Population Genetics I: Genetic Polymorphisms, Haplotype Inference, Recombination

02-710 Computational Genomics

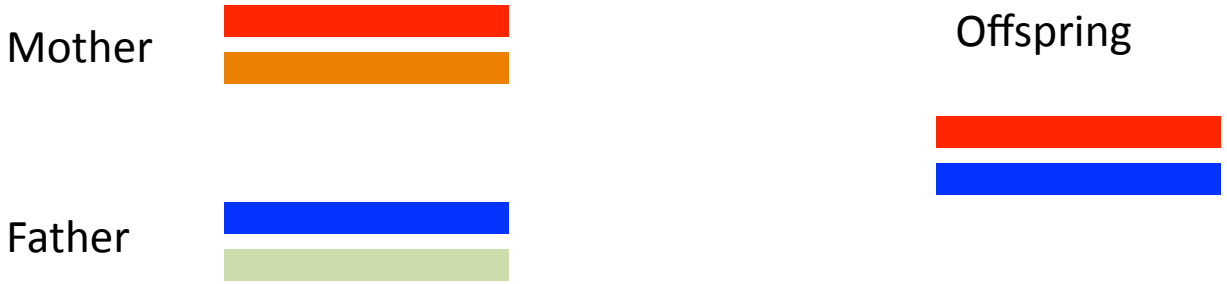
Seyoung Kim

Overview

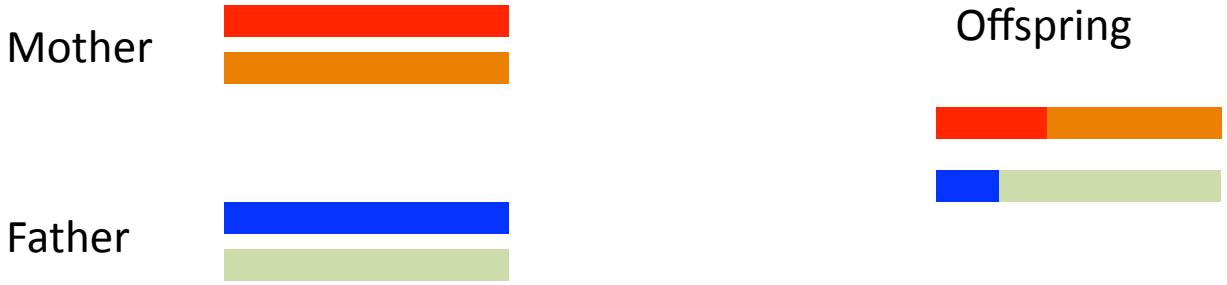
- Two fundamental forces that shape genome sequences
  - Recombination
  - Mutation, genetic polymorphisms
- A brief history: From Human Genome Sequencing Project to HapMap Project to 1000 Genome Project
- Haplotype inference, recombination rate estimation, linkage disequilibrium

Recombination

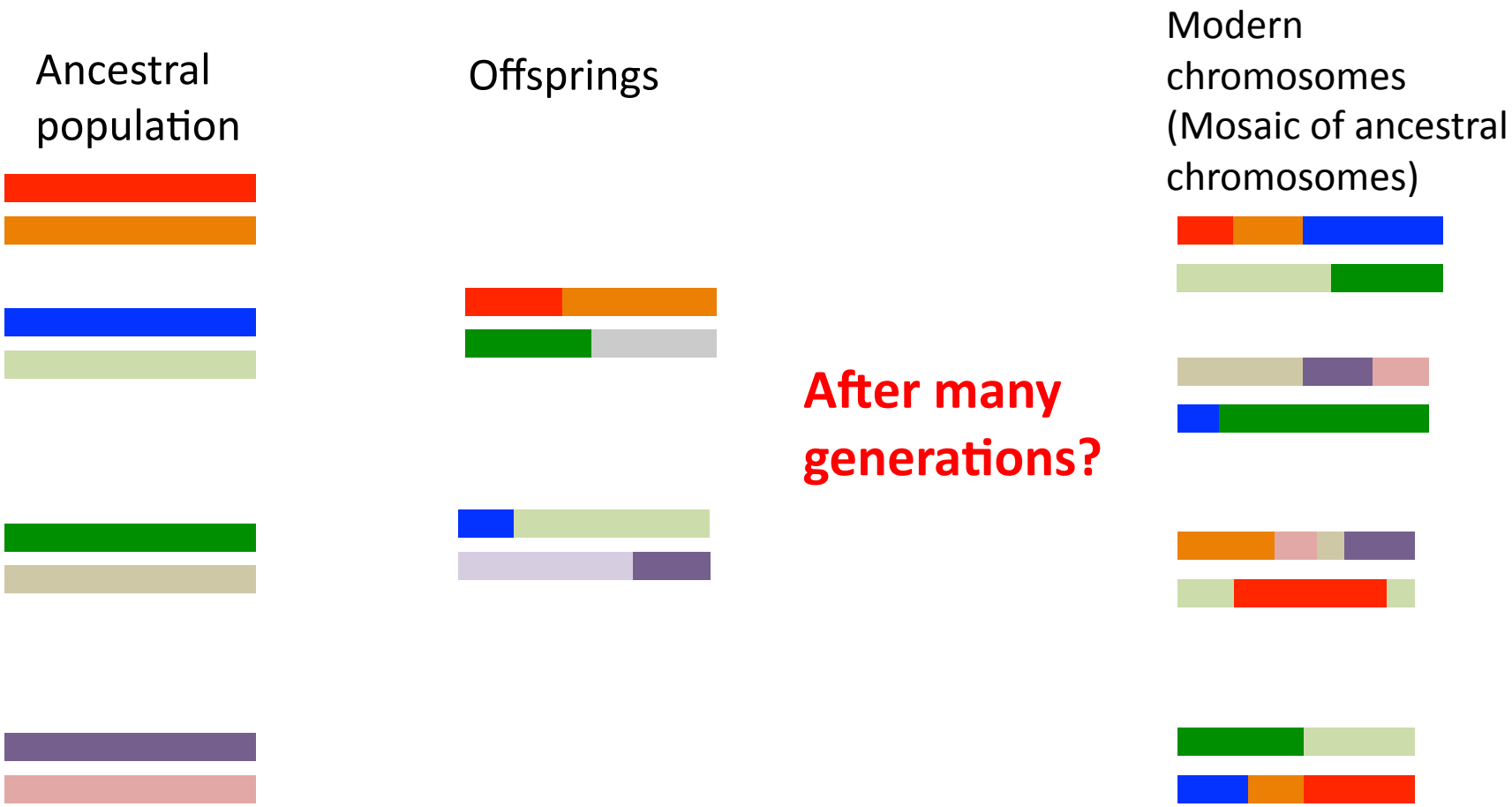
- Inheritance of genetic material without recombination



- Inheritance of genetic material with recombination



Recombination Shapes Genome Structure



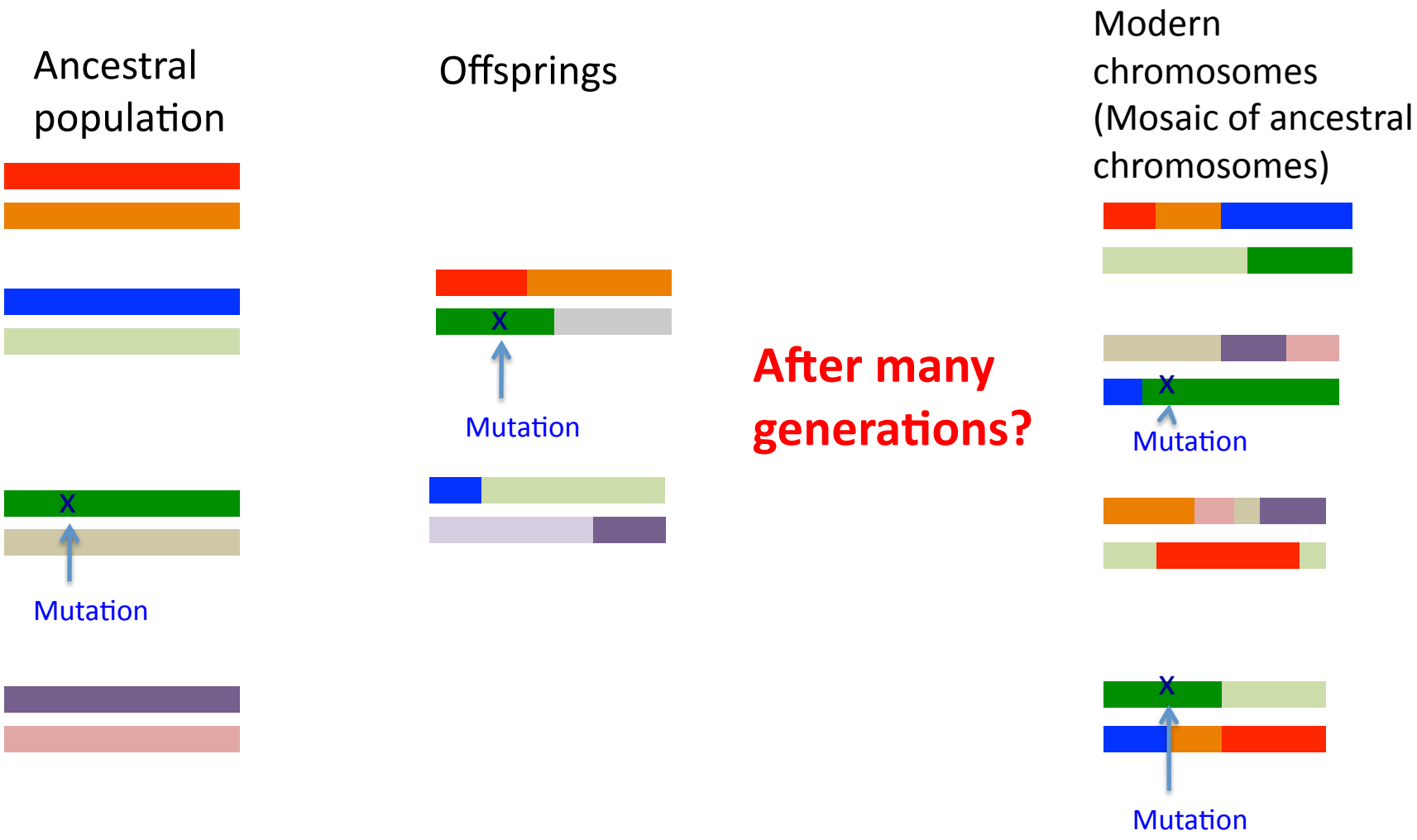
Non-uniform recombination frequencies across genomes
Recombination hotspots

Mutations

- A natural process that changes a DNA sequence
- As a cell copies its DNA before dividing, a "typo" occurs every 100,000 or so nucleotides
- “germline” mutations are inherited by the offsprings
- Some mutations are benign, others can be deleterious
- Mutations create genetic diversity in the population: genetic polymorphisms

http://learn.genetics.utah.edu/content/variation/mutation/

Mutations Create Genetic Diversity



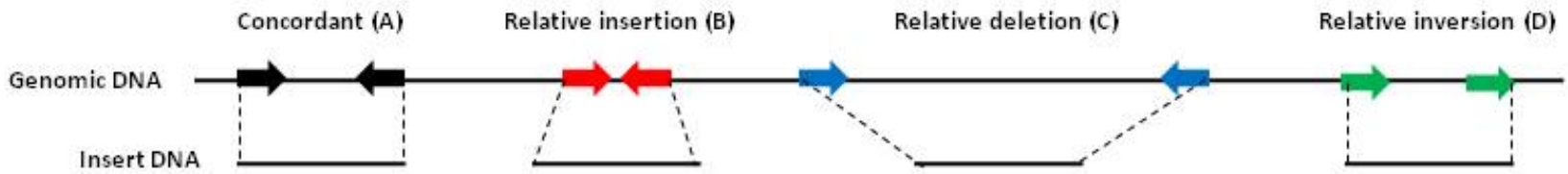
Single Nucleotide Polymorphisms (SNPs)

Involves a flip of a single nucleotide

```
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgaggcgtatcc···
···cagttaccgtgcacgatagctagcaatcatctagcactatgctgagacgtatcc···
···cagttaccgtgcatgatagctagcaatcatctagcactatgctgagacgtatcc···
```

Other Types of Genetic Polymorphisms

- Structural variants
  - insertions/deletions, duplications, copy number variations

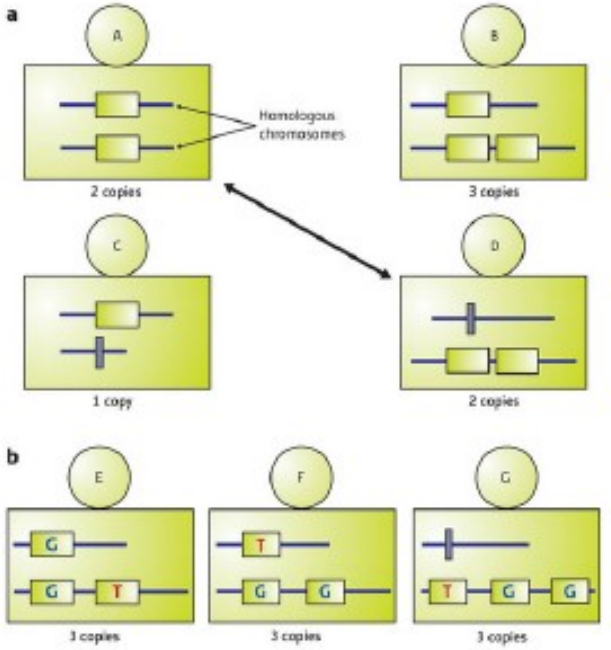


Other Types of Genetic Polymorphisms

- Insertion/deletion of a section of DNA
  - Minisatellites: repeated base patterns (10-60 base pair fragments repeated 5-50 times)
  - Microsatellites: 2-4 nucleotides repeated
  - Presence or absence of Alu segments

Genetic Polymorphisms

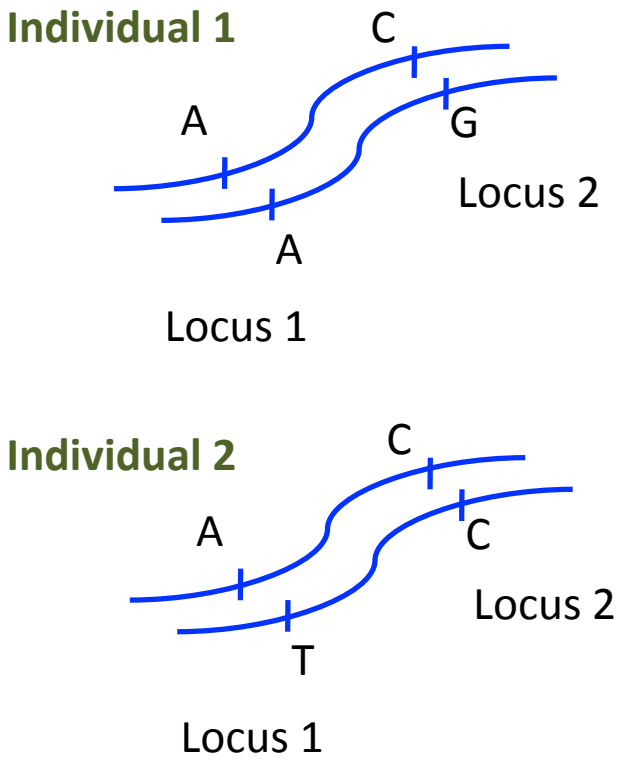
- Copy Number Variation
  - DNA segment whose numbers of occurrences in genomes differ in different individuals
    - Kilobases to megabases in size
  - Usually two copies of all autosomal regions, one per chromosome
  - Variation due to deletion or duplication



Copy-number variation (CNV) can occur in ambiguous patterns. (a) Individuals in a population may have different copy numbers on homologous chromosomes at CNV loci. For example, here individual A and D have two copies, although the patterns are different: A has one copy on each chromosome, whereas D has two on one chromosome and zero on the other. (b) Individuals may also have CNVs that contain SNPs. For example, individuals E, F, and G each have three copies, but the patterns can be distinguished by the numbers of copies on each chromosome and variations defined by SNPs.

Terminology

- Allele: different forms of genetic variations at a given gene or genetic locus
  - Locus 1 has two alleles, A and T, and Locus 2 has two alleles, C and G
- Genotype: specific allelic make-up of an individual's genome
  - Individual 1 has genotype AA at Locus 1 and genotype CG at Locus 2
- Heterozygous/Homozygous
  - Locus 1 of Individual 1 is homozygous, and Locus 2 is heterozygous



SNPs are bi-allelic.
Micro/minisatellites have many alleles, very informative because of the high heterozygosity (the chance that a randomly selected person will be heterozygous)

What Can We Learn from Genetic Variation?

- **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 linked to disease phenotypes
  - Association study: check for differences in SNP patterns between cases and controls

- **Forensic analysis**: the polymorphisms provide individual and familiar signatures

Overview

- Two fundamental forces that shape genome sequences
  - Recombination
  - Mutation, genetic polymorphisms
- A brief history: From Human Genome Sequencing Project to HapMap Project to 1000 Genome Project
- Haplotype inference, recombination rate estimation, linkage disequilibrium

Population Genetics

- Evolution vs. population genetics
  - Evolution studies genetic variation across species
  - Population genetics studies genetic variation within species

R.A. Fisher (1890-1929): geneticist & statistician

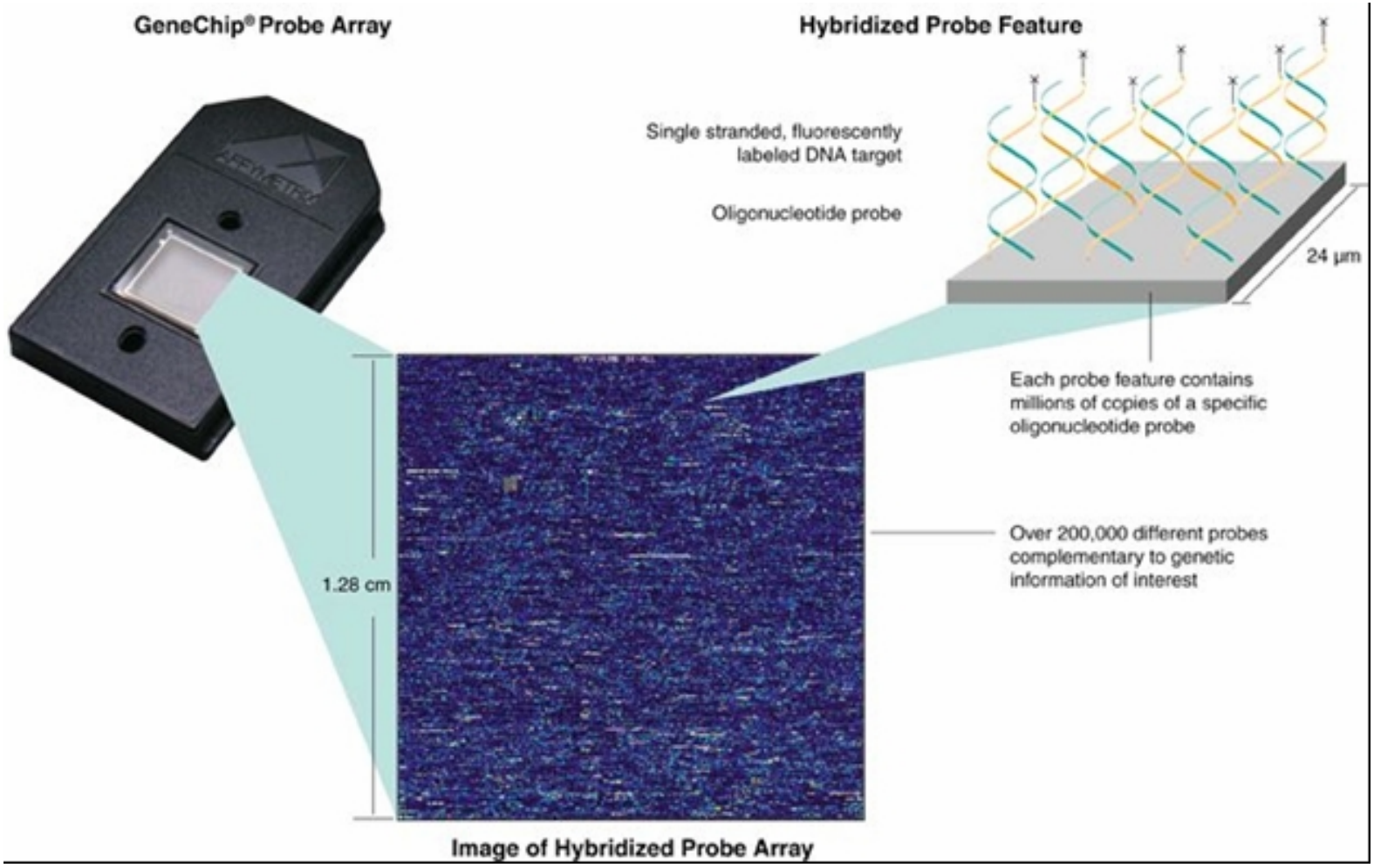
A Little Bit of History I

- 2001: The first draft of a human genome sequence become available
- 2001: The International SNP Map Working Group publishes a SNP Map of 1.42 million SNPs that contained all SNPs identified so far

Why SNPs?

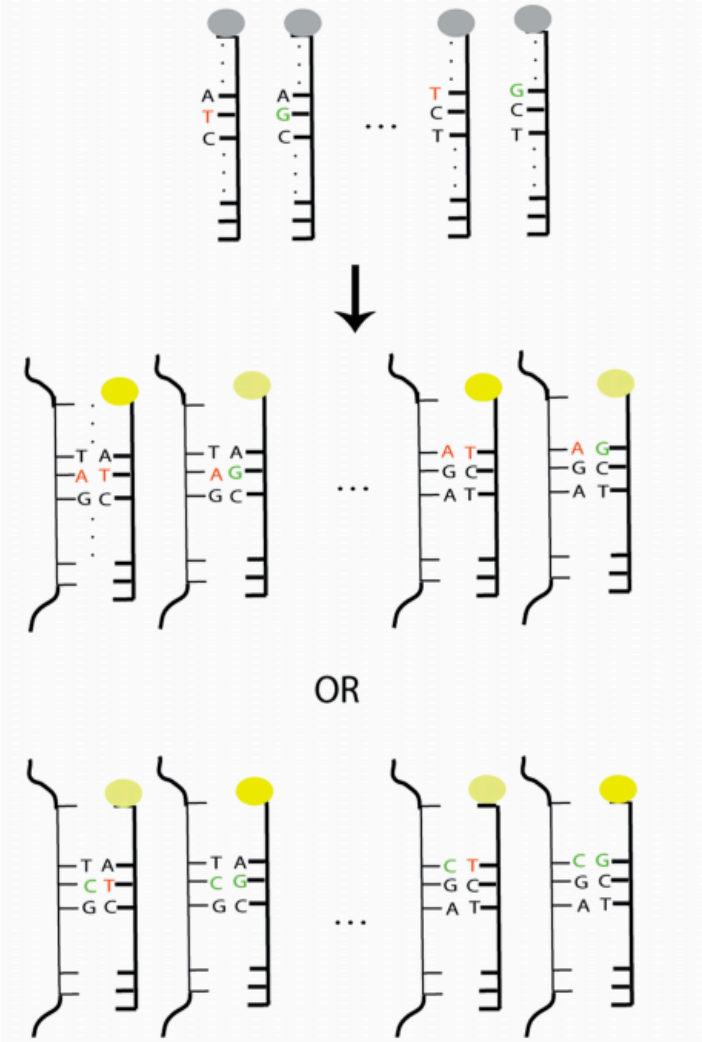
- Abundance: high frequency on the genome
- Position: throughout the genome
  - coding region, intron region, promoter site
- Ease of genotyping (high-throughput genotyping)
- SNPs account for around 90% of human genomic variation
- About 40 million or more SNPs exist in human populations
- Most SNPs are outside of the protein coding regions
- More than 5 million common SNPs each with frequency 10-50% account for the bulk of human DNA sequence difference
- It is estimated that ~60,000 SNPs occur within exons; 85% of exons are within 5 kb of the nearest SNP
- Account for most of the genetic diversity among different (normal) individual, e.g. drug response, disease susceptibility
- However, only two alleles at each locus, less informative than microsatellites. (Use haplotypes!)

Affymetrix GeneChip Probe Array



SNP Genotyping with SNP Array

- The SNP chip's basic design is similar to that of expression arrays
  - An array of 25 bp oligonucleotide sequences (features) is laid across the surface of the chip.
  - The sample's DNA is amplified, and hybridized to the array.
  - The array is scanned to quantify the relative amount of sample bound to each probe for different alleles.
- For SNPs, there is a pair of probes: one for each of the alleles.



A Little Bit of History II

- 2005: HapMap Phase I
  - Genotype at least one common SNP (minor allele frequency (MAF)>5%) every 5kb across 270 individuals
  - Geographic diversity
    - 30 trios from Yoruba in Ibadan, Nigeria (YRI)
    - 30 trios of European ancestry living in Utah (CEPH)
    - 45 unrelated Han Chinese in Beijing (CHB)
    - 45 nrelated Japanese (JPT)
  - 1.3 million SNPs


CBSNEWS.com

October 29, 2002 17:12:27 ▸ The Early Show ▸ CBS Evening News ▸

SCITECH ▪ Section Front

E-mail This Story Printable Version

The International 'HapMap' Project

WASHINGTON, Oct. 29, 2002

(AP) Looking for a quicker way to identify genes that cause disease, researchers are beginning a $100 million effort to identify blocks of DNA that contain common variations in the human genetic structure, officials announced Tuesday.

The project, expected to take three years, involves nine research groups in five

Hapmap.org

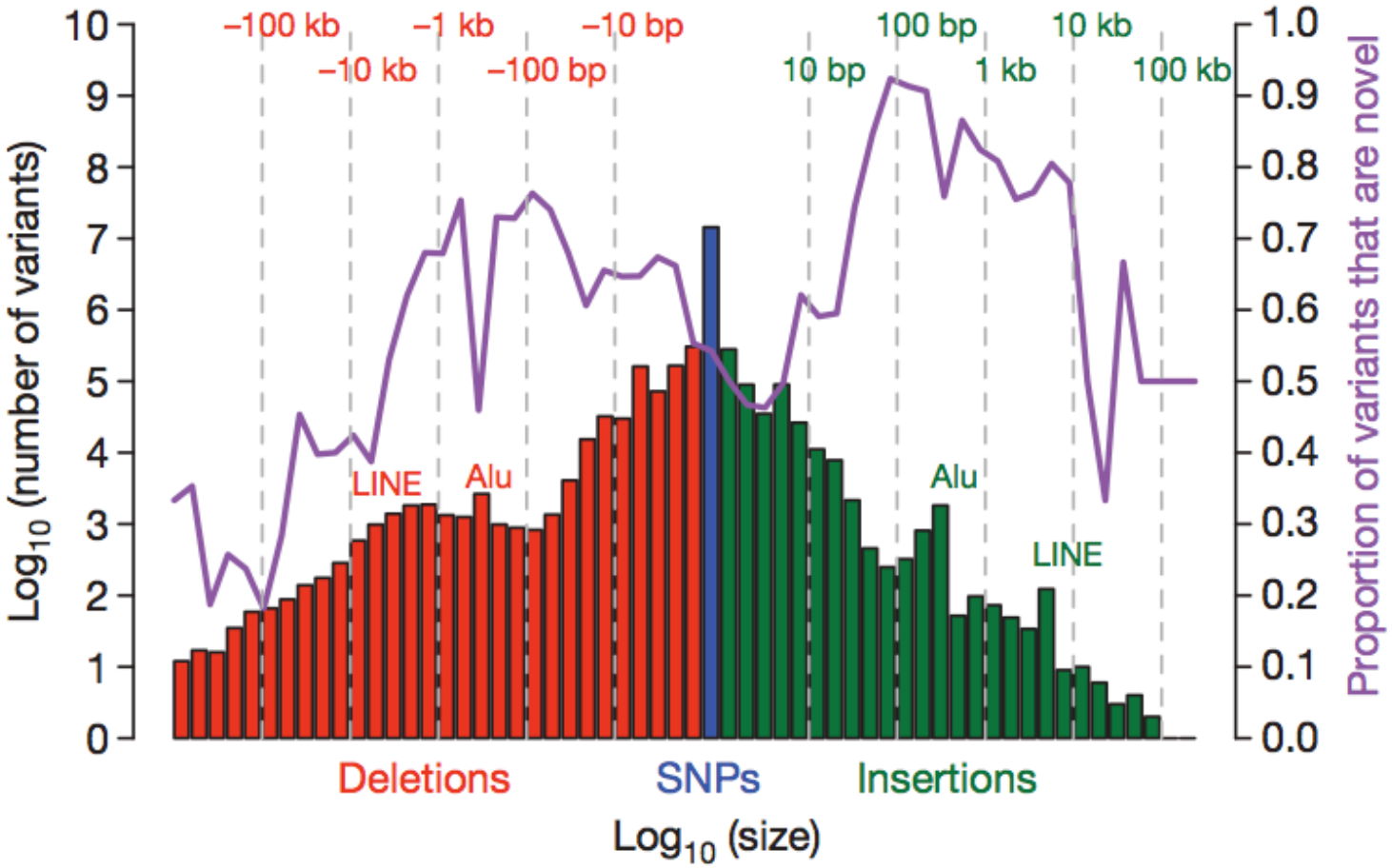
A Little Bit of History III

- 2007: HapMap Phase II
  - Genotype additional 2.1 million SNPs for the same individuals
  - SNP density about 1 per kb
  - Estimated to contain 25-35% of all 9-10 million common SNPs in assembled human genome.

- 2010: HapMap Phase III
  - 1184 individuals from 11 populations, including HapMap Phase I, II samples
  - Rare variants (MAF=0.05-0.5%), low frequency variants (MAF=0.5%-5%)
  - Copy number variations, resequencing of selected regions

- 2010 : 1000 Genome Pilot Project
  - A more complete characterization of human genetic variations

Variant Frequencies from 1000 Genome Pilot Project



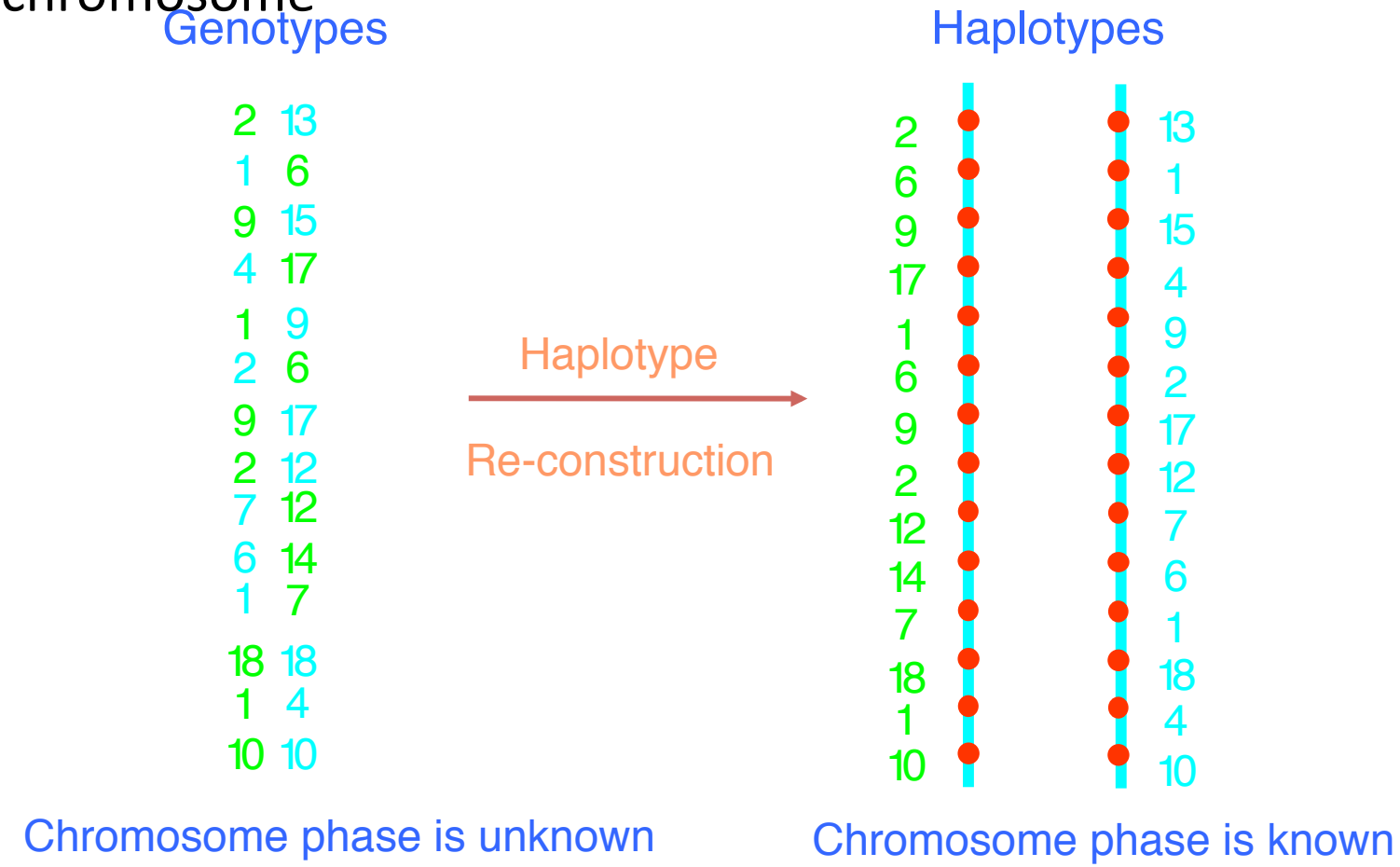
Frequency of SNPs greater than that of any other type of polymorphism

Overview

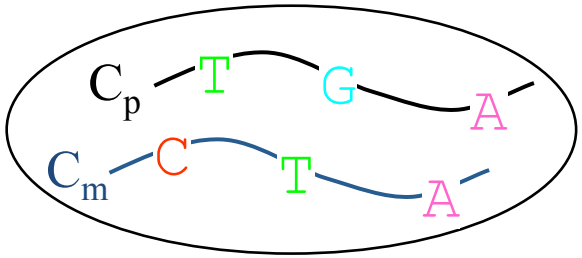
- Two fundamental forces that shape genome sequences
  - Recombination
  - Mutation, genetic polymorphisms
- A brief history: From Human Genome Sequencing Project to HapMap Project to 1000 Genome Project
- Haplotype inference, recombination rate estimation, linkage disequilibrium

Genotype and Haplotype

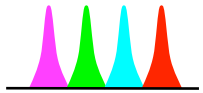
- Haplotypes: a collection of alleles derived from the same chromosome



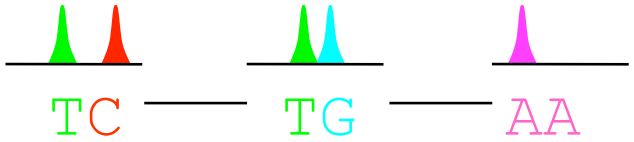
Phase ambiguity
-- haplotype reconstruction for individuals



Heterozygous diploid individual

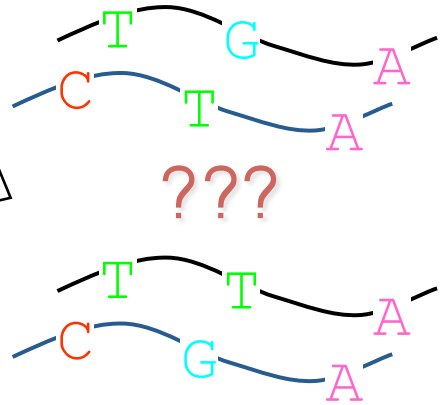


sequencing



Genotype g

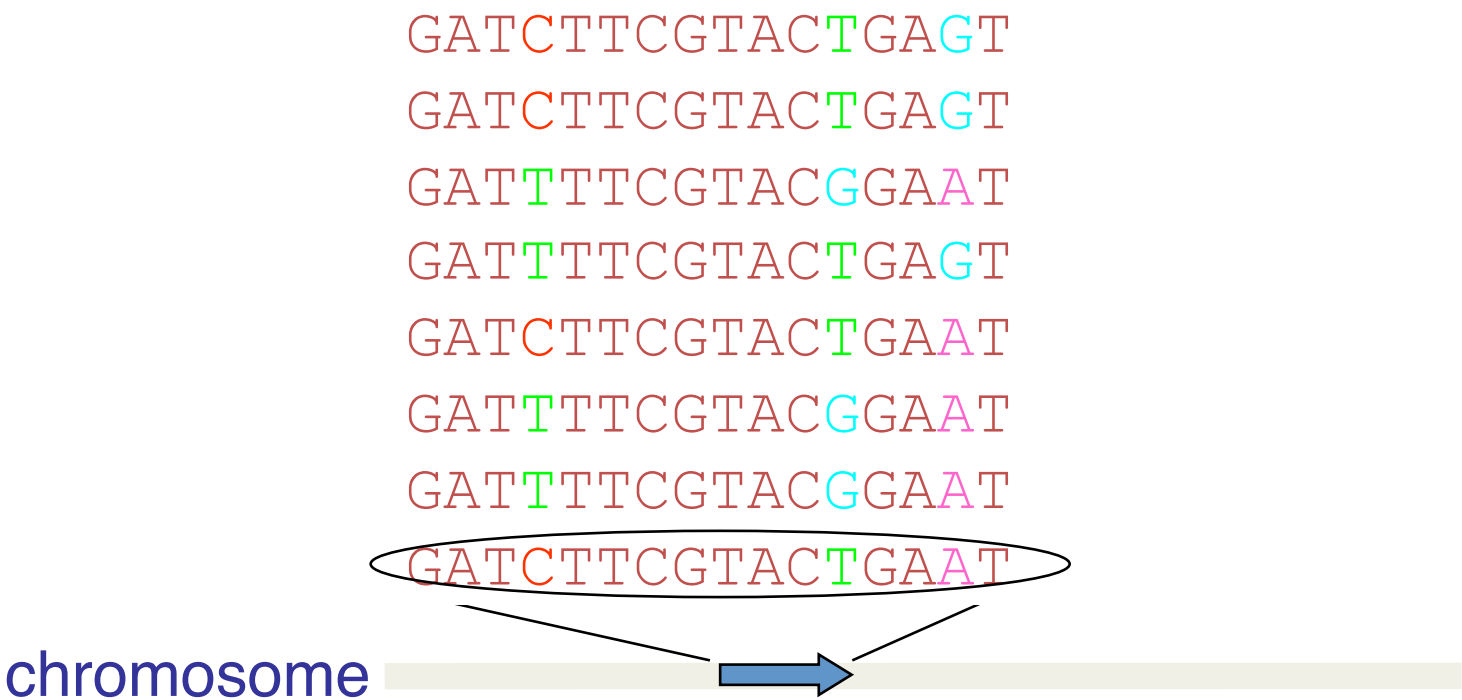
pairs of alleles with association of alleles to chromosomes unknown



haplotype $h \equiv (h_1, h_2)$

possible associations of alleles to chromosome

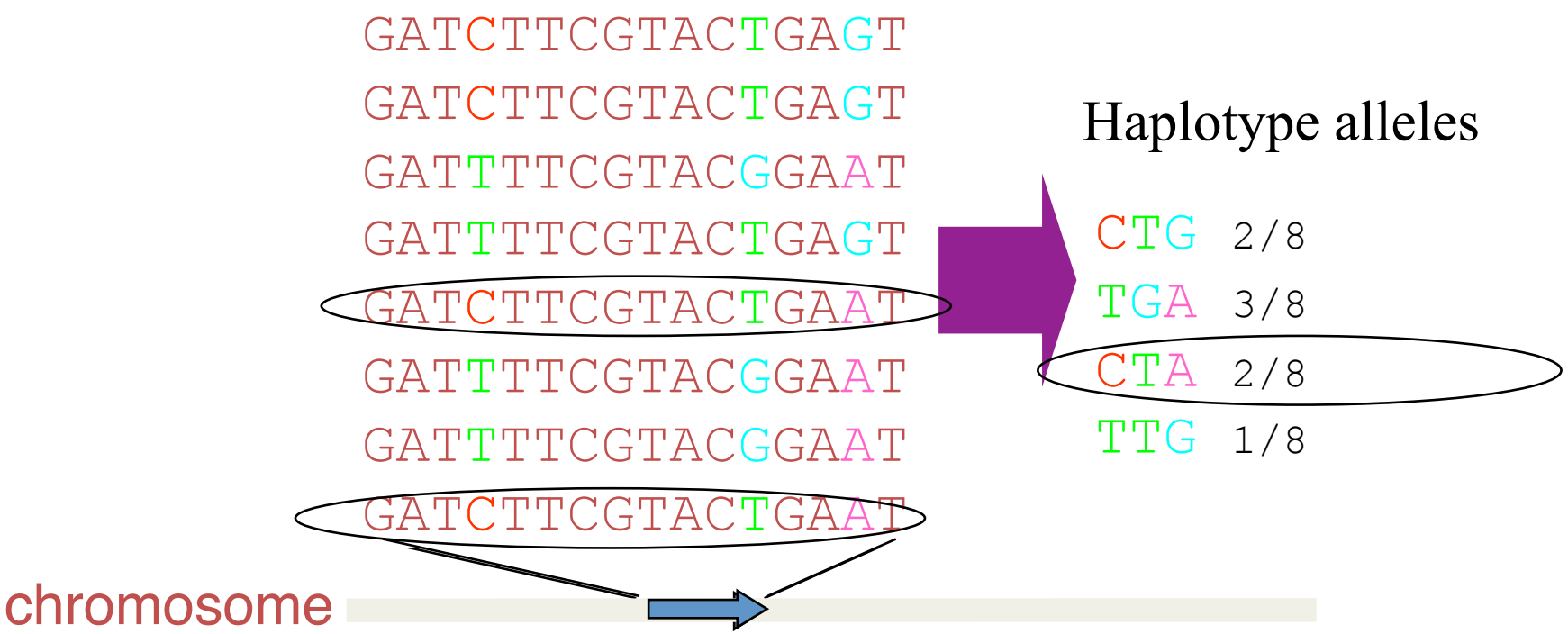
From SNPs to Haplotypes



- "Binary" nt-substitutions at a single locus on a chromosome: each variant is called an "allele" e.g., G vs. A alleles
- Haplotypes: a set of adjacent SNPs on the same chromosome

Why Haplotypes?

-- a more discriminative state of a chromosomal region



- Consider J SNPs (binary markers) in a genomic region
- There are 2^J possible haplotypes
  - but in fact, far fewer are seen in human population
- Good genetic marker for population, evolution and hereditary diseases …

Inferring Haplotypes

- Genotype: AT//AA//CG
  - Maternal genotype: TA//AA//CC
  - Paternal genotype: TT//AA//CG
  - Then the haplotype is AAC/TAG.

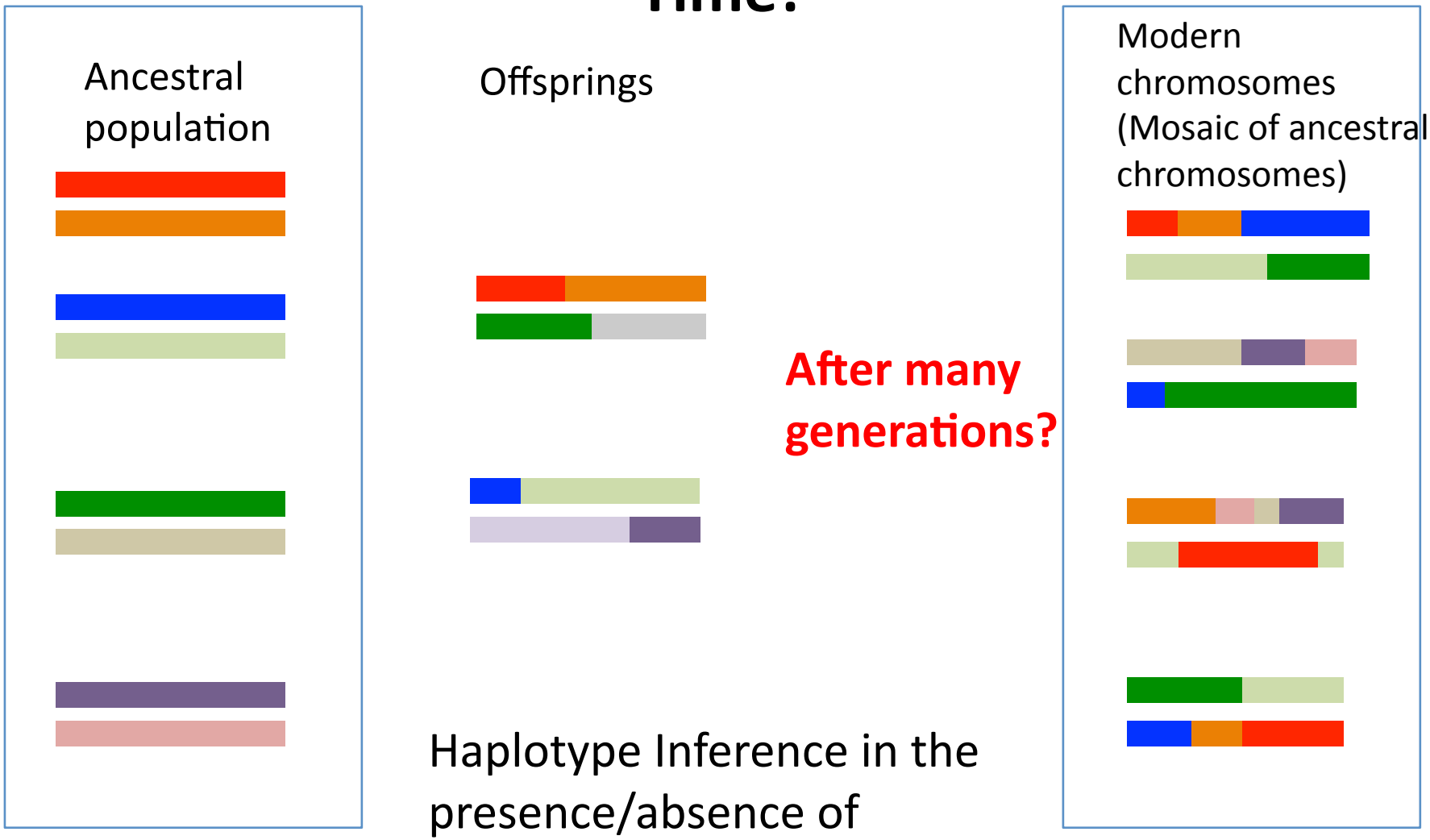
- Genotype: AT//AA//CG
  - Maternal genotype: AT//AA//CG
  - Paternal genotype: AT//AA//CG
  - Cannot determine unique haplotype

- **Problem**: determine Haplotypes without parental genotypes

Haplotype Inference (Phasing)

- Given a random sample of multilocus genotypes at a set of SNPs from a population of individuals
  - Frequency estimation of all possible haplotypes
  - Haplotype reconstruction for individuals

Recombination "Breaks" Long Haplotypes Over Time!



Haplotype Inference in the presence/absence of recombinations?

Haplotype Inference (Phasing)

- Haplotype reconstruction algorithm
  - Clark's parsimony algorithm (Clark, Mol. Biol. Evol. 1990)
    - Ignores recombination
  - PHASE (Li and Stephens, Genetics 2003)
    - Takes into account recombination
    - Recover haplotypes
    - Estimate recombination rate, recombination hotspots
    - Impute missing genotypes

Haplotype reconstruction: Clark (1990)

- Choose individuals that are homozygous at every locus (e.g. TT//AA//CC)
  - Haplotype: TAC
- Choose individuals that are heterozygous at just one locus (e.g. TT//AA//CG)
  - Haplotypes: TAC or TAG
- Tally the resulting known haplotypes.
- For each known haplotype, look at all remaining unresolved cases: is there a combination to make this haplotype?
  - Known haplotype: TAC
    - Unresolved pattern: AT//AA//CG
    - Inferred haplotype: TAC/AAG. Add to list.
  - Known haplotype: TAC and TAG
    - Unresolved pattern: AT//AA//CG
    - Inferred haplotypes: TAC and TAG. Add both to list.
- Continue until all haplotypes have been recovered or no new haplotypes can be found this way.

Problems: Clark (1990)

- Many unresolved haplotypes at the end
- Error in haplotype inference if a recombination of two actual haplotypes is identical to another true haplotype
- Clark (1990): algorithm "performs well" even with small sample sizes.

PHASE

(Stephens et al., AJHG 2001)

- Key idea: Construct an HMM p(haplotypes, genotypes)that models the observed genotype sequences as observed, and models the unknown haplotypes as unovserved variables

- Haplotypes can be inferred as haplotypes that maximize p(haplotypes | genotypes): Viterbi algorithm!

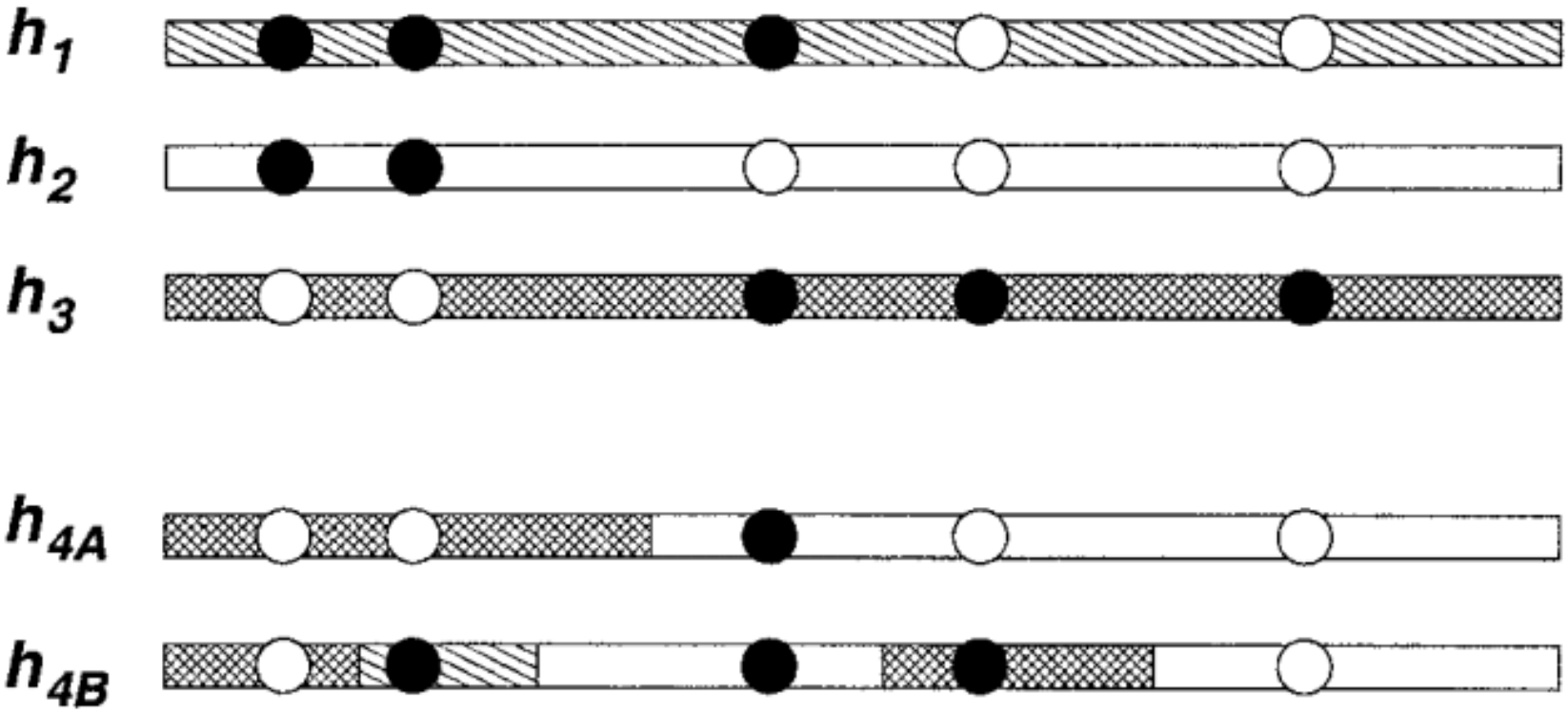
PHASE

(Stephens et al., AJHG 2001)

- HMM for PHASE
  - States: ancestral haplotypes
  - Transition probabilities: models recombination events
  - Emission probabilities: generates genotypes and models mutations

PHASE

- h_1, h_2, h_3: unobserved ancestral haplotypes
- h_{4A}, h_{4B}: unobserved haplotypes for individuals
- Circles: alleles, mutations



PHASE

(Stephens et al., AJHG 2001)

- Given the genotype data G, start with some initial haplotype reconstruction $H^{(0)}$. At each iteration $t = 0, 1, 2, \ldots,$ obtain $H^{(t+1)}$ from $H^{(t)}$ using the following three steps:

  1. Choose an individual, i, uniformly and at random from all ambiguous individuals (i.e., individuals with more than one possible haplotype reconstruction).

  2. Sample $H_i^{(t+1)}$ from $P(H_i | G, H_{-i}^{(t)})$, where H_{-i} is the set of haplotypes excluding individual i.

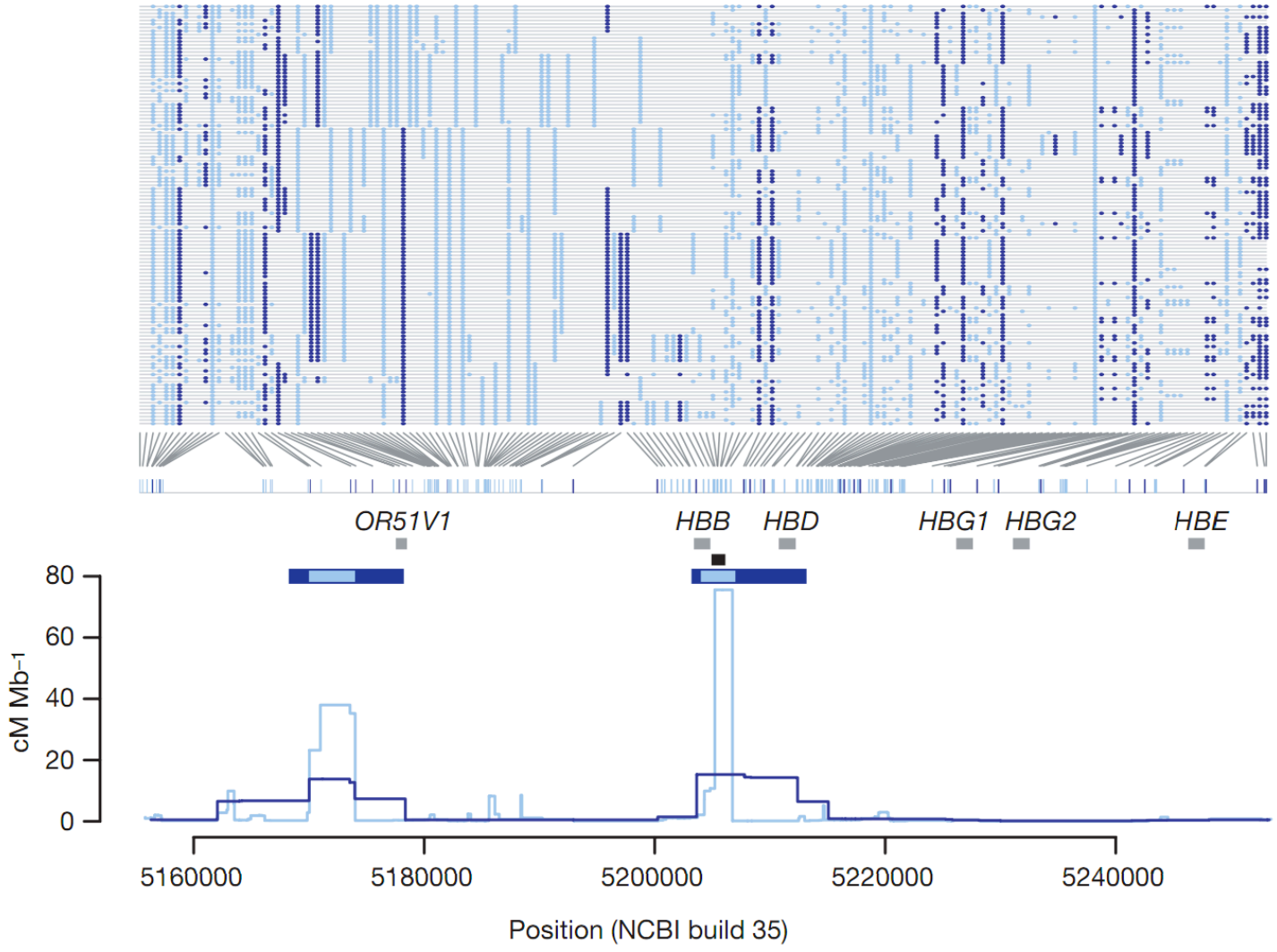
  3. Set $H_j^{(t+1)} = H_j^{(t)}$ for $j=1,\ldots,n$, $j \neq i$

Parameter Estimation in PHASE

- Gibbs sampling approach
  - Start with initial guesses on haplotypes
  - Iteratively reconstruct the haplotype of each individual assuming the haplotypes of other individuals have been correctly reconstructed.

- Robust to “moderate” levels of recombinations
- More accurate than Clark’s and many other algorithms
- Provides estimates of the uncertainty associated with each phase call

Haplotype Structure and Recombination Rate Estimates: HapMap I vs. HapMap II

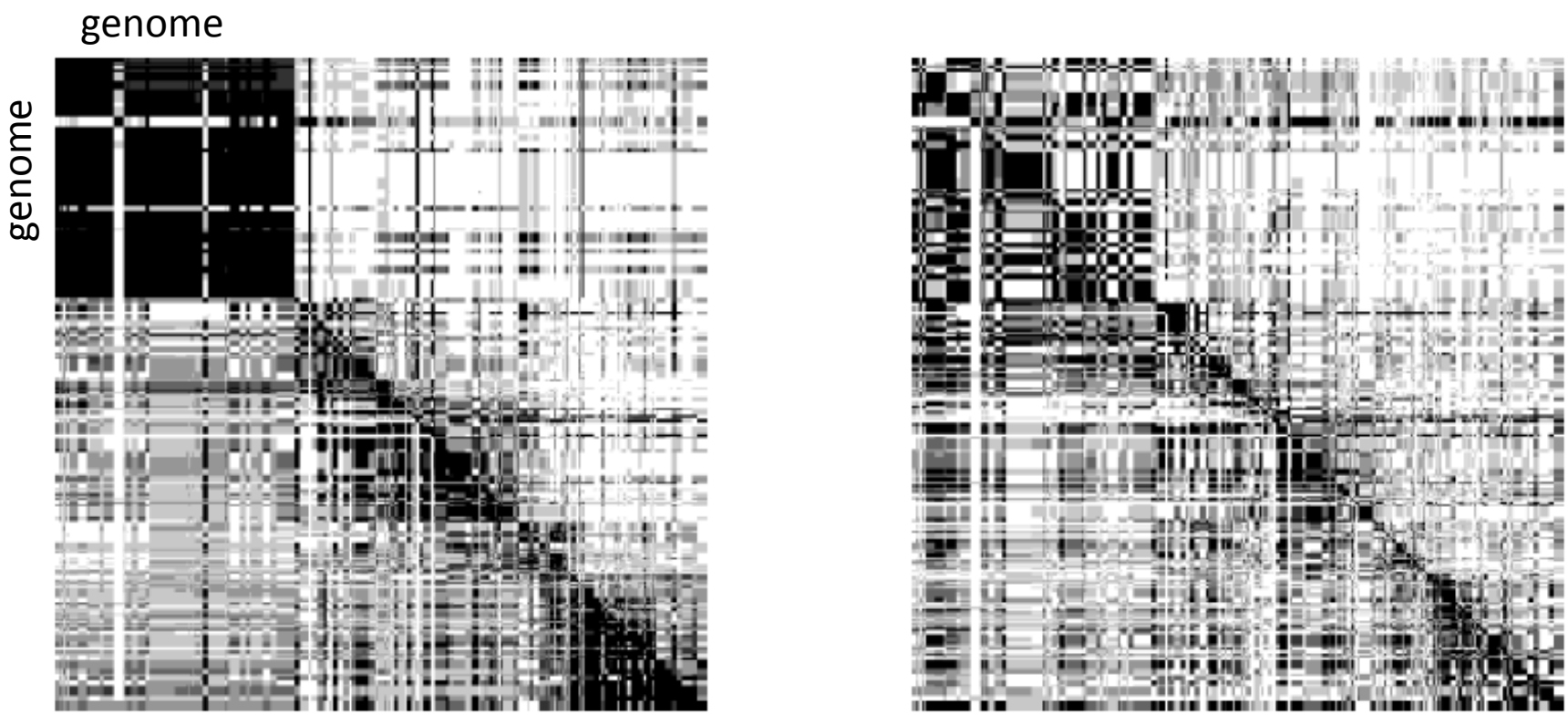


Linkage Disequilibrium (LD)

- LD reflects the relationship between alleles at different loci.
  - Linkage equilibrium: alleles at different loci are NOT linked and inherited to offsprings independently
  - Linkage disequilibrium: alleles at different loci ARE LINKED. LD is an allelic association measure

Linkage Disequilibrium in HapMap Data

- r^2 in HapMap Data



Two different populations in upper/lower diagonal

Haplotype Analyses

- Haplotype analyses
  - Linkage disequilibrium assessment
  - Disease-gene discovery
  - Genetic demography
  - Chromosomal evolution studies

Summary

- Recombination and mutation shape genomes over time in population
- Different types of genetic polymorphisms and why they are useful
- What is LD and how is LD structure created in the genomes
- Haplotype inference
  - PHASE is the most commonly used software for this task