# STAT588/BIOL588: Genomic Data Science
# Lecture 4: Introduction to Population Genetics

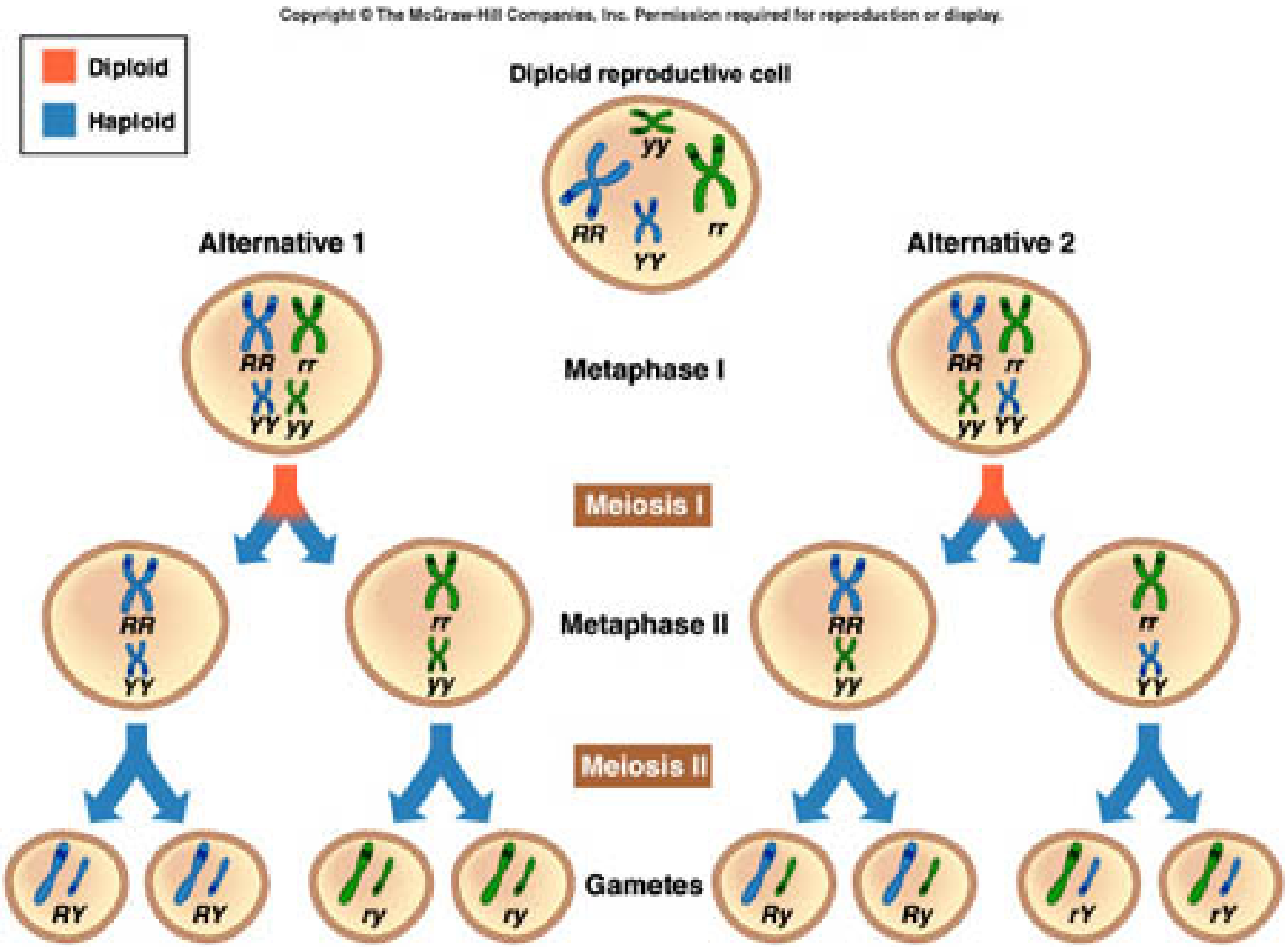
Guest Lecture: Dr. Shannon Davis
Department of Biological Sciences

# Objectives of Lecture 4

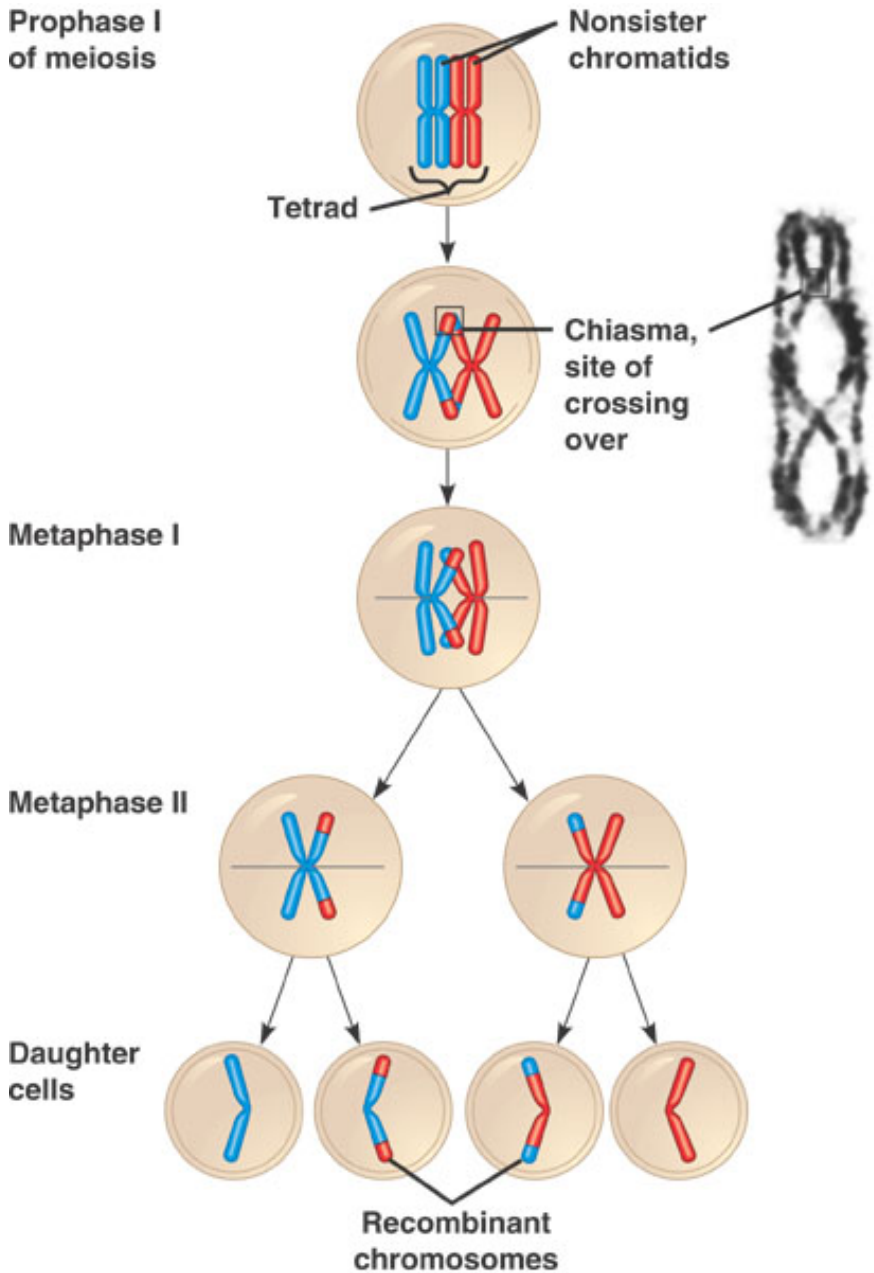
Review basic terminology of population genetics

- Crossing Over
- DNA Recombination
- Genetic Markers
- Genetic Association Analysis
- Online Resources

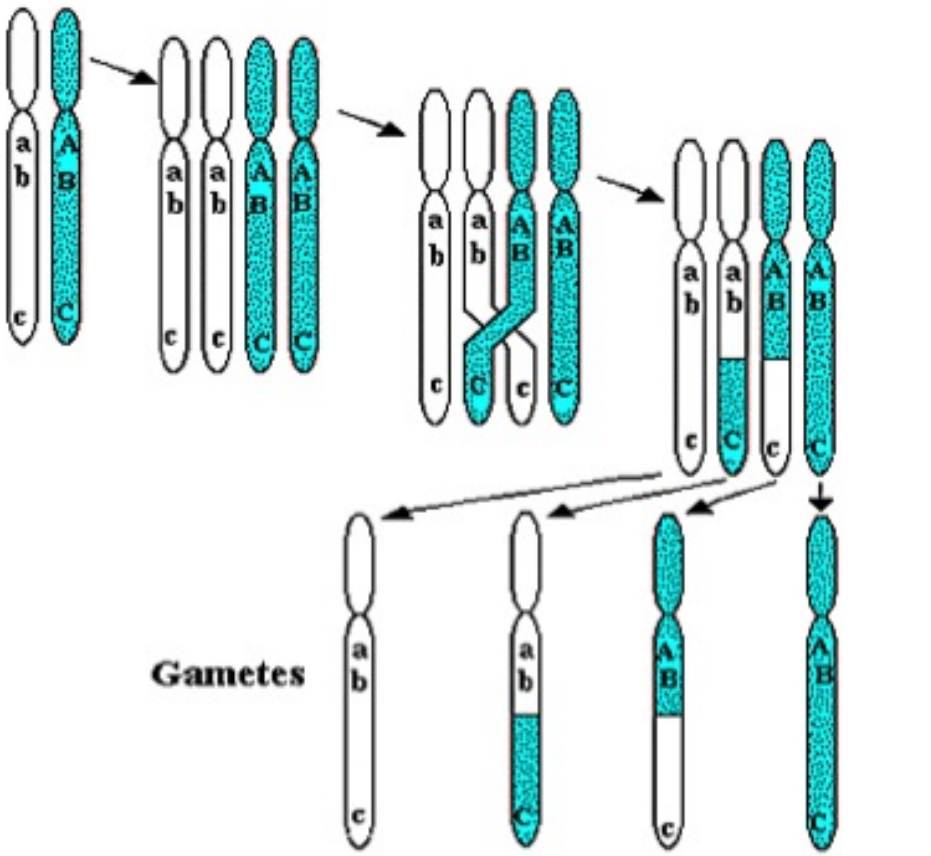
# Random Combinations of Gametes



# Crossing Over



# DNA Recombination

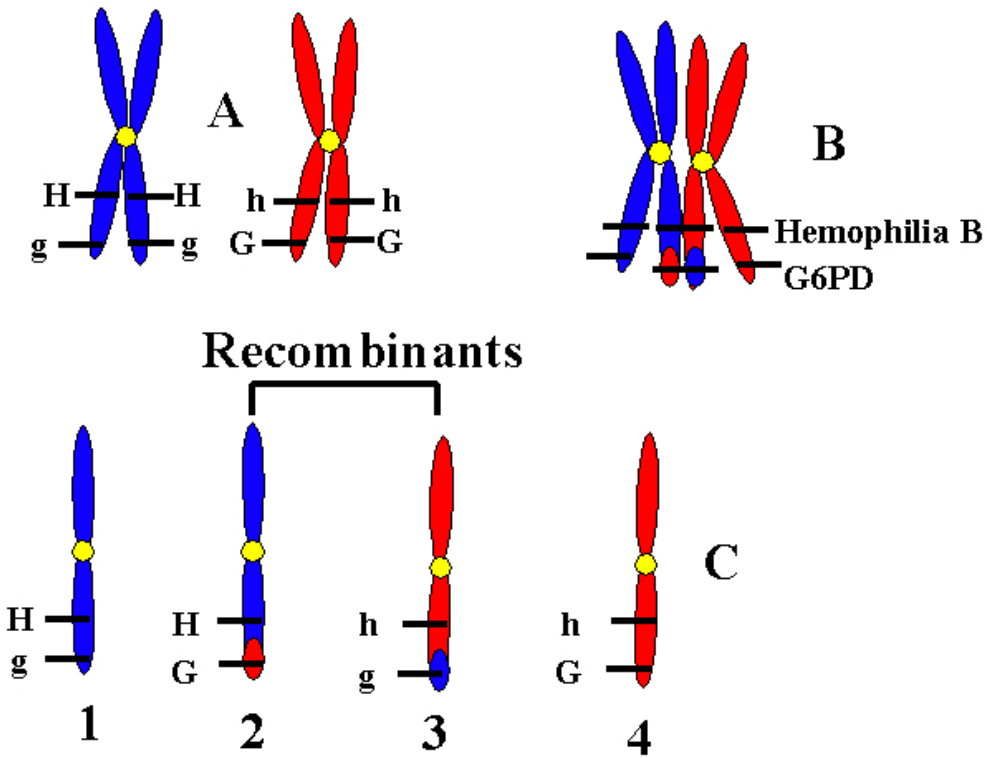



Crossing-over and recombination during meiosis

Haplotype: a set of DNA variations, or polymorphisms, that tend to be inherited together.

# Linkage

- 2 genes close together on the same chromosome pair do not assort independently at meiosis.
- Recombination frequency is the frequency that you will observe recombinant DNA among all gametes.
- A recombination frequency much less than 50% between 2 genes shows that they are linked.



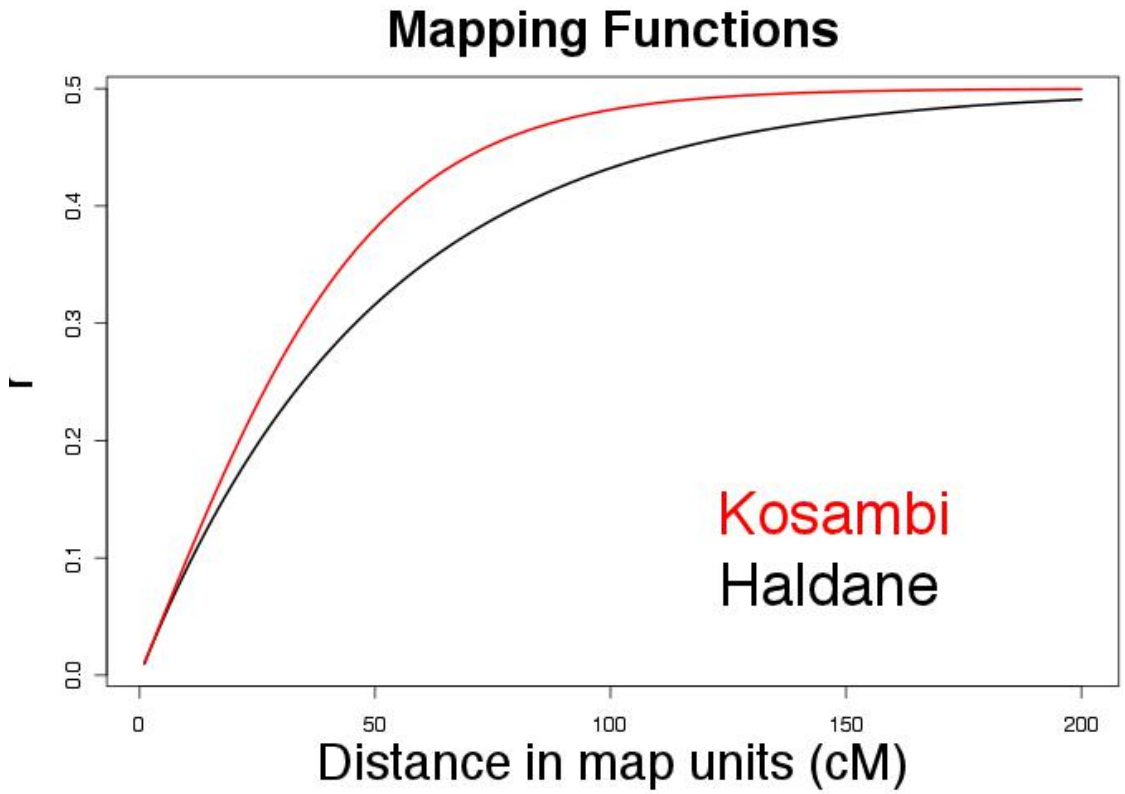
# Recombination Fraction

The recombination fraction (r) between two loci is the probability that a recombination occurs between the two loci.
**In human,**

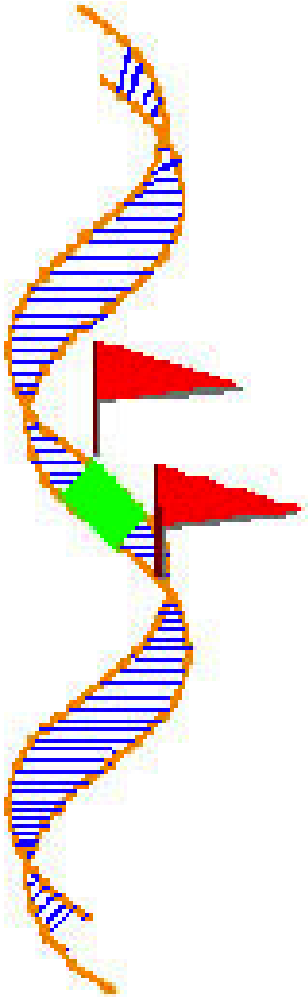
Kosambi
$r = 1/2 \times \frac{e^{d/25}-1}{1+e^{d/25}}$

Haldane
$r = 1/2 \times (1 - e^{-d/50})$

d: map units (cM)
1cM $= 1\% \approx 10^6$ base pairs.



# Genetic Markers



A genetic marker is a DNA sequence with a known physical location on a chromosome.

# Types of Genetic Variations

- 99% of DNA is shared between two individuals
- Variation in the remainder explains all our predisposition differences
- Remaining phenotypic variation: environmental/stochastic differences

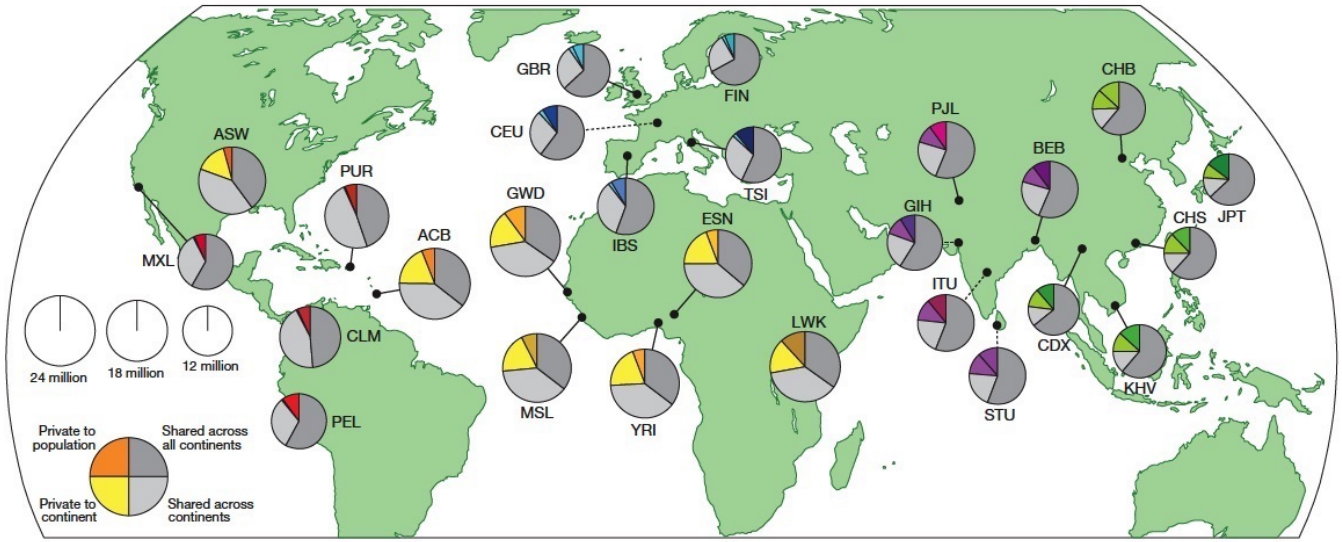
| Name | Example | Frequency |
|---|---|---|
| SNPs | GAGAACG[C/G]AACTCCG | 1 per 1,000 bp |
| Insertions / deletions | TATTC[C/CTATGG]TGTCT | 1 per 10,000 bp |
| Short tandem repeats (STRs) | ACGGCAGTCGTCGTCGTCACCGTAT | 1 per 10,000 bp |
| Structural variants (SVs), Copy Number Variants (CNVs) | Large (median 5,000 bp) deletions, duplications, inversions | 1 per 1,000,000 bp |

# Variant alleles: Distinguishing the two alleles in a SNP marker

- Matching the human reference sequence (reference/alternate)
- Being more frequent in the population (major/minor)
- Based on their disease association (risk/non-risk)

Example: rs189107123
GAGGAGAACG[C/G]AACTCCGCCG
Reference allele: C
Minor allele: G (frequency 0.03 in Europeans)

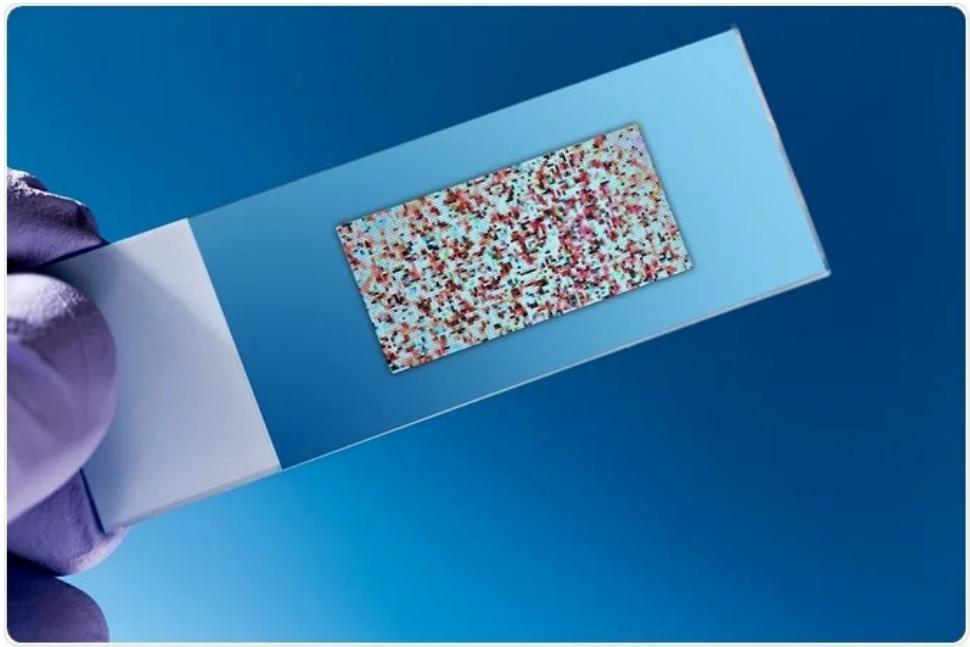
# A global reference for human genetic variation : The 1000 Genome Project




- 2,504 whole genome sequences across 26 subpopulations spanning the globe.
- The area of each pie is proportional to the number of polymorphisms within a population.
- The four slices in a pie representing whether the variants are shared/private across continents and subpopulation groups.

Thousand Genomes Consortium *Nature*, 2016

# Measuring known genetic variation: genotyping

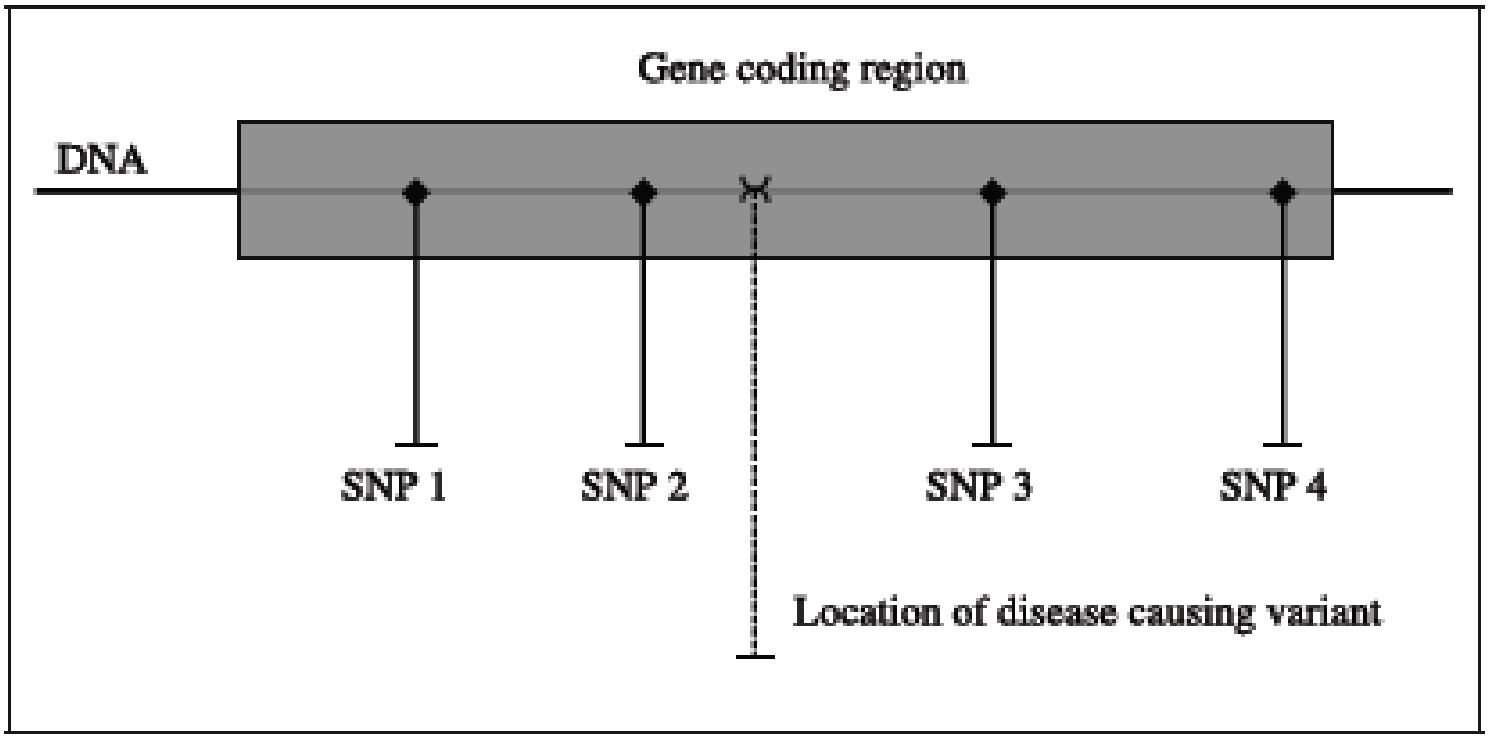


- Most genetic variants in an individual are recurrent in the population. Once they've been discovered/catalogued, a **common array** can be built for measuring them
- DNA microarrays were the key technological advance of the 1990s.
- We will cover how to analyze these image data generated from microarray later in this course.

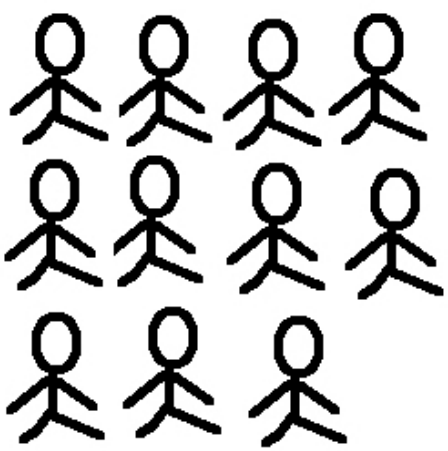
Image credit: sciencephoto/ Shutterstock.com

# Gene Association Analysis

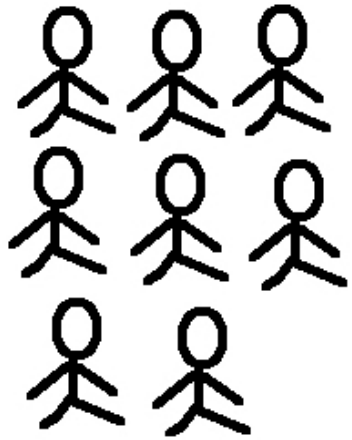
## SNP markers



# Gene Association Analysis



552 Type I diabetes cases



395 non-Type I diabetes controls

| | | |
|---|---|---|
| Frequency of a specific allele on a genetic marker | 10% | 7% |

We can compare the frequency of a specific allele on a genetic marker between participants in the case and control group and report a p value (Lecture 6).

# Genome-Wide Association Analysis (GWAS)

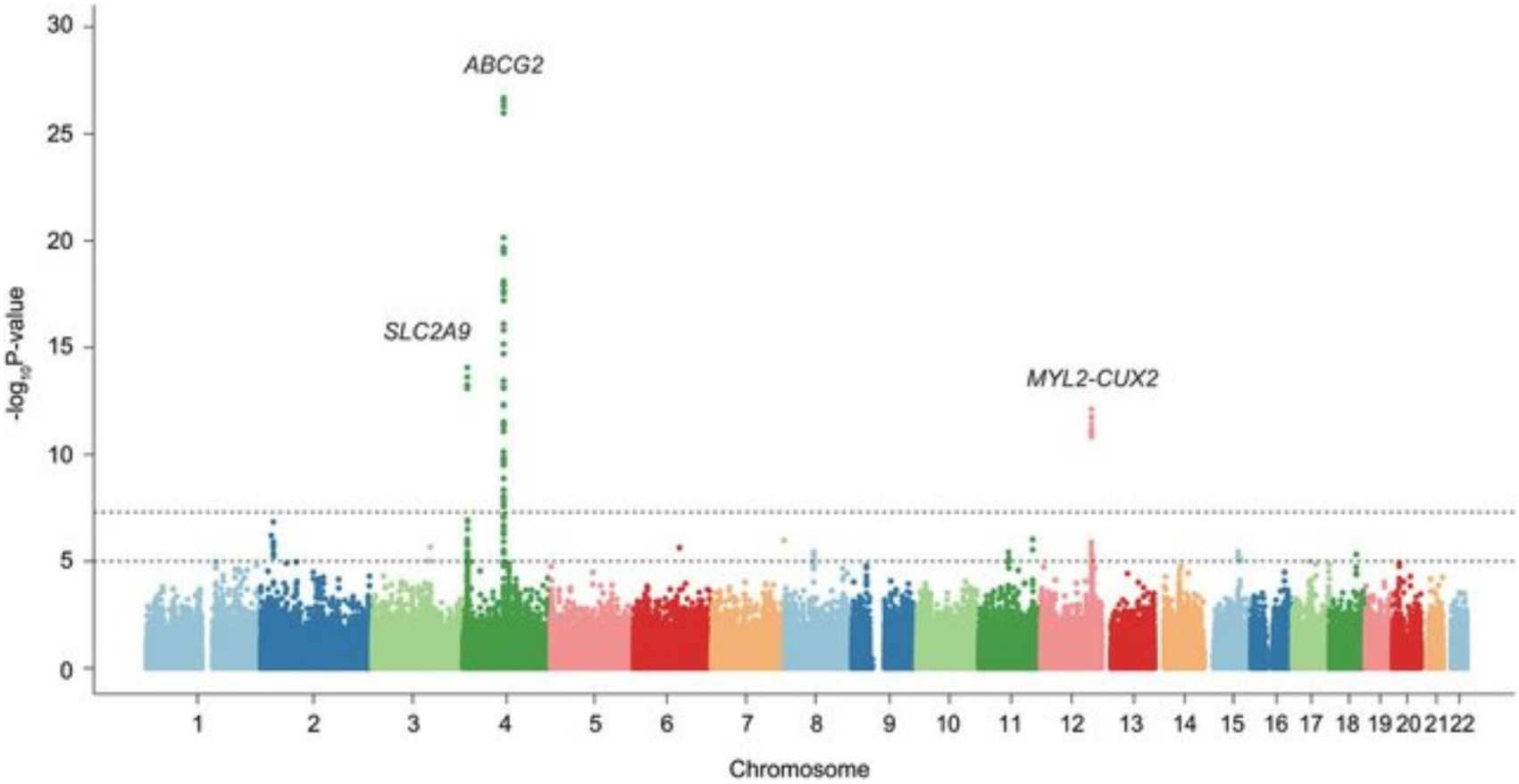
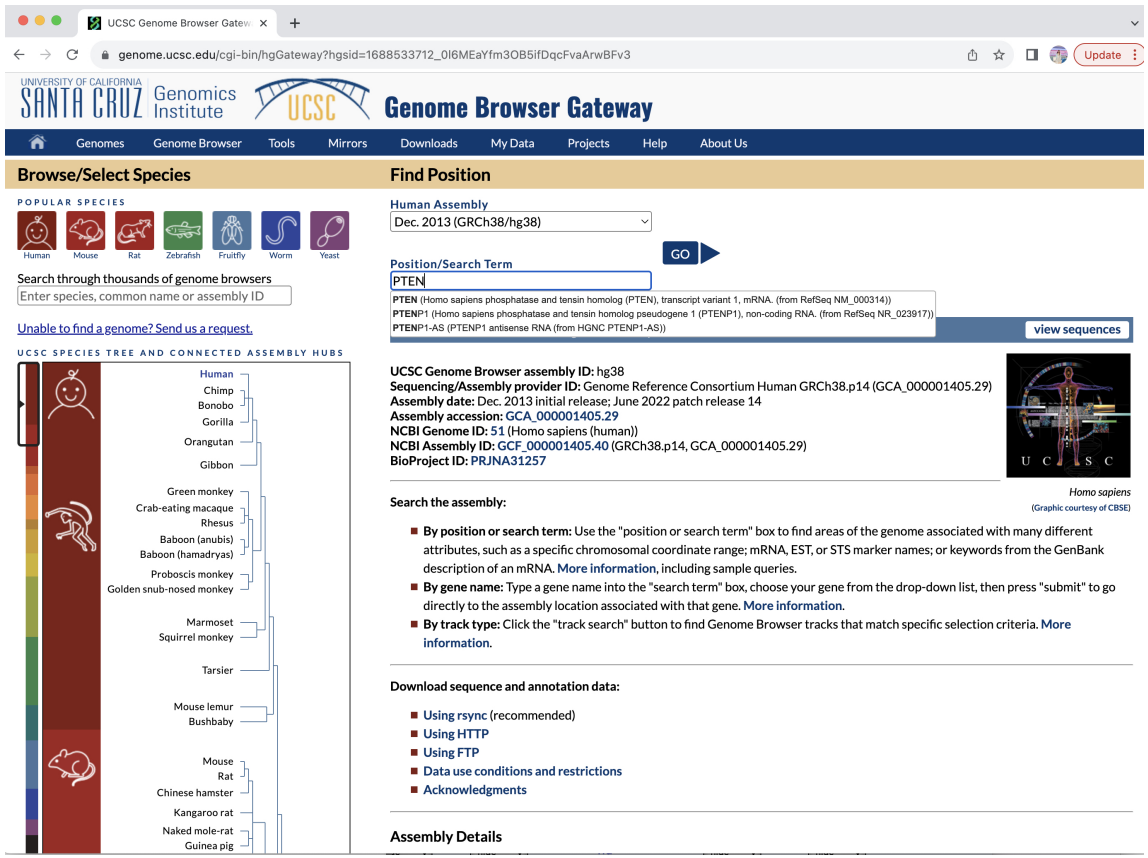


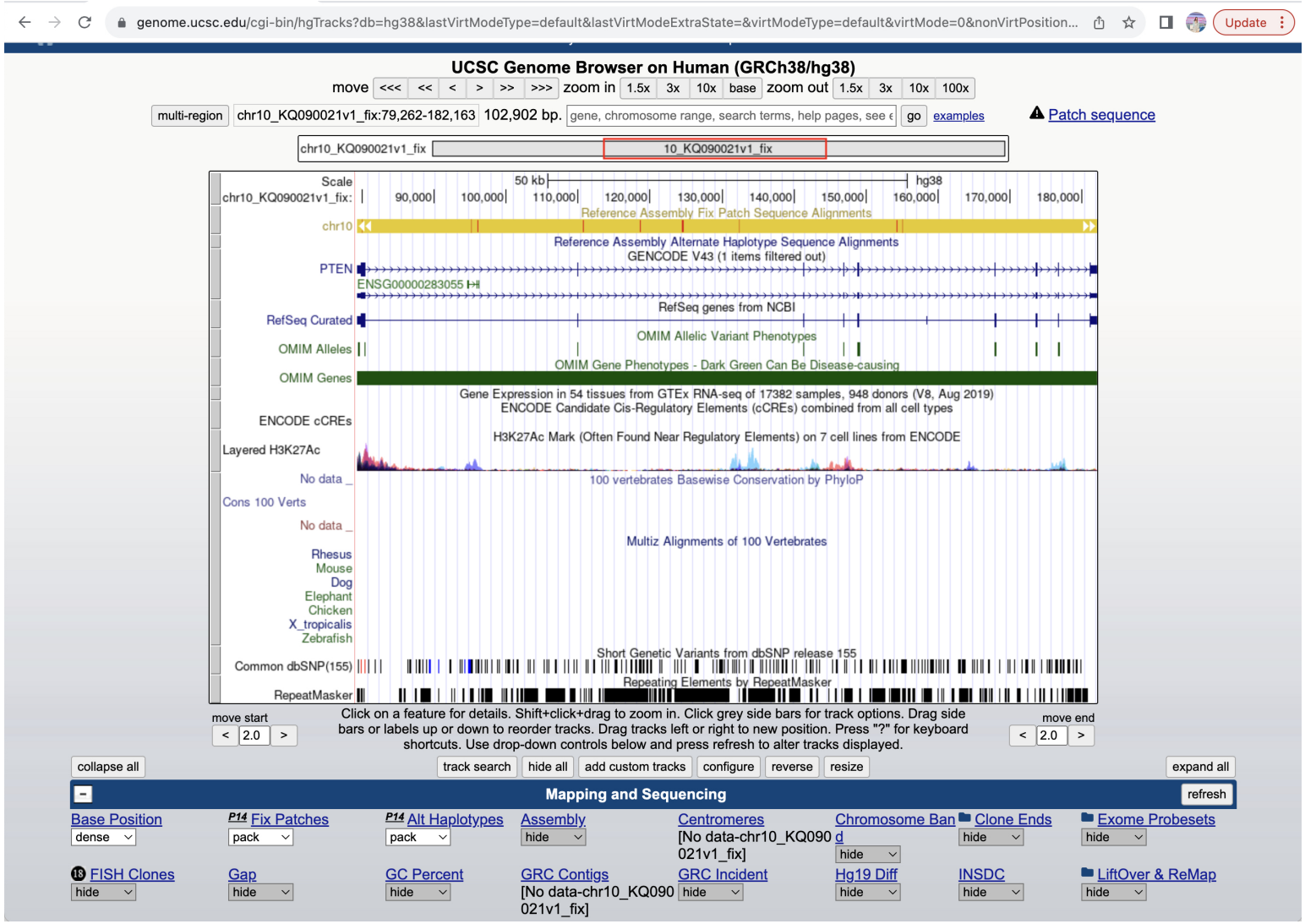

Figure: GWAS analysis of gout

# Online resources: genome browser and public data repositories

- UCSC genome browser: host genomic annotation data for many species.



# Online resources: genome browser and public data repositories



# Public high-throughput data repositories

- GEO: Gene expression omnibus.
  - Funded by NCBI
  - Host array- and sequencing-based data.
- ArrayExpression: European version of GEO
  - Better curated than GEO but has less data.
- SRA: sequence read archive.
  - Designed for hosting large scale high-throughput sequencing data (high speed file transfer).

# Other public data resources

- TCGA (The Cancer Genome Atlas)
  - Host data generated by TCGA, a big consortium to study cancer genomics.
  - Huge collection of cancer related data: different types of genomic, genetic and clinical data for many different types of cancers.
- ICGC (International Cancer Genome Consortium): Similar to TCGA but have a larger collection of studies.
- ENCODE (the ENCyclopedia Of DNA Elements) data coordination center
  - Host data generated by ENCODE, a big consortium to study functional elements of human genome.
  - Rich collection of genomic and epigenomic data.
- Many others ...