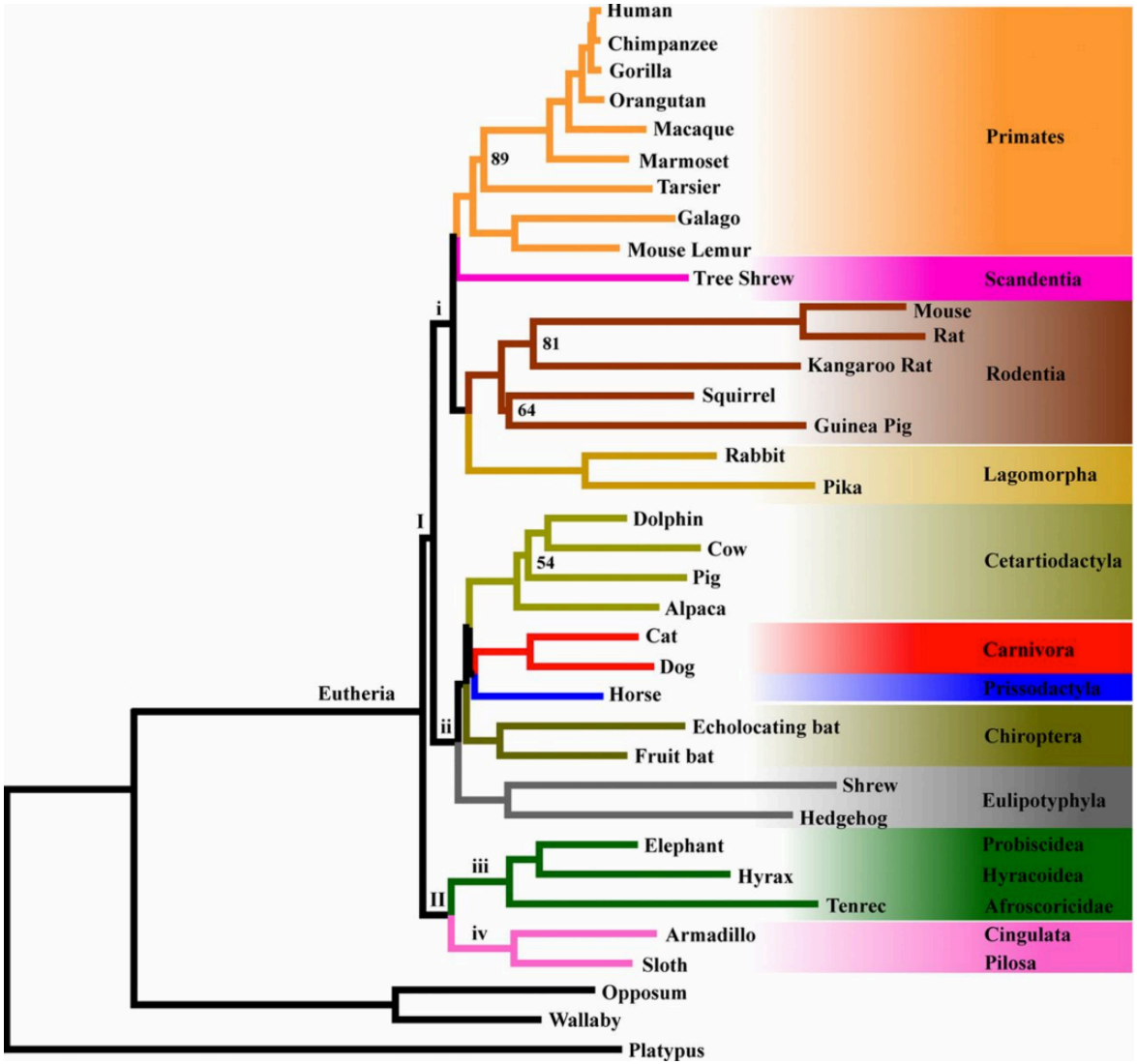
Evolution, Population Genetics, and Natural Selection

02-710 Computational Genomics

Seyoung Kim

Phylogeny of Mammals



Phylogenetics vs. Population Genetics

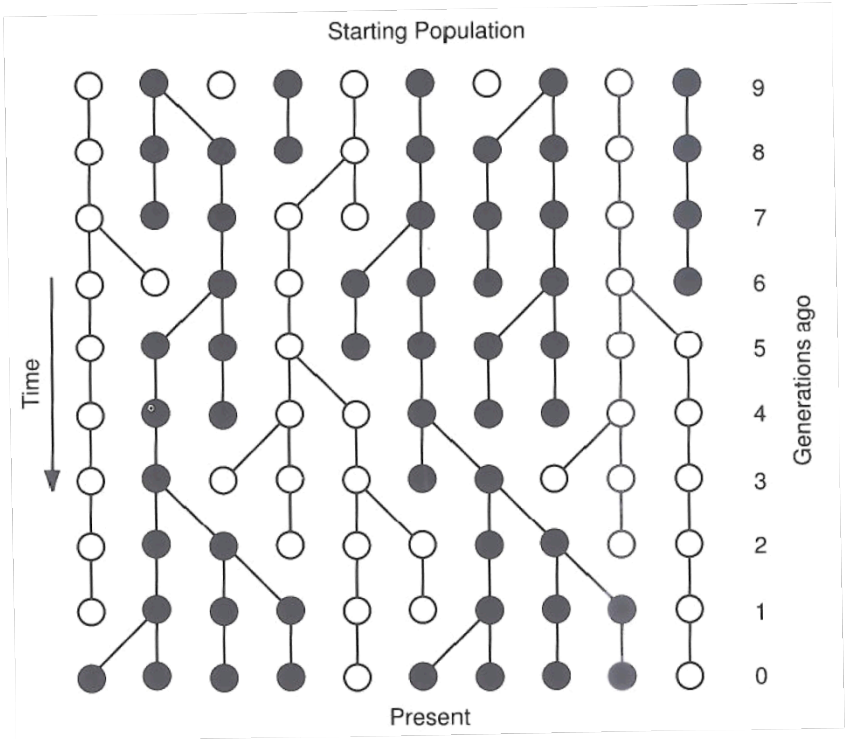
- Phylogenetics
  - Assumes a single correct species phylogeny that holds across genomes
  - Reduces the entire population of a species into a single individual
  - Ignores variations among individuals of the same species or assumes a negligible variability within species

- Population genetics
  - Usually concerned with within-species variation in genomes
  - Individuals within a species are related by genealogies

Siepel, A. Genome Res. 19(11):1929-41. 2009.
Phylogenomics of primates and their ancestral populations.

Write-Fisher Model

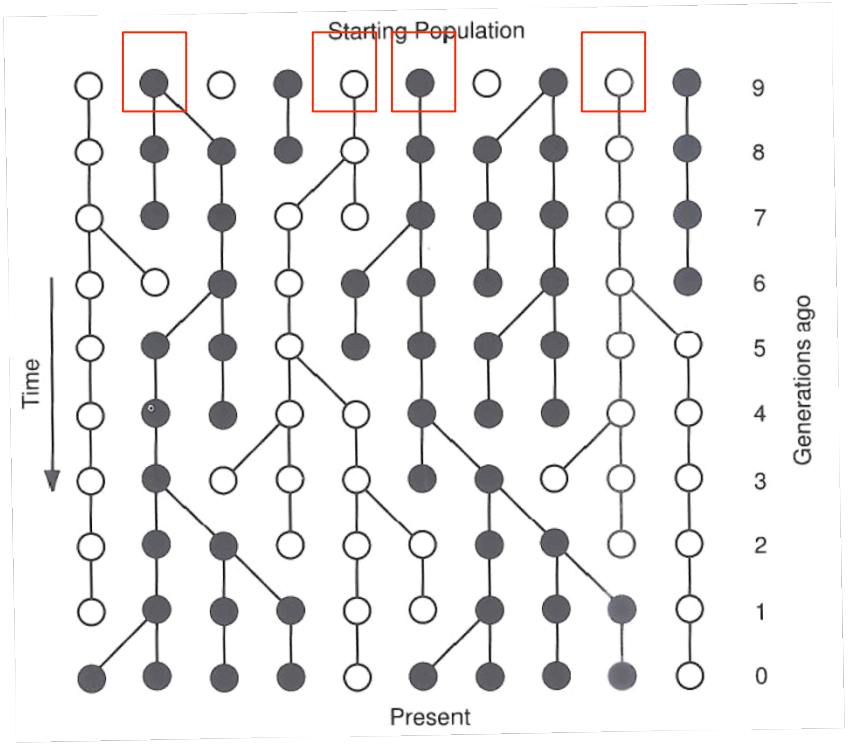
- Stochastic model for gene frequencies in finite populations
  - Assume the same population size *N* at each generation. Thus, 2*N* copies of genes. (no recombination)
  - *p*, *q*: allele frequencies of two alleles
  - the probability of having *k* copies of one allele (with frequency *p* in the current generation) in the next generation is given as:



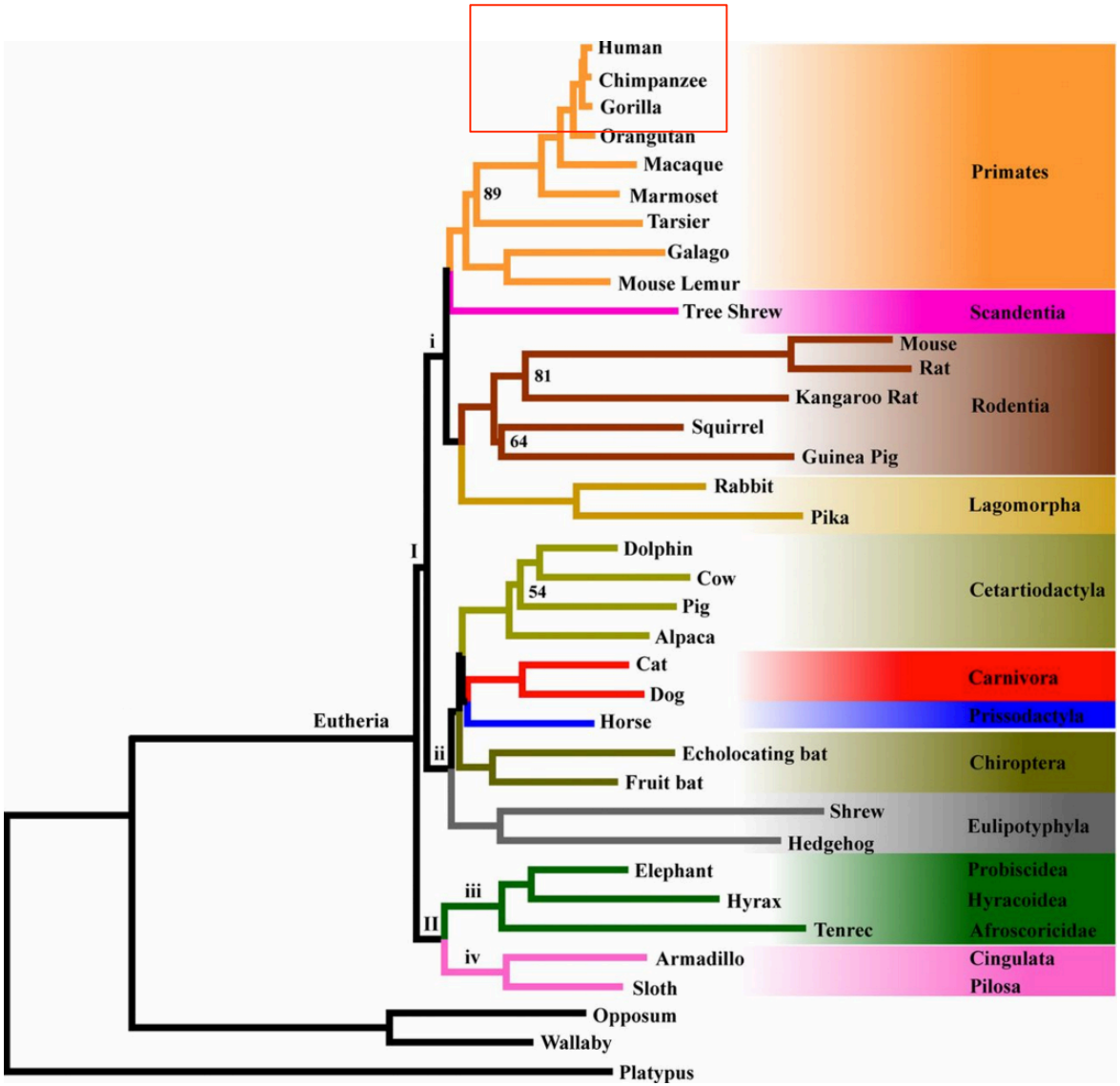
$$\binom{2N}{k} p^k q^{2N-k}$$

Write-Fisher Model

- An individual at generation *t* randomly samples two individuals at generation *t-1* as parents
- We can trace the ancestry of each individual at the current generation backward
- In finite number of generations, all individuals will coalesce into a single individual



Phylogeny of Mammals



Population-aware Phylogenetics

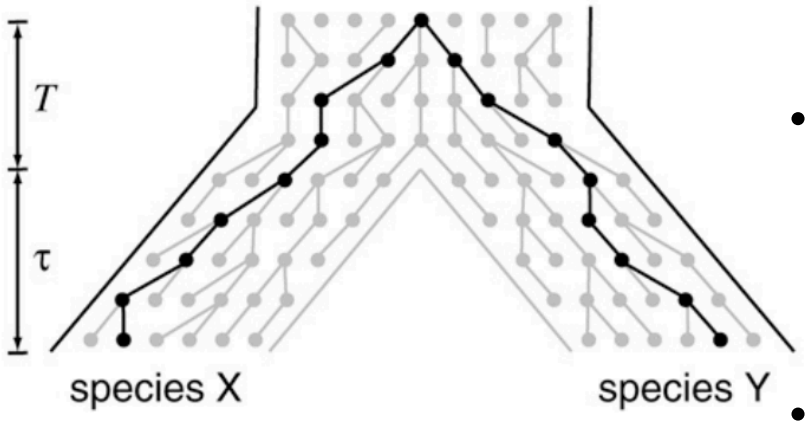
- Primate species
  - Divergence time is short relative to ancestral population sizes
  - Phylogenetics assumptions do not hold
  - Non-negligible population genetic effects

- Interspecies comparison, taking into account selective forces within species

Population Genetics Interpretation of Speciation

τ: speciation time
T: coalescent time
N_e: Population size

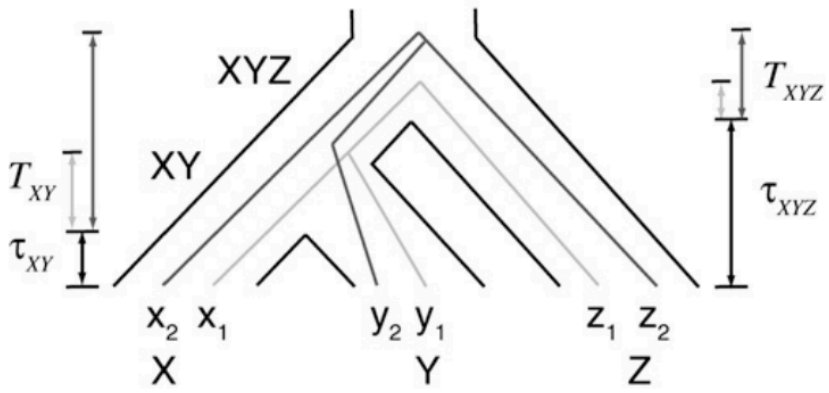
$T \sim \exp(2N_e)$



- $\tau >> T$ ($\tau >> N_e$):
  - the phylogenetics assumption holds
  - Divergence between individual chromosomes as an estimate of speciation time
- $\tau << T$ ($\tau << N_e$):
  - Equivalent to the coalescence in population genetics
  - Coalescent time dominates
- $\tau \sim T$ ($\tau \sim N_e$):
  - Both ancestral population dynamics and interspecies divergence must be considered
  - Population-aware phylogenetics

Three-Species Phylogeny

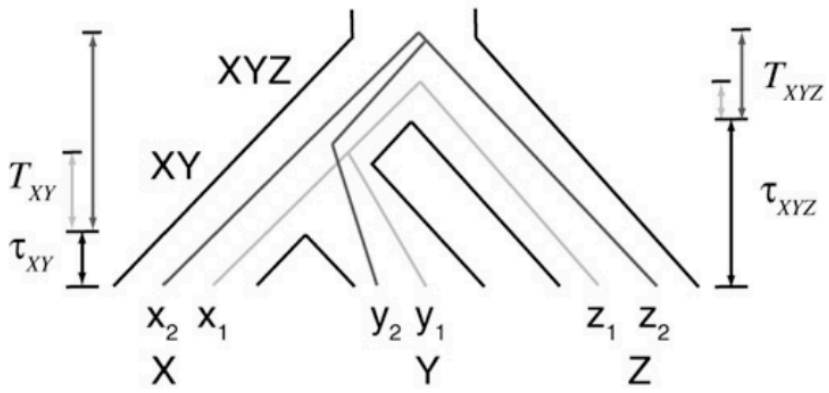
τ: speciation time
T: coalescent time



- Gray phylogeny: concordant with the phylogeny among the three species
- Black phylogeny: discordance with the phylogeny among the three species
- ILS: incomplete lineage sorting with deep coalescent

Three-Species Phylogeny

τ: speciation time
T: coalescent time

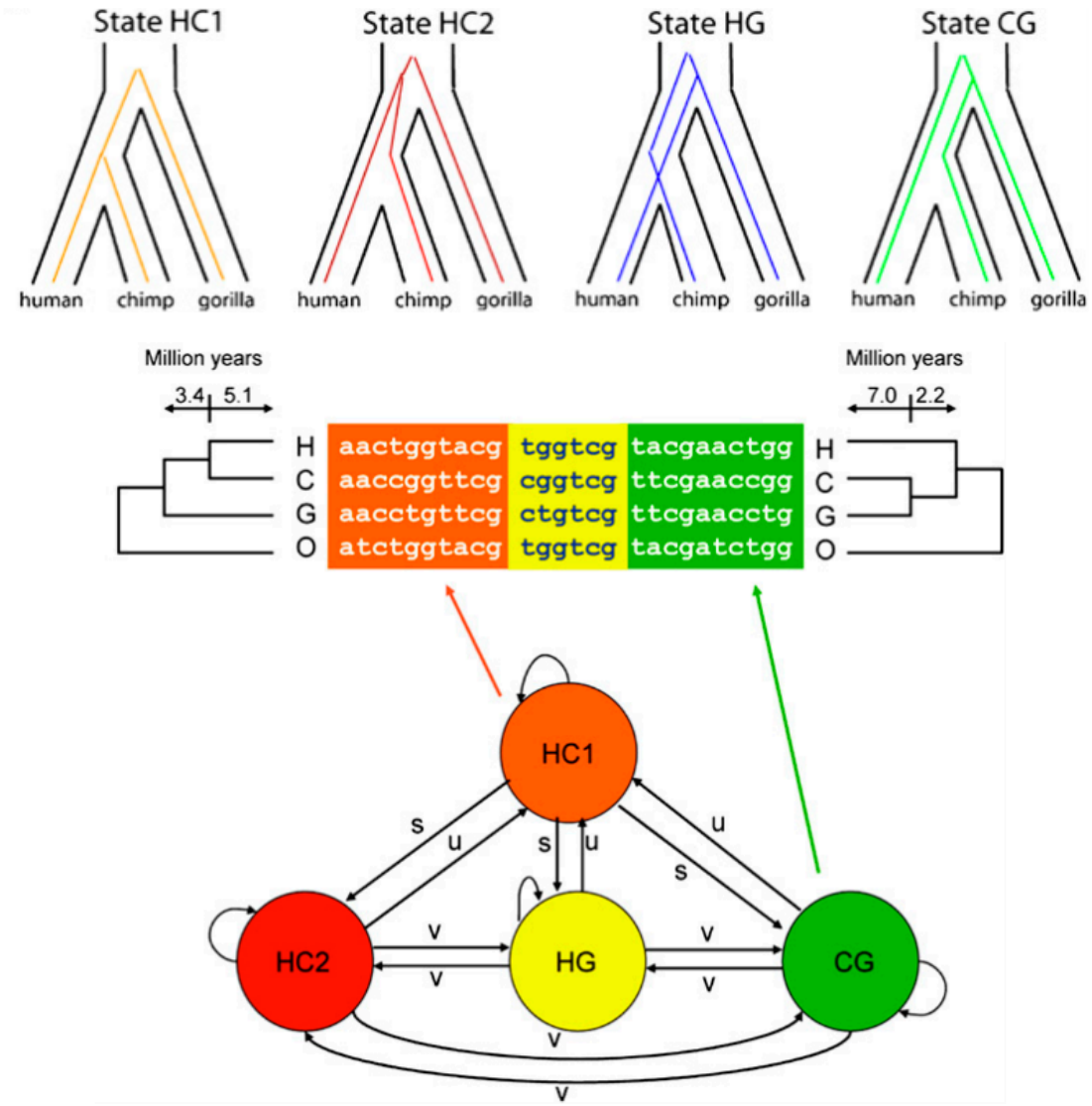


- When N_{xy}, N_{xyz} are small, τ_{xy} and τ_{xyz} approximate the divergence time well
- Otherwise, the coalescent time T_{xy}, T_{xyz} need to be taken into account

What if We Ignore Incomplete Lineage Sorting

- Aligned human (Hom), chimpanzee (Pan), gorilla (Gor), orangutan (Pon) sequences
- Two different estimated lineages
- Without consideration of ILS, substitution rates are overestimated

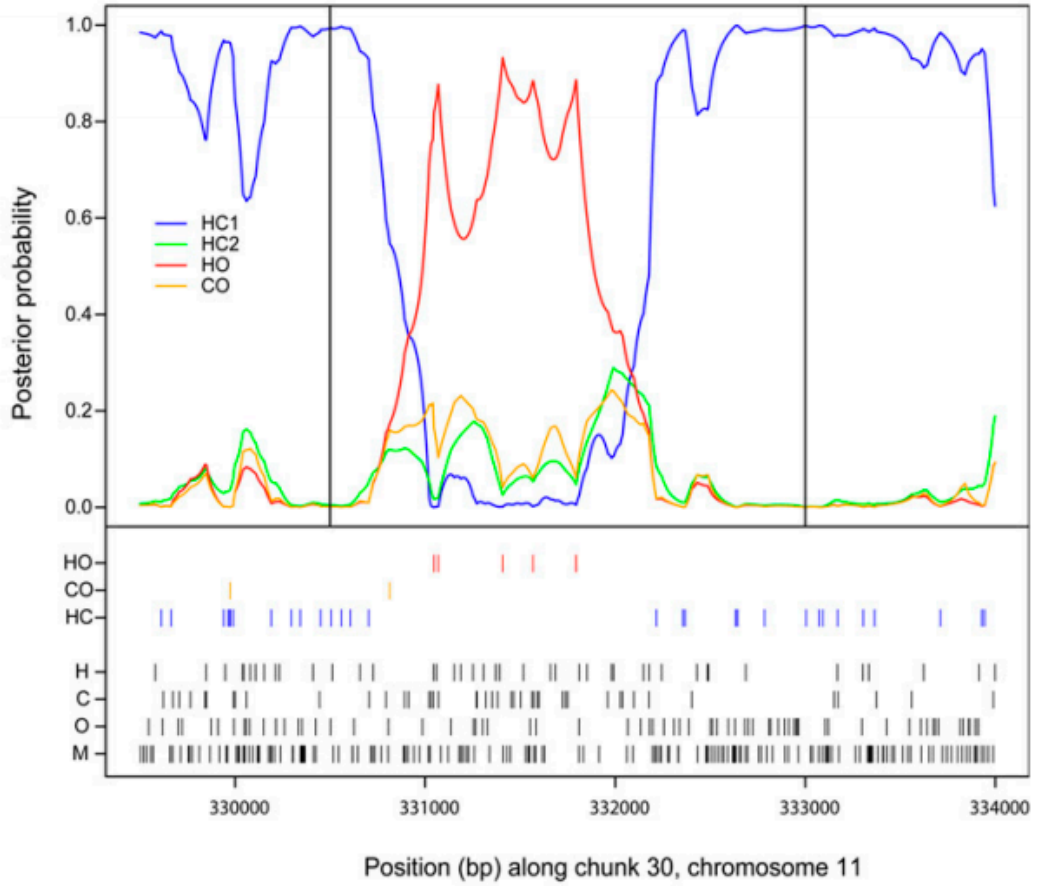
Coalescent-HMM (Hobolth et al., 2009,2011)



- Four states corresponding to different phylogenies with ILS
- Transitions to other states correspond to recombinations

Coalescent-HMM

- HC1 state (with no ILS) explains only ~50% of the loci
- Remaining states explain the other 50% proportioned roughly equally



Summary

- For a large population with a relatively short population history, the concepts of population genetics and evolution for speciation are closely related.

- Incomplete lineage sorting is observed among species when divergence happens in a relatively short period of time compared to the population size and coalescent time

Natural Selection

Divergence and Selection

- **Genetic drift**: the stochastic change in population frequency of a mutation due to the sampling process that is inherent in reproduction.

- **Selective sweep**: The process by which new favorable mutations become fixed so quickly that physically linked alleles also become fixed

- **Fixation**: the state where a mutation has achieved a frequency of 100% in a natural population

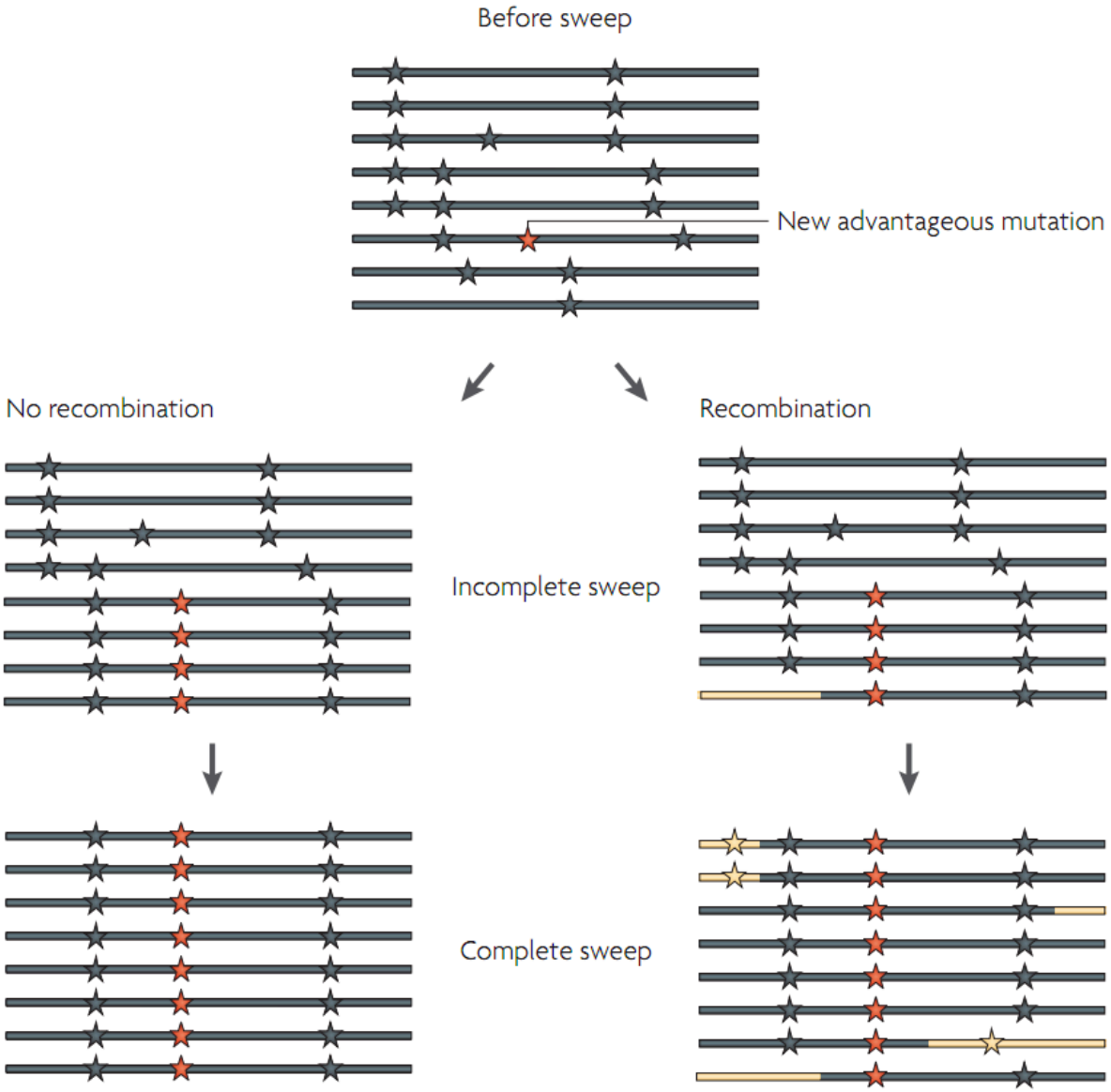
- **Hitchhiking** of neighboring polymorphisms

Complete vs Incomplete Sweep

- **Complete sweep**: the favored allele reaches a fixation
  - Local variation is removed except the ones that arise by mutation and recombination during the selective sweep

- **Incomplete sweep**: positively selected alleles are currently on the rise, but have not reached a frequency of 100%

Selective Sweep



Different Types of Natural Selection

- Directional selection
  - **Positive selection**
    - Reduces genetic diversity
    - The small region around the selected variation is over-represented
    - Hitch-hiking of neighboring alleles
  - **Negative selection** (purifying selection)
    - Can lead to background selection, where linked variations are also removed
    - Reduces genetic diversity, but no skewing in allele frequencies

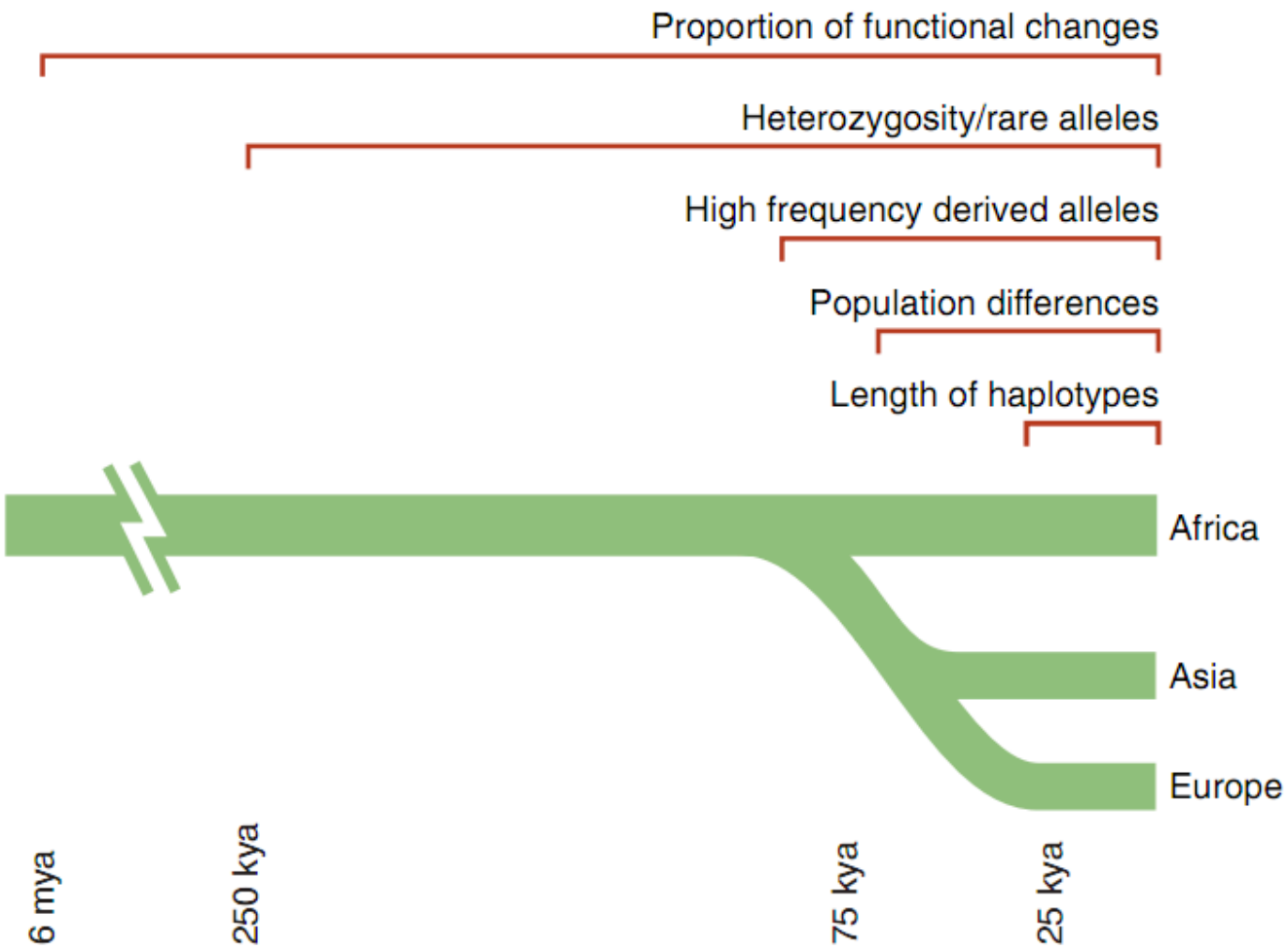
Different Types of Natural Selection

- **Balancing selection**: a selection for the maintenance of two or more alleles at a single locus in a population
  - Heterozygote advantage over homozygotes
    - G6PD mutation
      - G6PD enzyme deficiency and sickle-cell anemia in homozygote state
      - Partial protection against malaria in the heterozygous state
    - CFTR mutation
      - Causes cystic fibrosis in the homozygous state
      - Protects against asthma in the heterozygous state

- **Disruptive (diversifying) selection**: extreme values of the traits are favored than intermediate values

Time Scales for the Signatures of Selection

How do we detect the loci under natural selection from genome data?



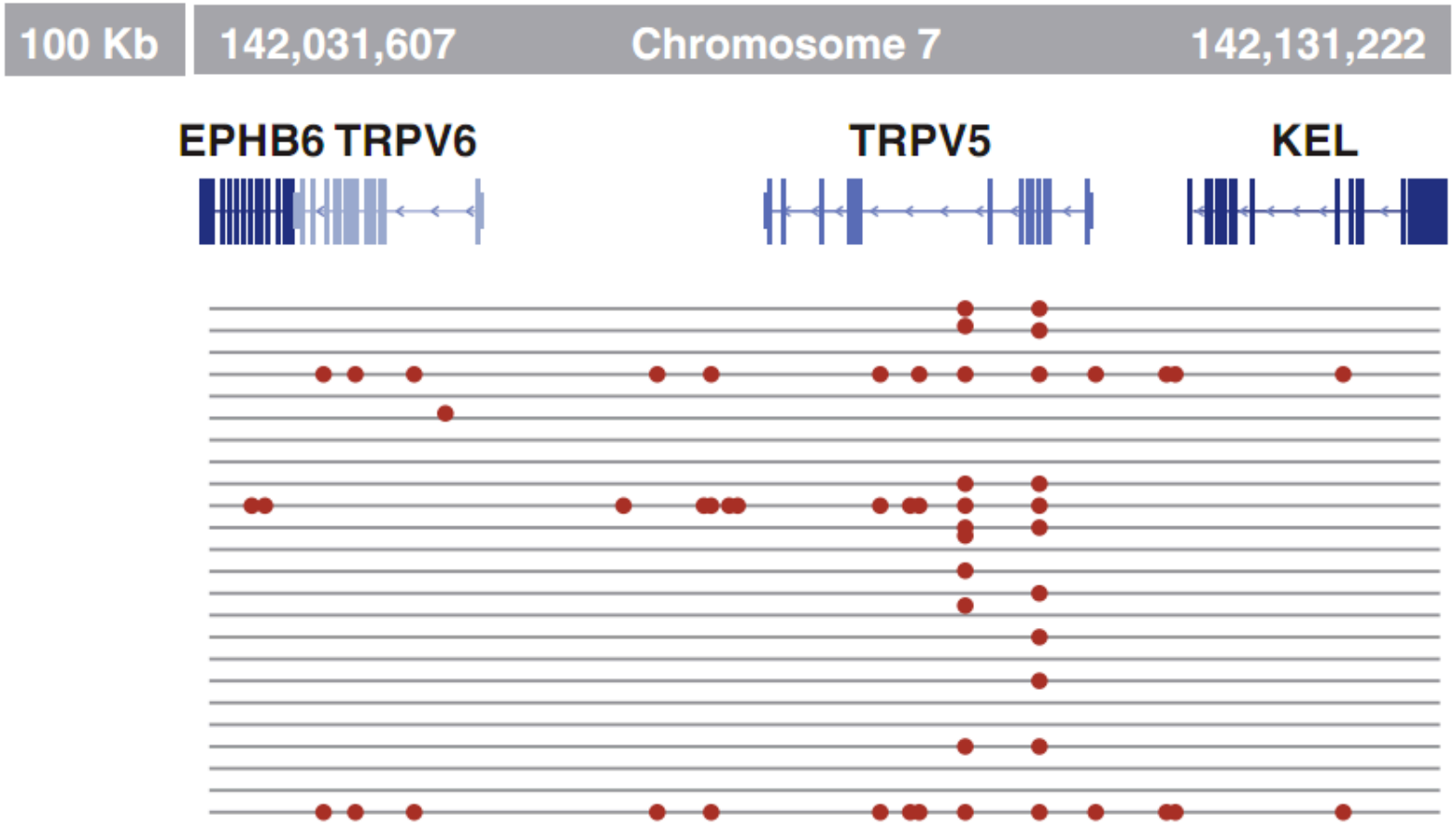
1. Reduction in Genetic Diversity

- Hitch-hiking of neutral mutations near beneficial mutations
- Selective sweep lowers the genetic diversity
- Subsequent mutations restores genetic diversity
  - Rare alleles at low frequency
- ~ 1 million years for a rare allele to drift to a high frequency allele

Low Diversity and Many Rare Alleles

- Speed of positive selection
  - Rapid selection
    - larger selected region
    - easier detection but harder to distinguish between hitchhiking and beneficial variations
- Recombination rate
  - Lower recombination rate leads to a larger selected region

Low Diversity and Many Rare Alleles



2. High-frequency Derived Alleles

- Derived vs ancestral alleles
  - Use alleles in closely related species for distinction
- Derived alleles typically have lower frequency, but under selective pressure, the frequency rises
- Many high-frequency derived alleles as a signature for positive selection
- High-frequency derived alleles rapidly drift to near fixation, thus persists for a shorter period time.

High-frequency Derived Alleles

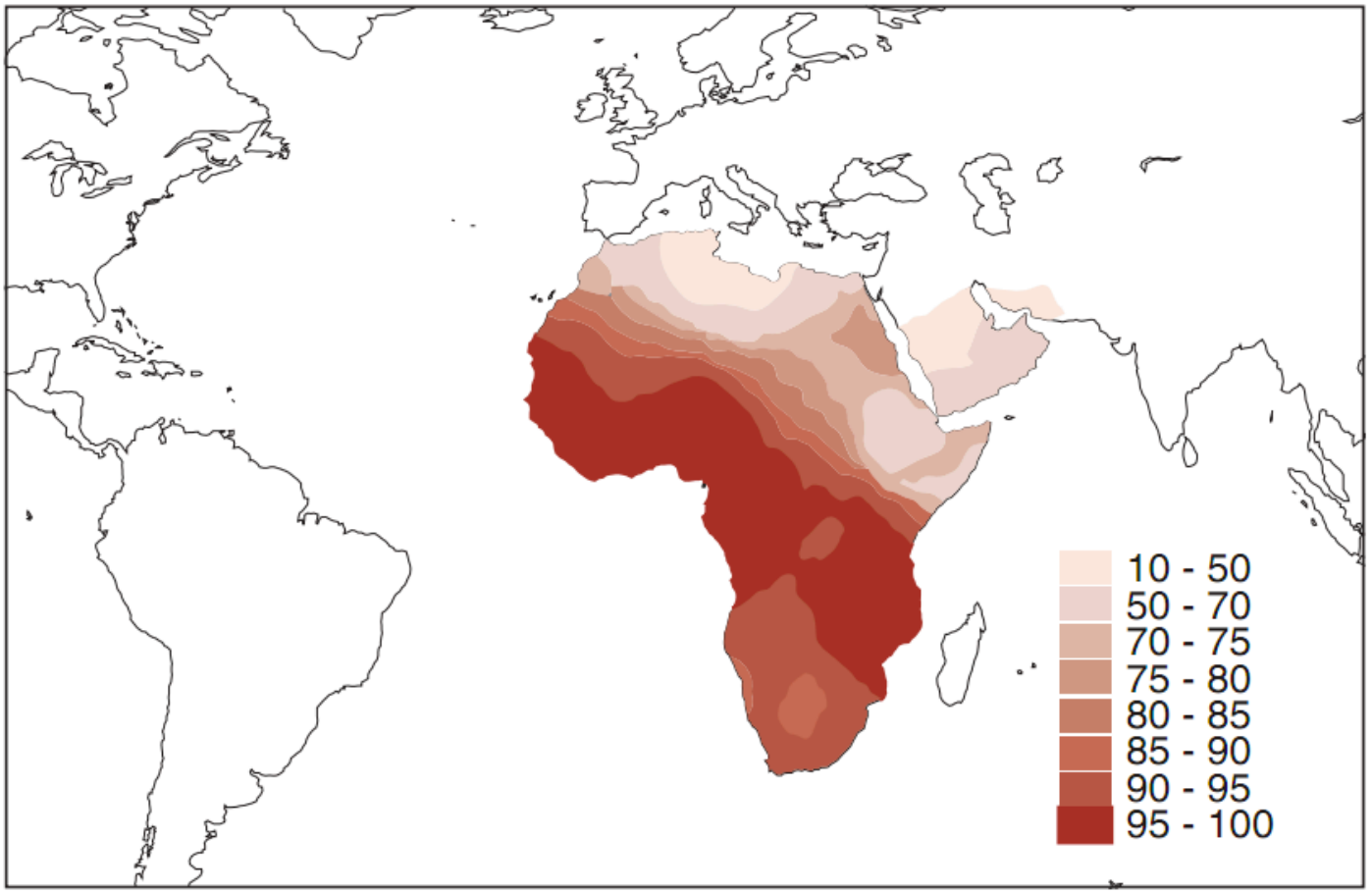
- Duffy red cell antigen in Africans: selection for resistance to P. vivax malaria (red: derived, grey: ancestral)

3. Population Differences

- Geographically separated populations and different selective pressure on different populations.
- Relatively large differences in allele frequencies between populations as a signature for positive selection.
- Useful only for positive selection after population differentiation (e.g., human migration out of Africa 50,000-75,000 years ago)
- F_{ST} can be used as test statistic

Extreme Population Differences

- *FY*O* allele for resistance to P. vivax malaria



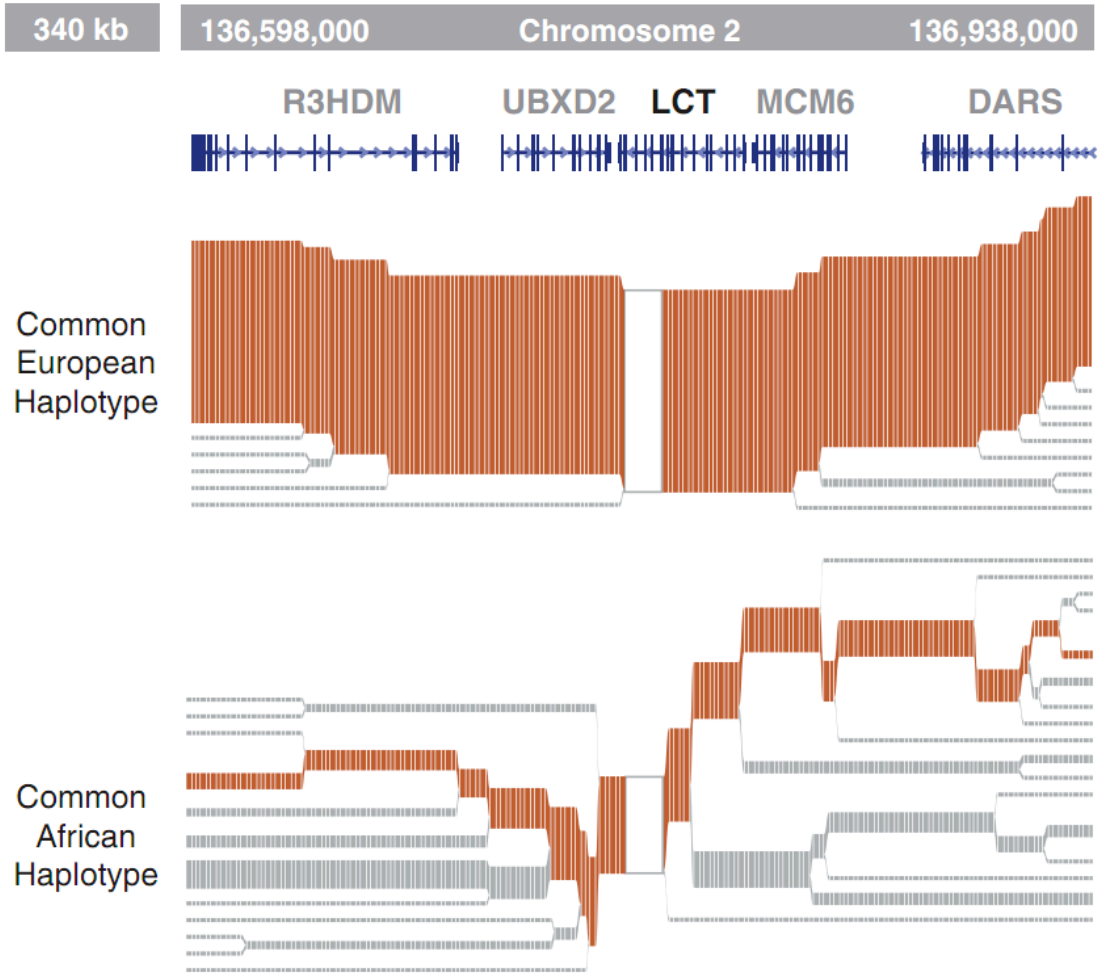
4. Long Haplotypes

- Alleles under recent positive selection, thus little break-downs by recombination
  - long haplotypes with high-frequency alleles

- This signature persists for a short period of time before recombination breaks down the chromosome.

Long Haplotypes

- *LCT* allele for lactase persistence (high frequency ~77% in European populations but long haplotypes)



Determining Function of the Candidate Loci

- What is the associated trait that is being selected, given the candidate locus for selection?
  - Examine the annotations of the genes
  - Look for associations to phenotypes in a population
  - Use comparative genomics
    - e.g., SLC24A5 gene positively selected in human population. Zebrafish homolog of this gene determines pigmentation phenotype.
  - It is harder to identify function for variations that reached fixation in human population
    - e.g., speech-related genes

Natural Selection and Diseases

- Positive selection
  - Response to pathogens, environmental conditions, diet.
  - Beneficial mutations for infectious diseases can be selected rapidly

- Negative selection
  - Deleterious mutation with only small or moderate effects can reach a low-level frequency with relative low selective pressure

Difficulties in Detecting Natural Selection

- Confounding effects of demography
  - Population bottleneck and expansion can leave signatures that look like a positive selection
- Ascertainment bias for SNPs
  - Regions where many sequences were used for ascertainment may appear to have more segregating alleles at low frequencies with more haplotypes.
- Recombination rate
  - Strong signature for selection for regions with low recombination rates