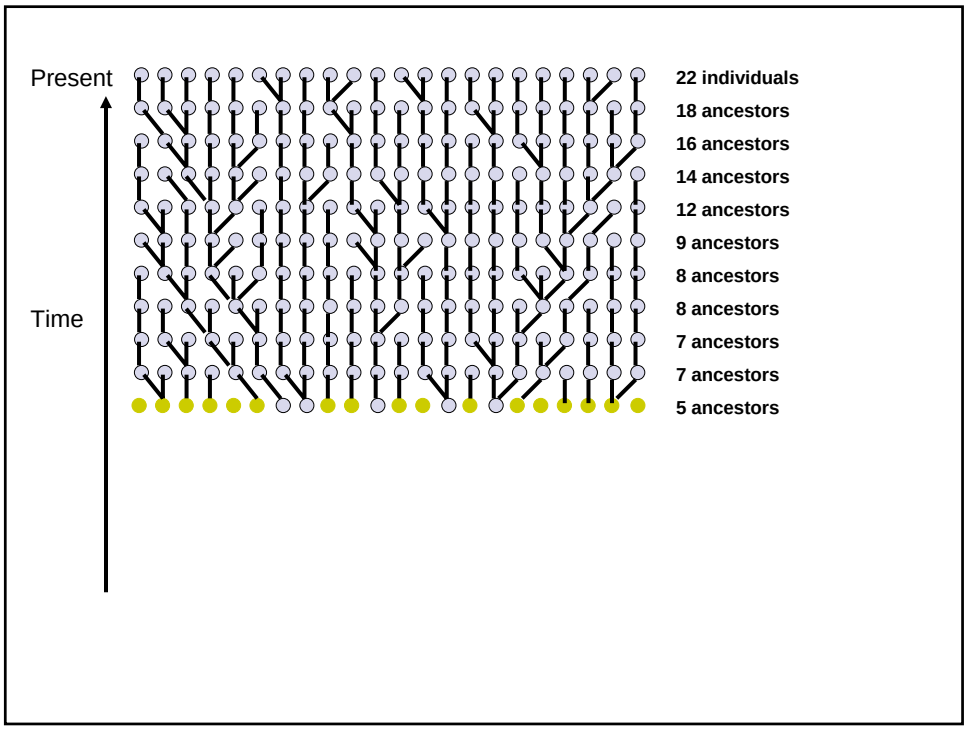

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors

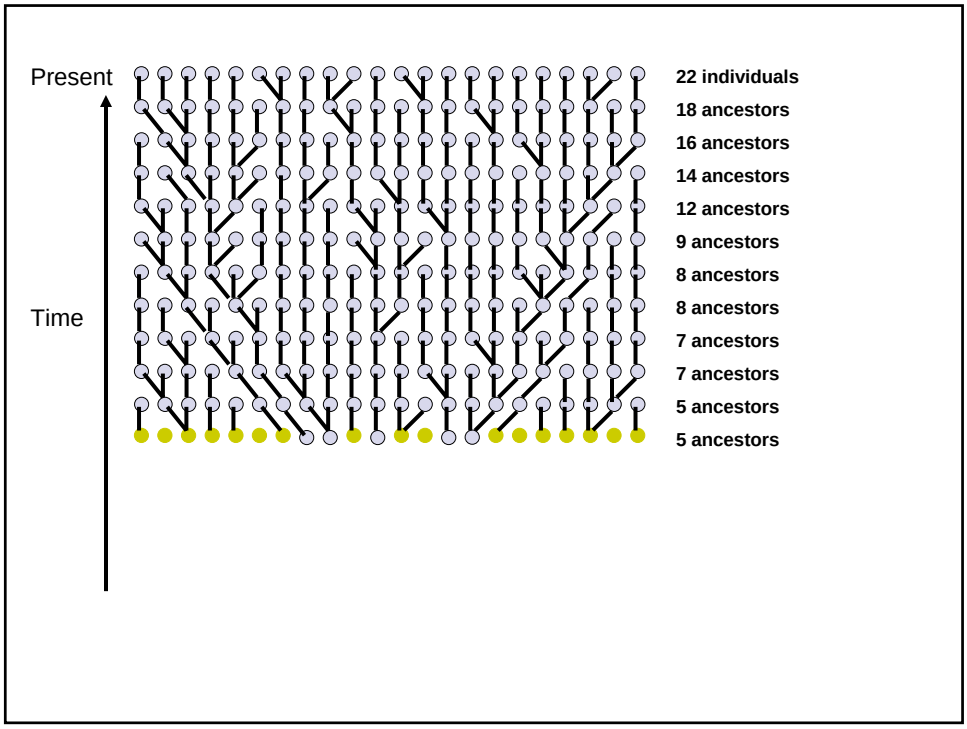

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors

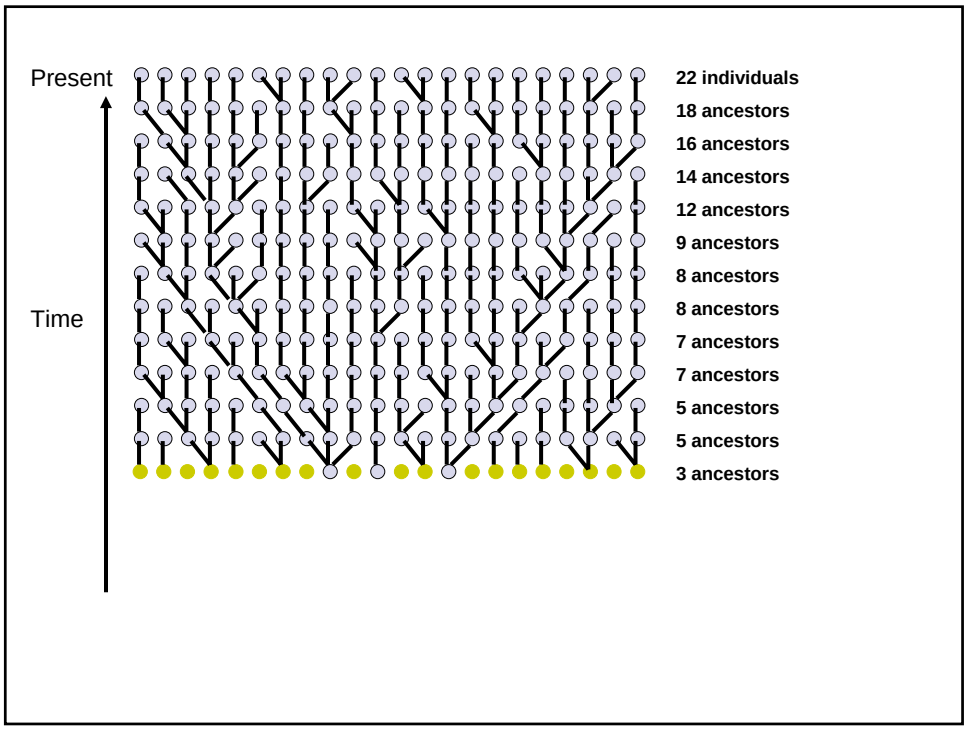

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors

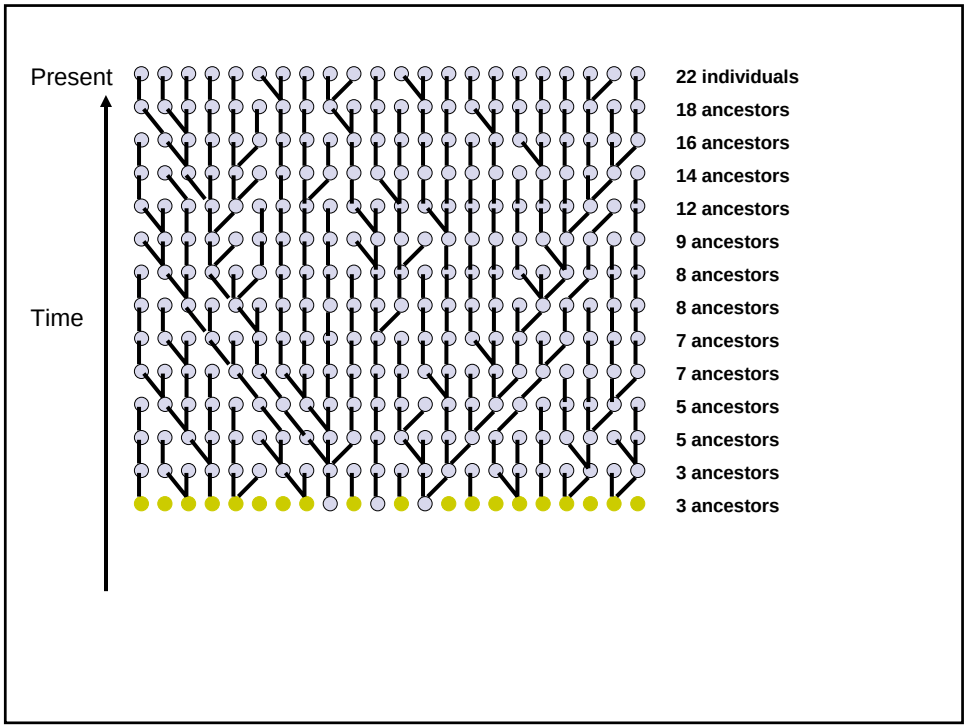

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors

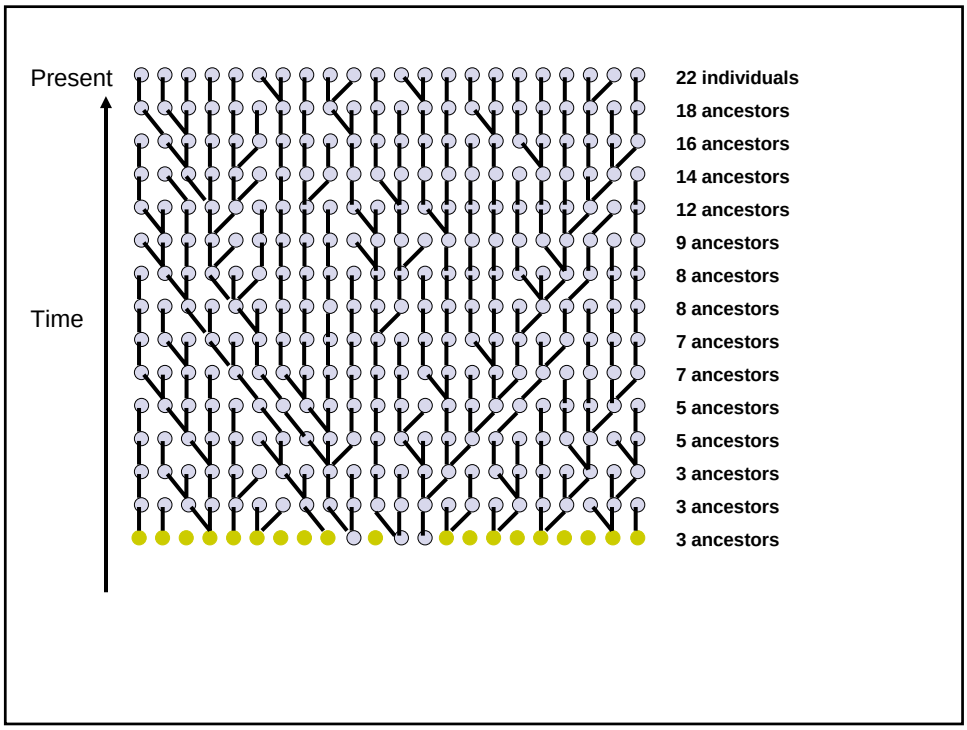

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors

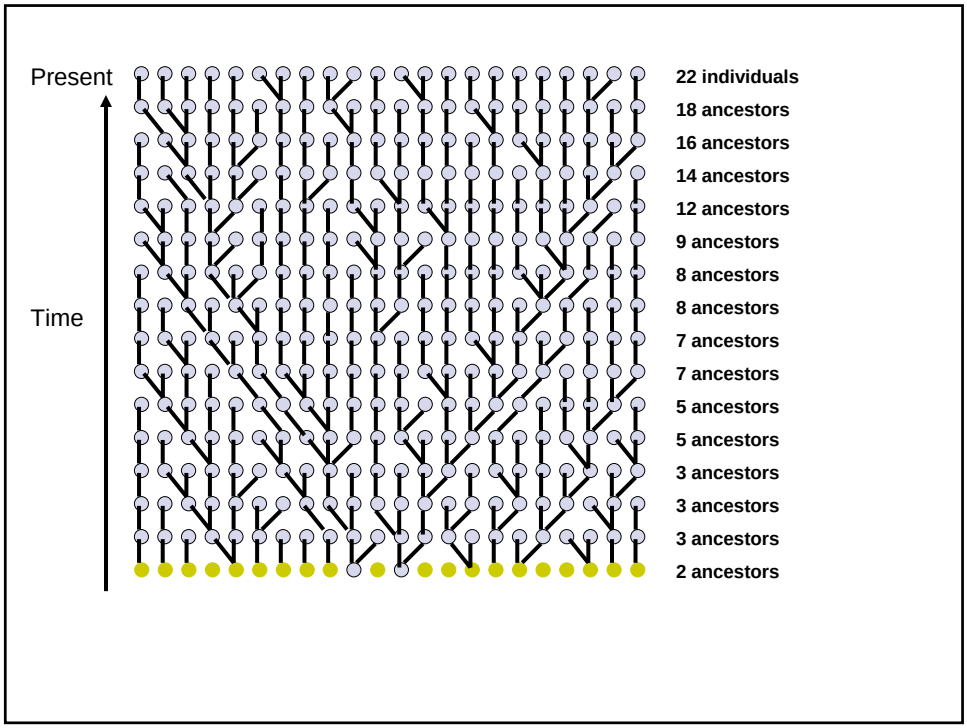

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors

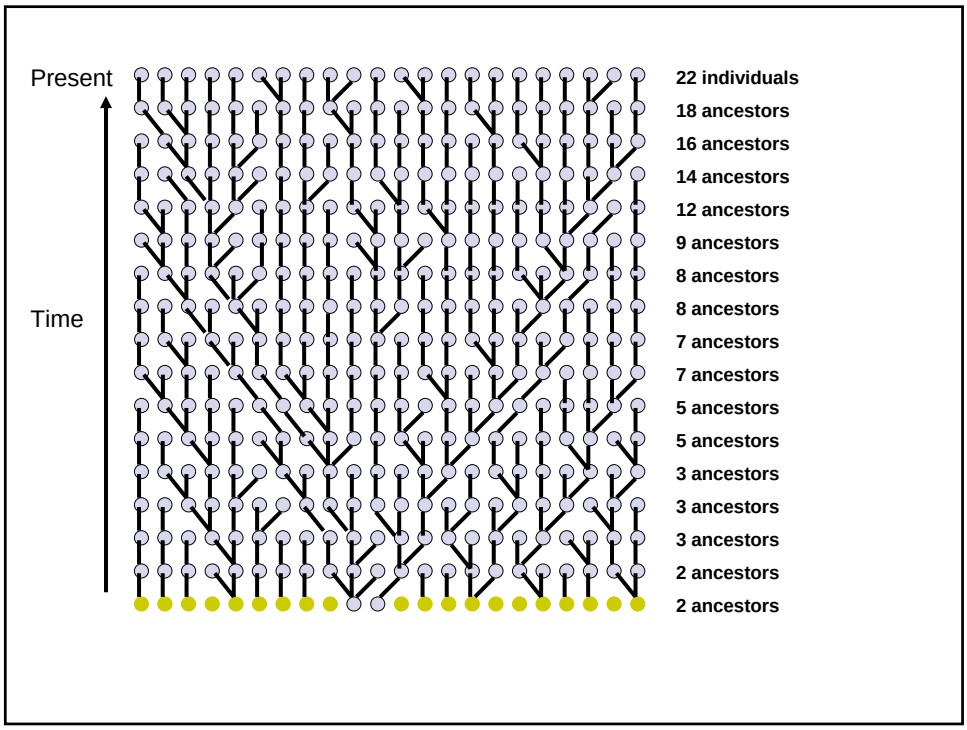

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors

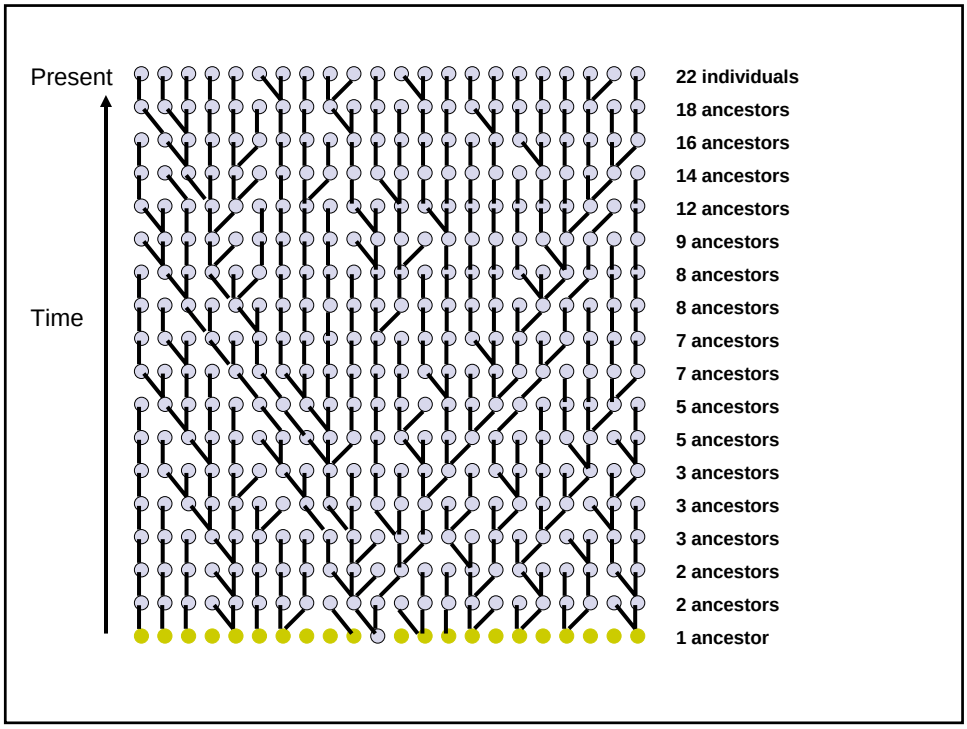

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors

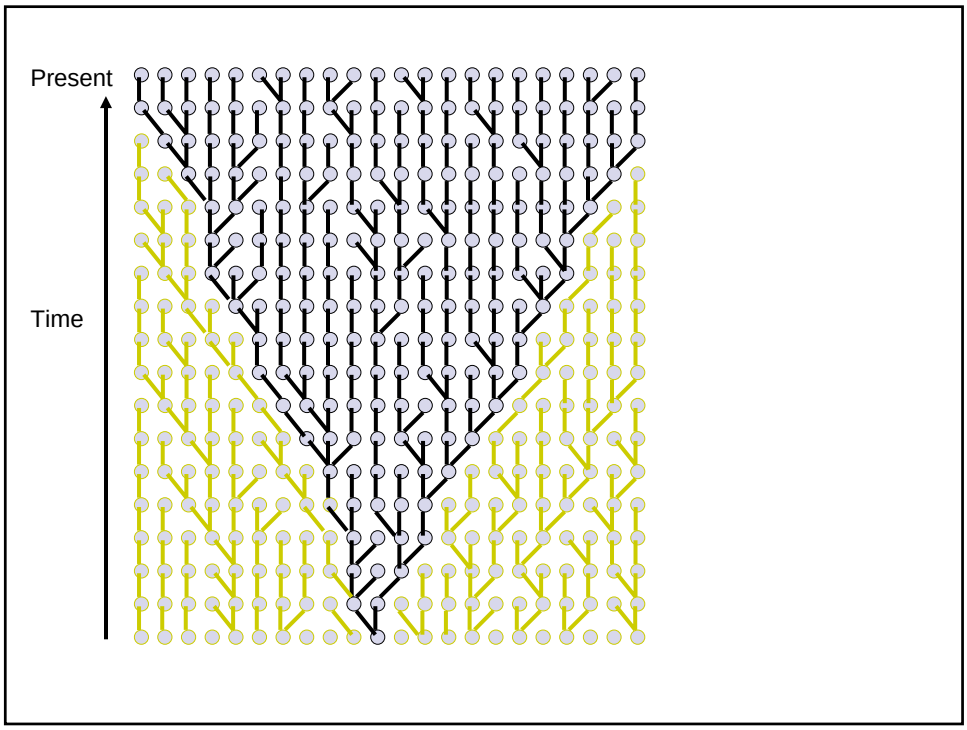

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors
3 ancestors

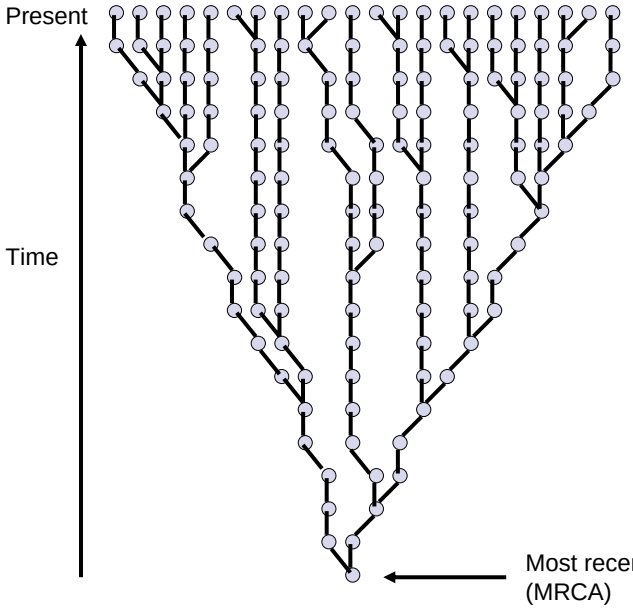

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors
3 ancestors
3 ancestors


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

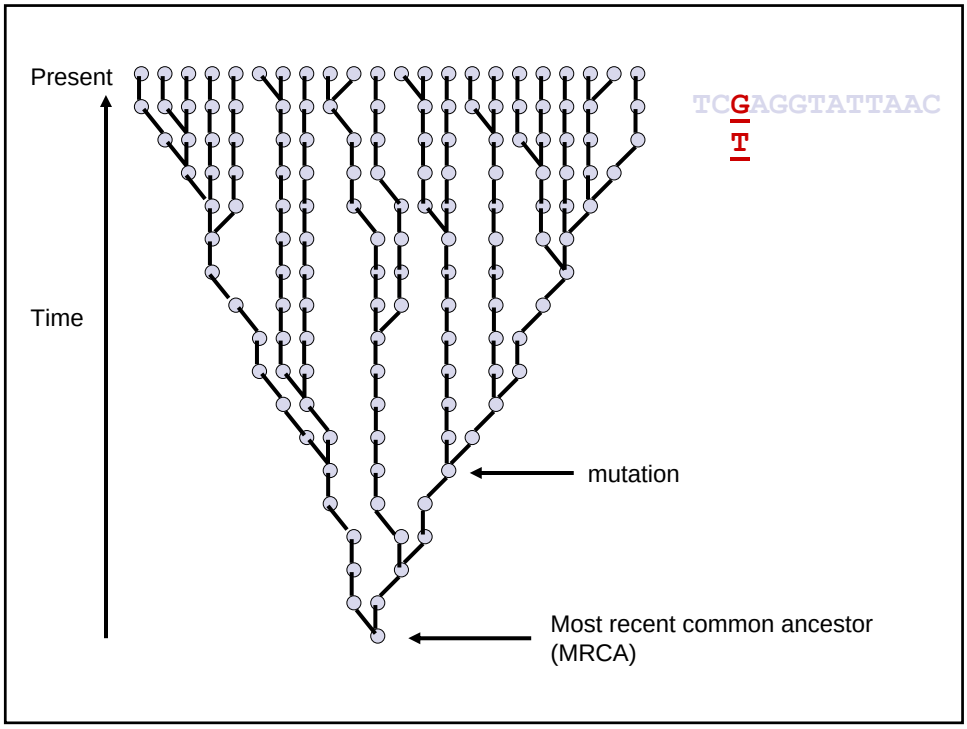

Present
Time
22 individuals
18 ancestors
16 ancestors
14 ancestors
12 ancestors
9 ancestors
8 ancestors
8 ancestors
7 ancestors
7 ancestors
5 ancestors
5 ancestors
3 ancestors
3 ancestors
3 ancestors
2 ancestors
2 ancestors
1 ancestor

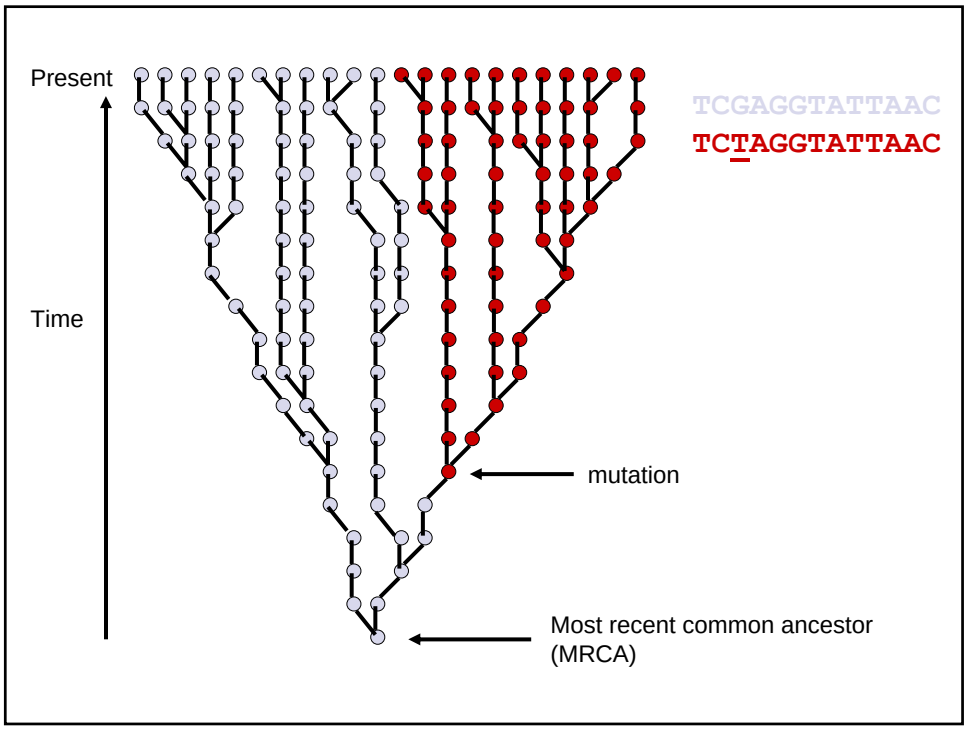

Present
Time

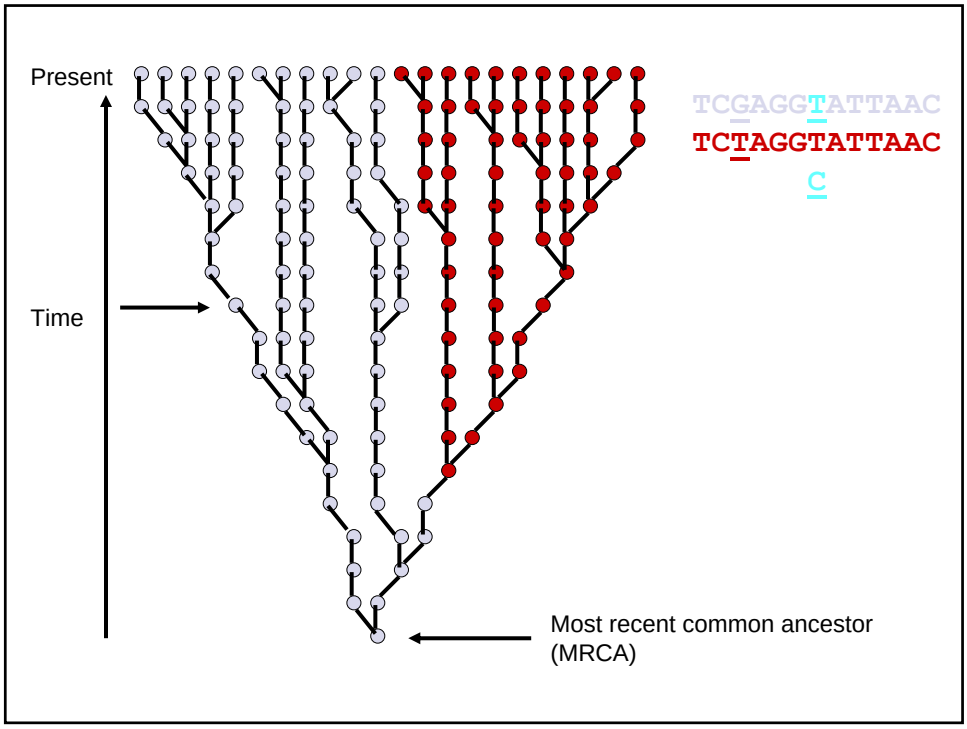


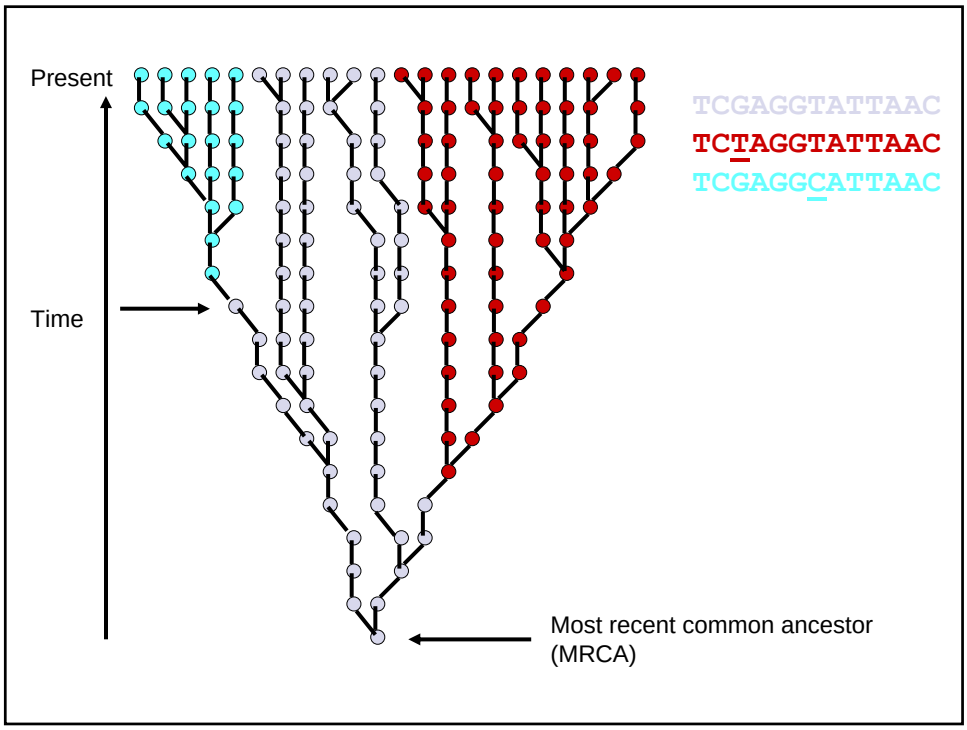
- Mutational events can now be added to the genealogical tree, resulting in **polymorphic** sites. If these sites are typed in the modern sample, they can be used to split the sample into **sub-clades** (represented by different colors)

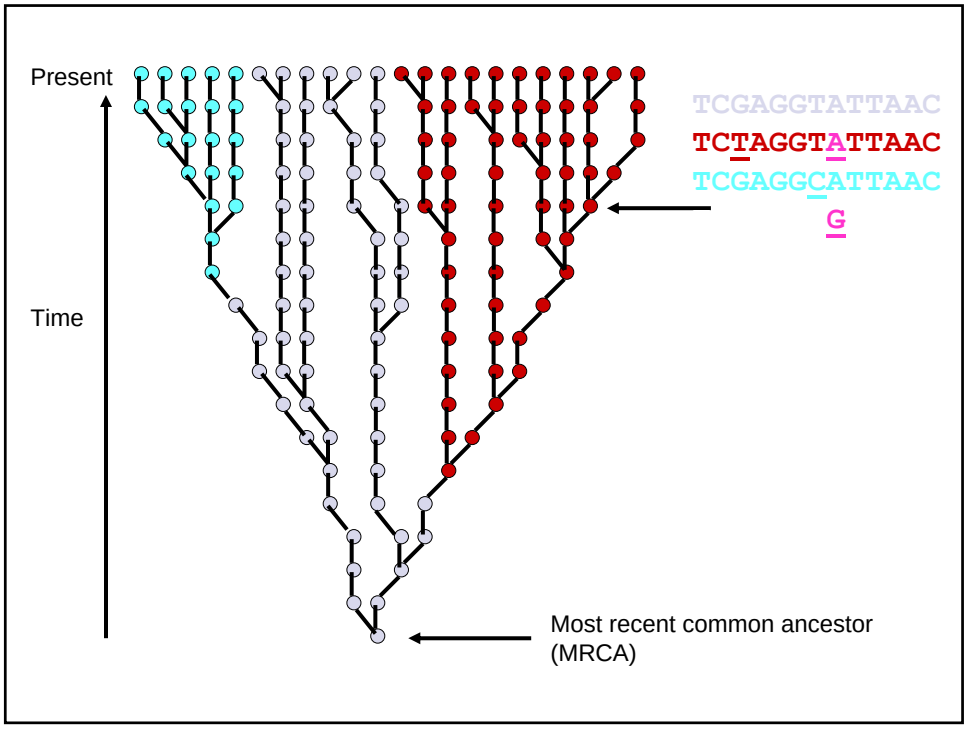
40

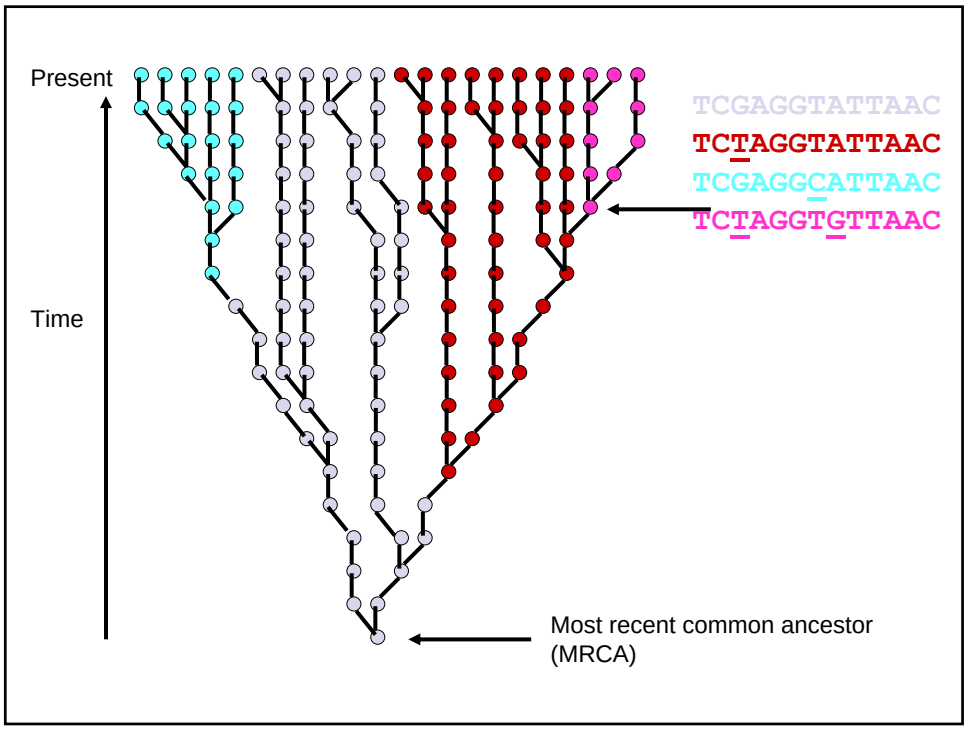

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
mutation
Most recent common ancestor
(MRCA)

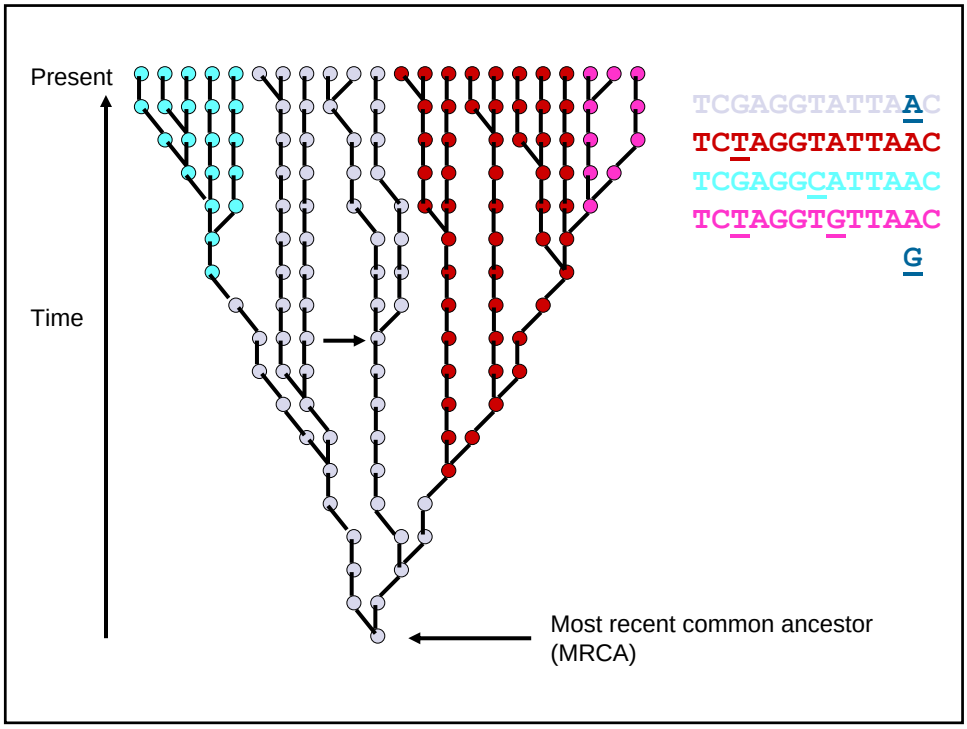

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
mutation
Most recent common ancestor
(MRCA)

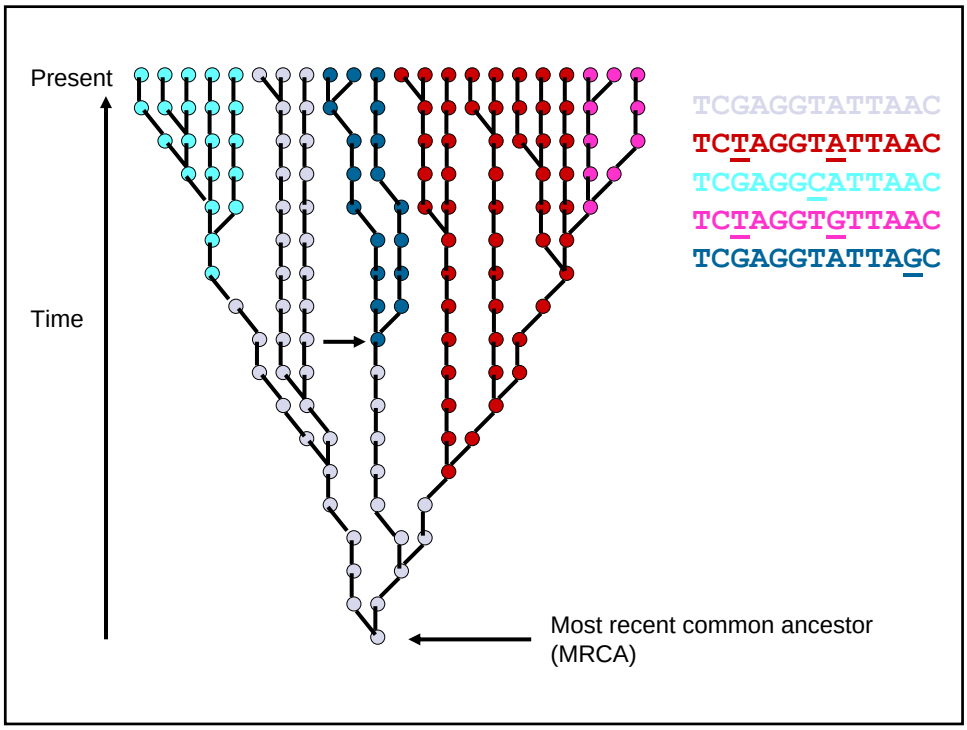

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
C
Most recent common ancestor
(MRCA)

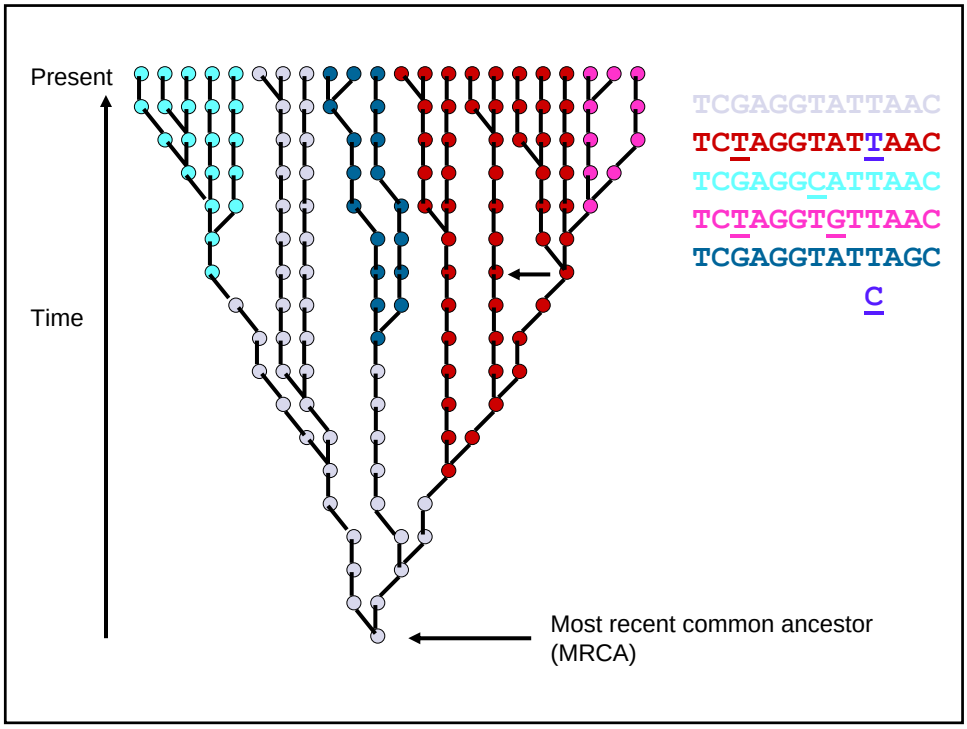

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
Most recent common ancestor
(MRCA)

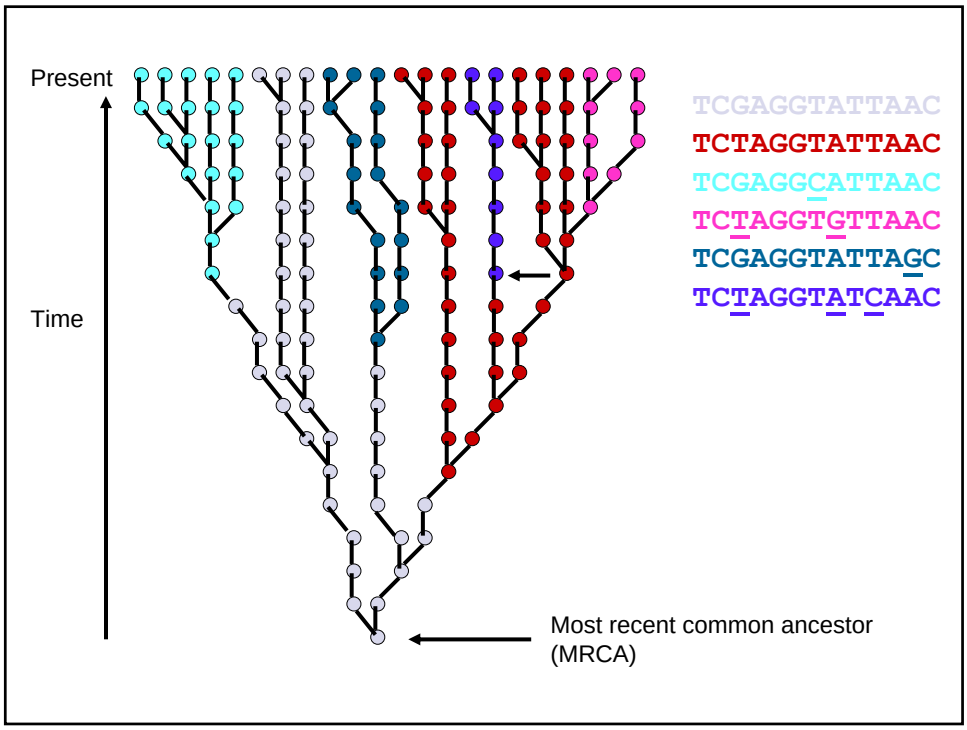

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
G
Most recent common ancestor
(MRCA)

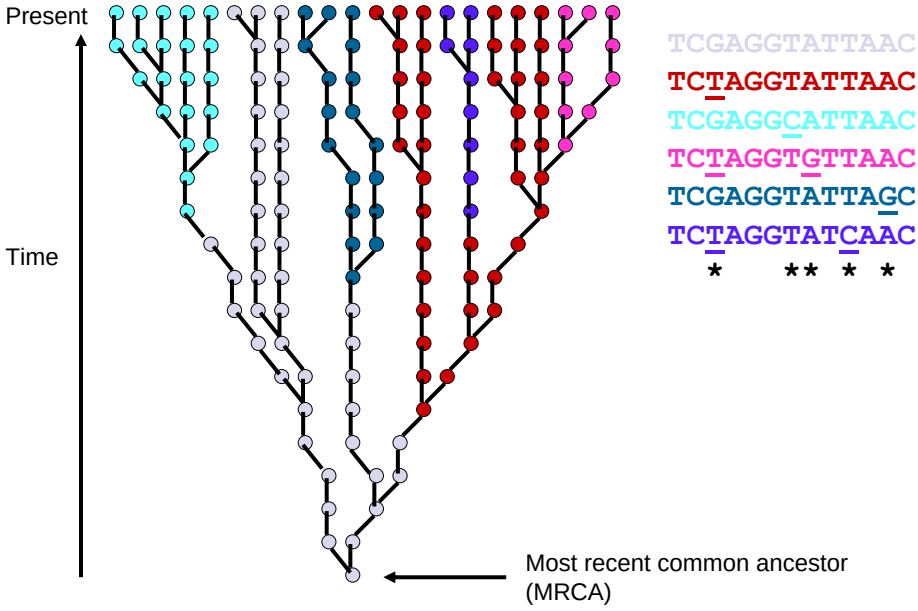

Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
Most recent common ancestor
(MRCA)


Present
Time
TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
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)



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,

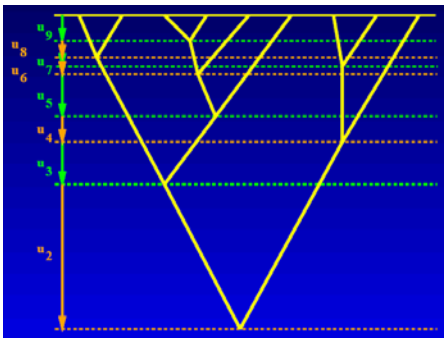
Average time for k copies to coalesce to k-1 copies:

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

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!

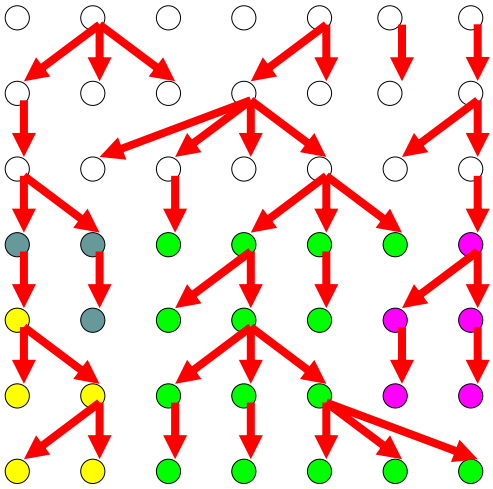
Hint of the derivation

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

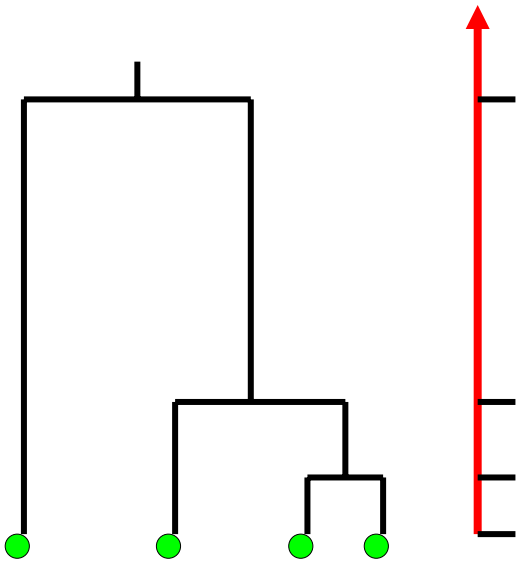
- It is a forwards-in-time model of a **neutral** locus in a constant-size, random-mating, haploid population evolving in **discrete** generations.
- Each individual in generation t has a random number (possibly 0) of offspring in generation $t+1$. Each is:
  - identical to the parent with probability $1-\mu$;
  - 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*.

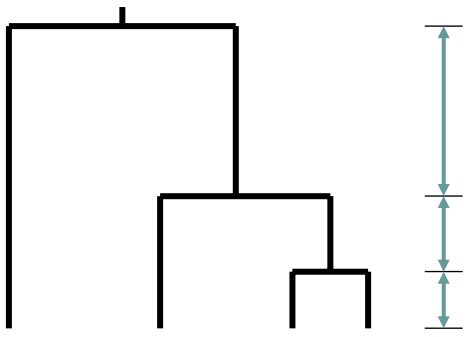
- A coalescent is the backwards-in-time "cousin" of the WF model: similar assumptions, but traces the ancestry of n observed alleles.
- Ancestry is represented via a genealogical tree: leaves are observed alleles, root is the most recent common ancestor (MRCA).



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. $k = 4$:

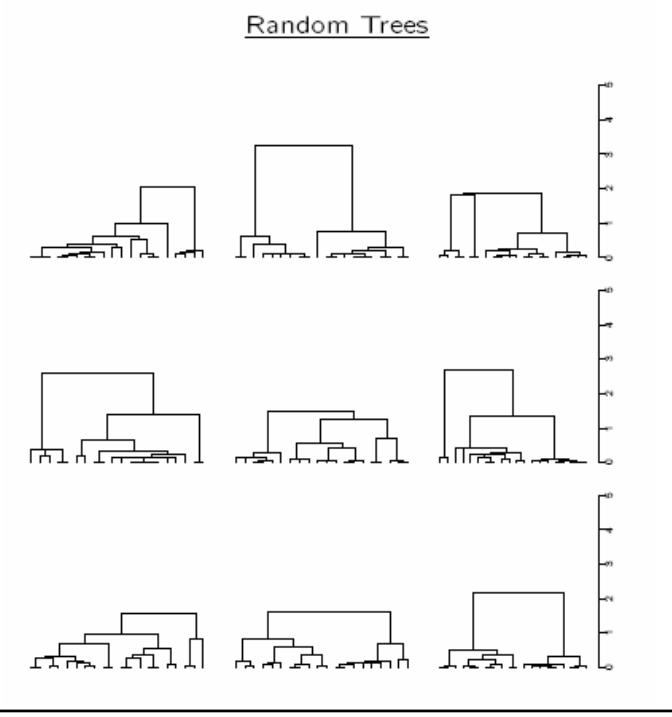


Height of tree: mean = 3/2 sd = 1·07

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

The TMRCA under the coalescent

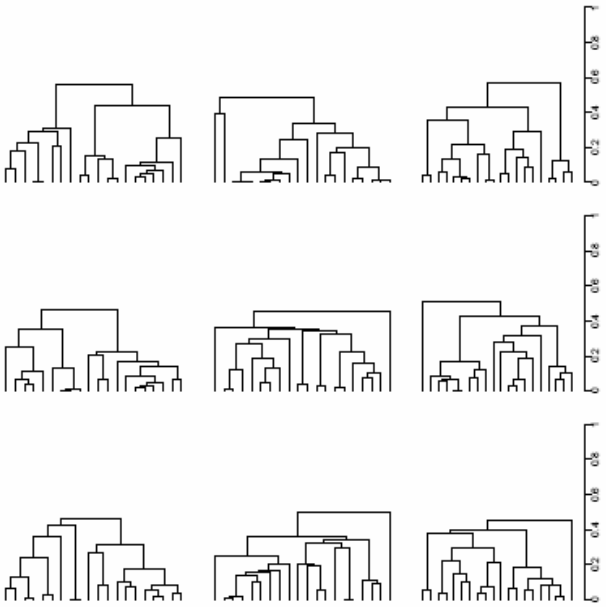
- The TMRCA (height of the genealogical tree) is on average 2(*n*-1)*/n;* the average time in which there are just two ancestral lineages is 1.
  - the number of ancestors of a sample drops rapidly (backwards in time);
  - for more than half its history, on average, a sample has only two ancestors;
  - 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).

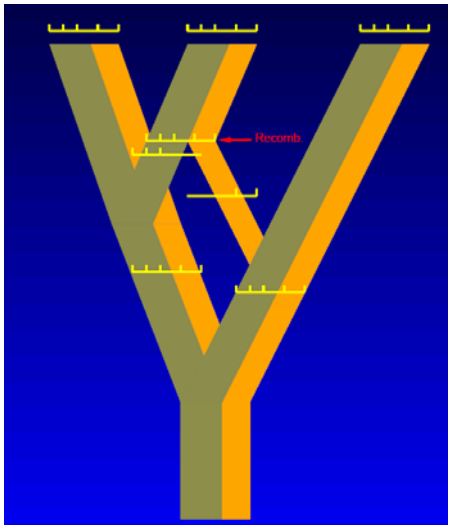
Exponentially Growing Populations



Generalisations of the standard coalescent model

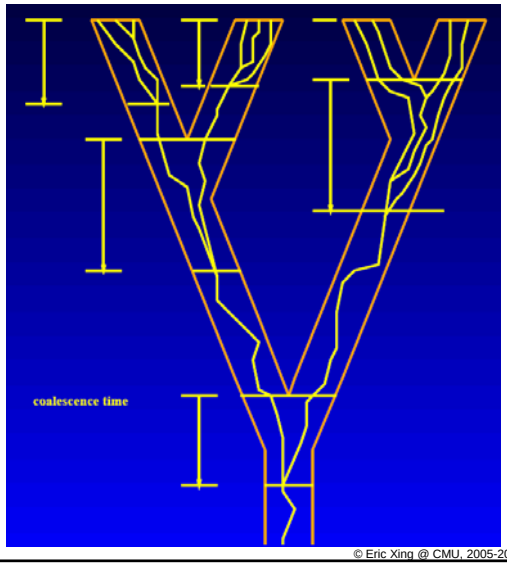
- Variable population size: coalescences occur more rapidly when the population size is small.
- Population subdivision with migration.
- Some forms of selection.
- Recombination: the ancestral recombination graph (ARG)

A recombining coalescent



Different markers have slightly different coalescent trees

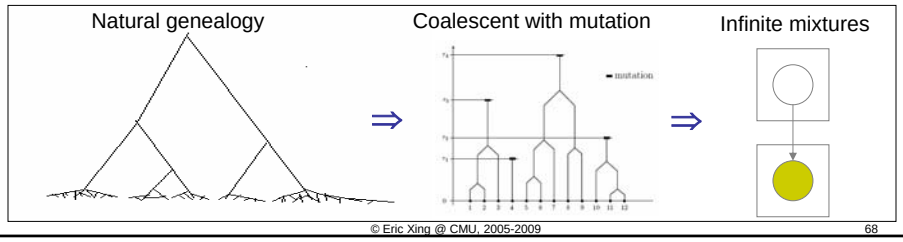
Coalescents in related species

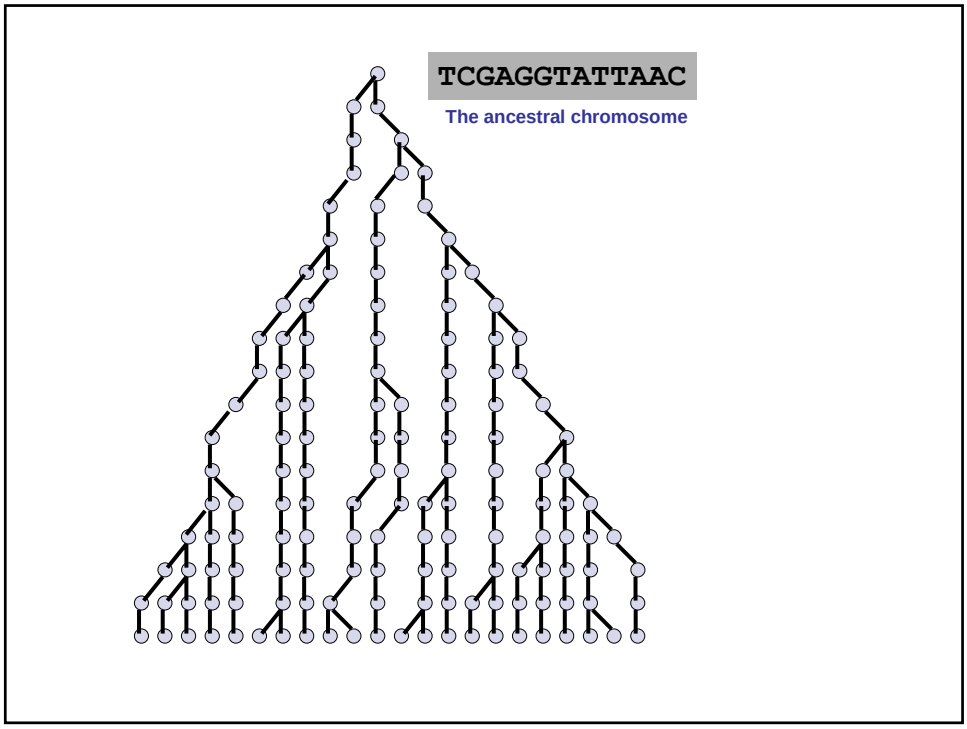


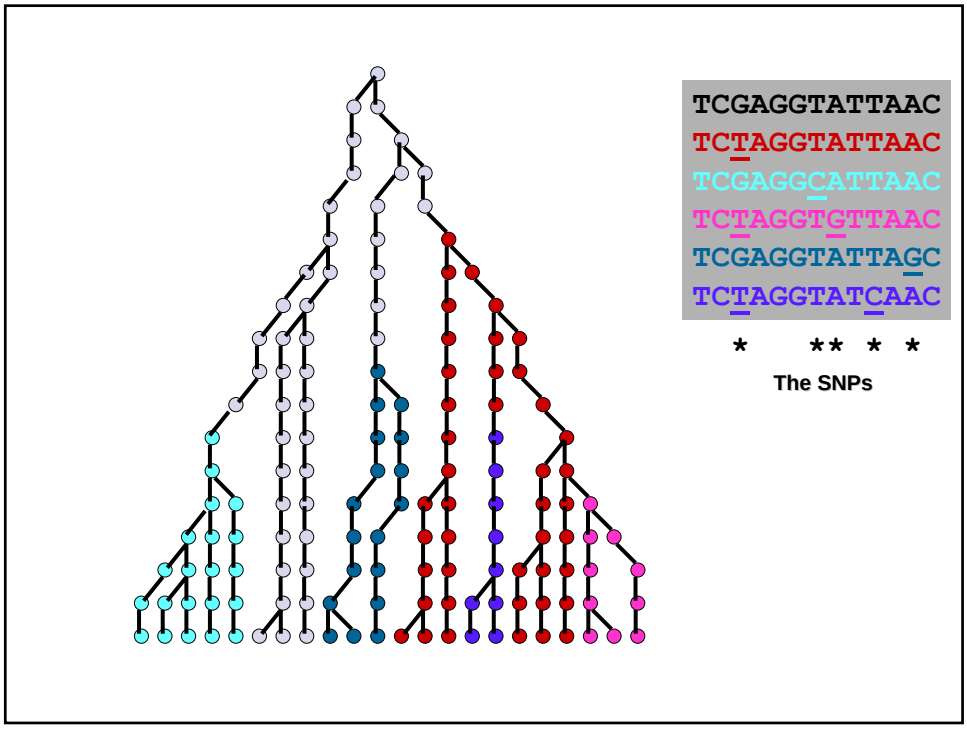
Consistency of gene tree with species tree

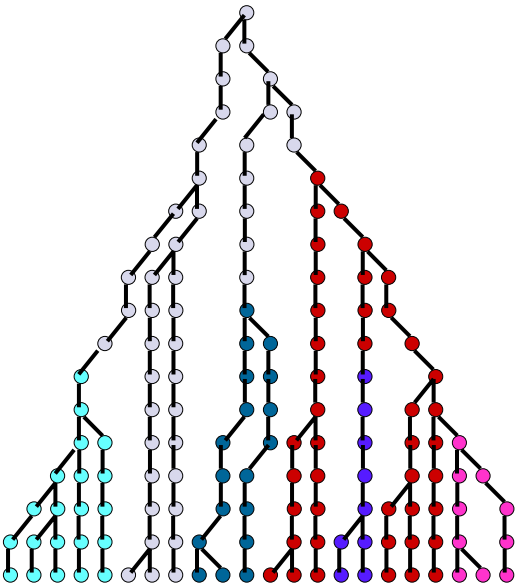
How to approximate a coalescent?

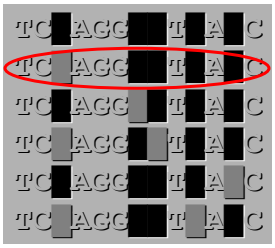
- Kingman coalescent process with binary lineage merging
- New population haplotype alleles emerge along all branches of the coalescence tree at rate a/2 per unit length
- This can be approximated by an infinite mixture model (aks, Dirichlet process mixture)




TCGAGGTATTAAC
The ancestral chromosome

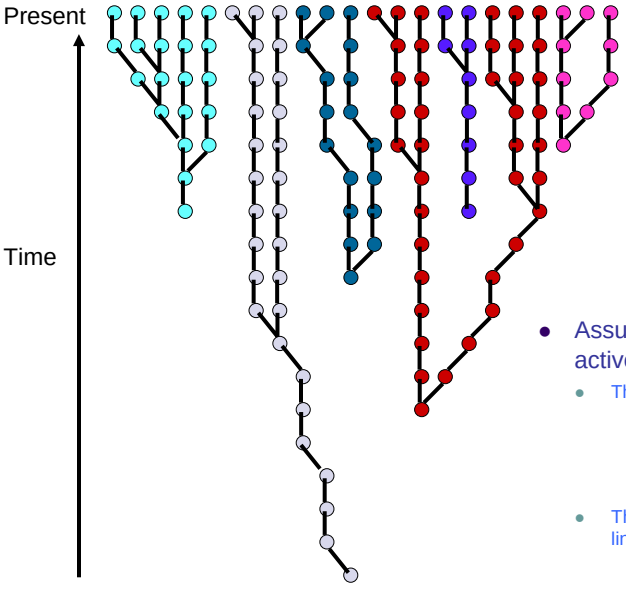

TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
TCTAGGTATCAAC
* ** * *
The SNPs





The haplotypes

useful markers for studying disease association or genome evolution:
-- landmarks, indicators, co-variates, causes …



TCGAGGTATTAAC
TCTAGGTATTAAC
TCGAGGCATTAAC
TCTAGGTGTTAAC
TCGAGGTATTAGC
TCTAGGTATCAAC
* ** * *

- Assuming there are presently k active lineages:
  - The probability of coalescence:

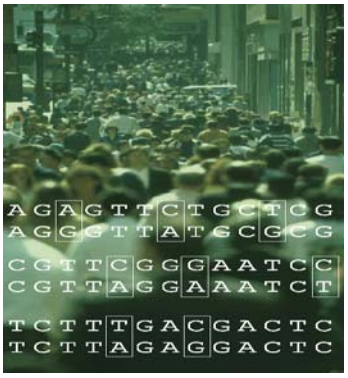
$$\frac{k-1}{k-1+\alpha}$$

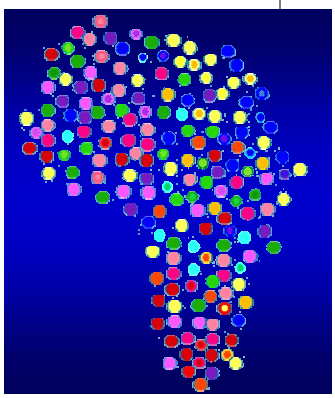
  - The probability of mutation (killing a lineage):

$$\frac{\alpha}{k-1+\alpha}$$

How to derive? ---HW!

Genetic Demography





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

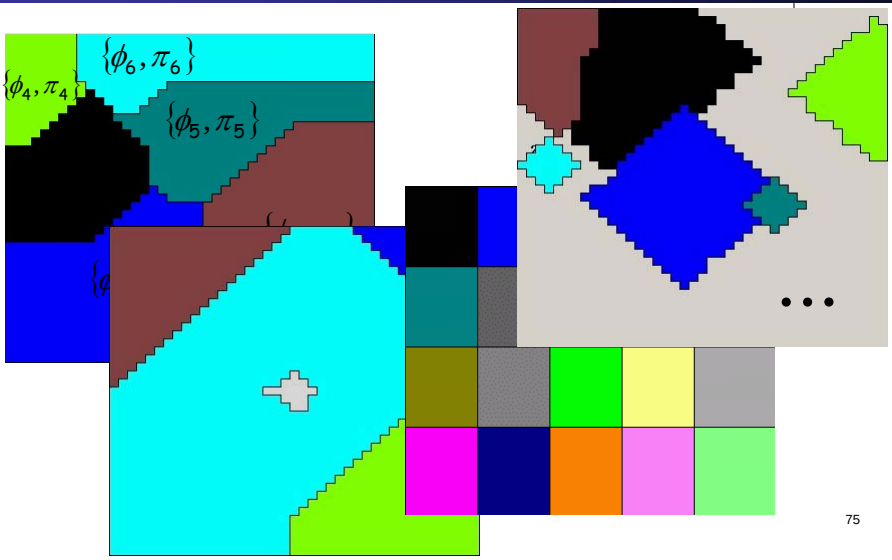
73

Clustering

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

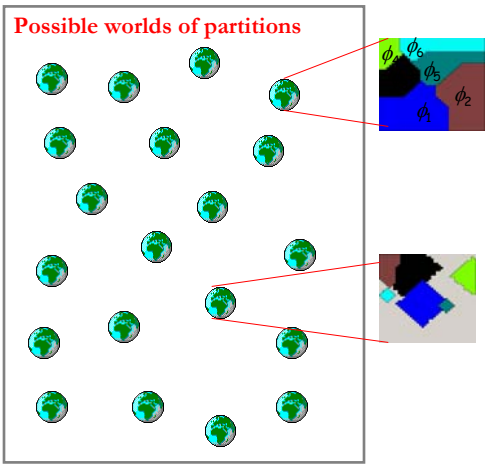
74

Random Partition of Probability Space



75

Dirichlet Process



- A *CDF*, G, on possible worlds of random partitions follows a Dirichlet Process if for any measurable finite partition $(\phi_1, \phi_2, .., \phi_m)$:

$(G(\phi_1), G(\phi_2), \ldots, G(\phi_m)) \sim$ Dirichlet($\alpha G_0(\phi_1), \ldots, \alpha G_0(\phi_m)$)

where G_0 is the base measure and α is the scale parameter

Thus a Dirichlet Process G defines a distribution of distribution

76

DP as a Stick-breaking Process

$G \sim \mathrm{DP}(\alpha, G_0)$

$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k)$

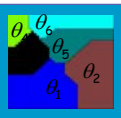
$\theta_k \sim G_0$ — Location

$\sum_{k=1}^{\infty} \pi_k = 1$

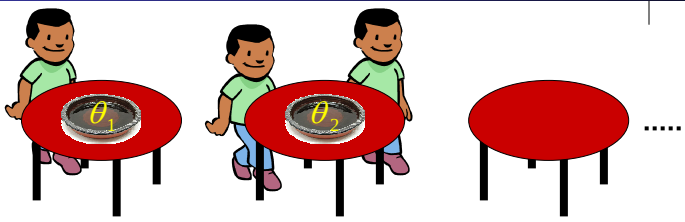
$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_k)$ — Mass

$\beta_k \sim \mathrm{Beta}(1, \alpha)$

$\prod_{j=1}^{k-1}(1-\beta_j)$	β_k	π_k
0	0.4	0.4
0.6	0.5	0.3
0.3	0.8	0.24



Chinese Restaurant Process



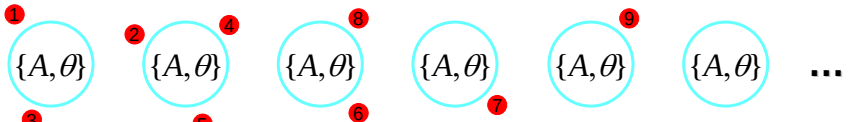
$P(c_i = k \mid \mathbf{c}_{-i}) =$

1	0	0
$\frac{1}{1+\alpha}$	$\frac{\alpha}{1+\alpha}$	0
$\frac{1}{2+\alpha}$	$\frac{1}{2+\alpha}$	$\frac{\alpha}{2+\alpha}$
$\frac{1}{3+\alpha}$	$\frac{2}{3+\alpha}$	$\frac{\alpha}{3+\alpha}$
$\frac{m_1}{i+\alpha-1}$	$\frac{m_2}{i+\alpha-1}$	$\frac{\alpha}{i+\alpha-1}$

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

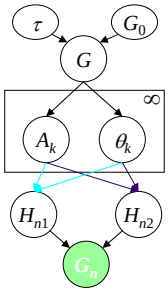
The DP Mixture of Ancestral Haplotypes

- 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



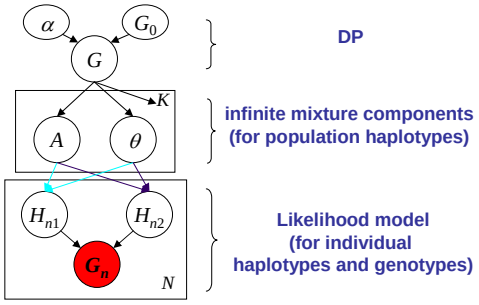
- Assume an individual haplotype h is stochastically derived from a population haplotype a_k with nucleotide-substitution frequency θ_k:

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

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

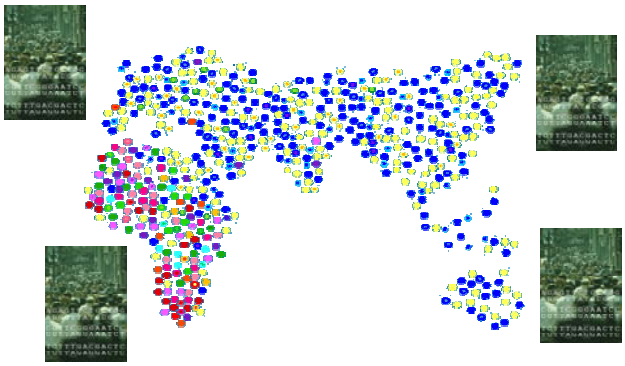
- The number and identity of the population haplotypes are unknown
  - use a Dirichlet Process to construct a prior distribution G on $\mathcal{H}\times\mathcal{R}^J$.

DP-haplotyper



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

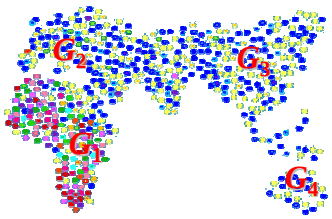
Multi-population Genetic Gemography



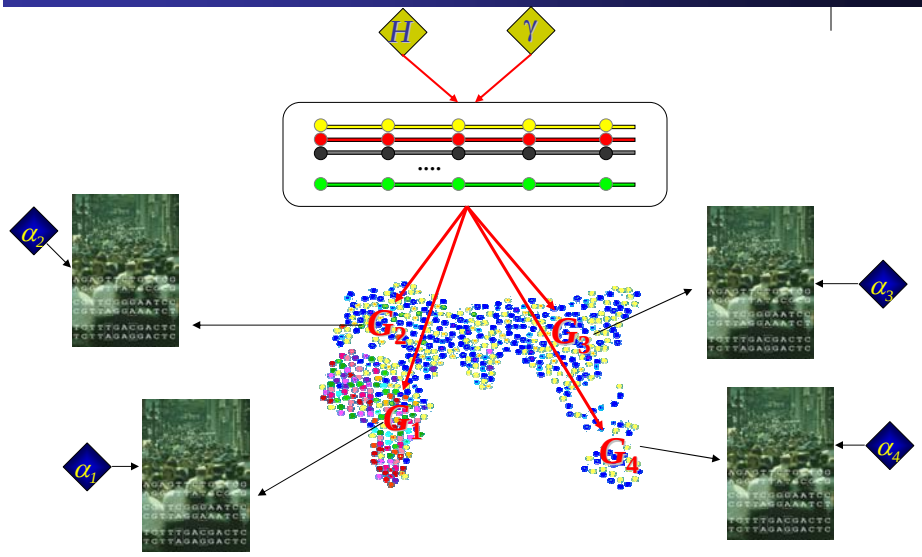
- Pool everything together and solve 1 hap problem?
  - --- ignore population structures
- Solve 4 hap problems separately?
  - --- data fragmentation
- Co-clustering … solve 4 *coupled* hap problems jointly

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



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.