

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

Guest Lecture: Dr. Shannon Davis  
Department of Biological Sciences

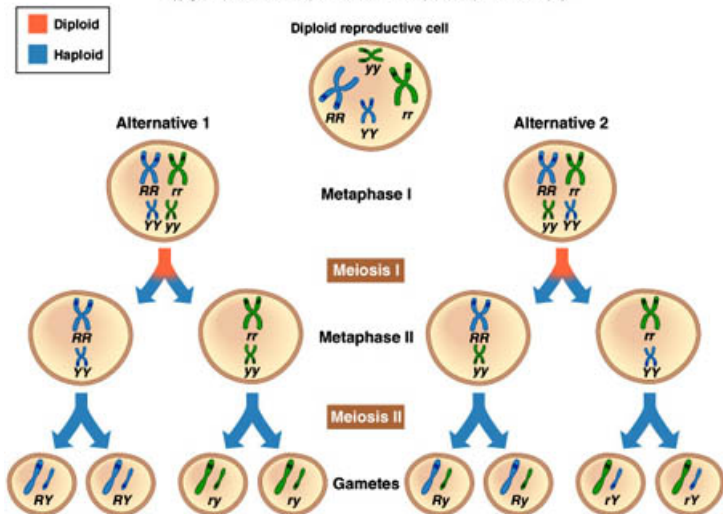
## Objectives of Lecture 4

Review basic terminology of population genetics

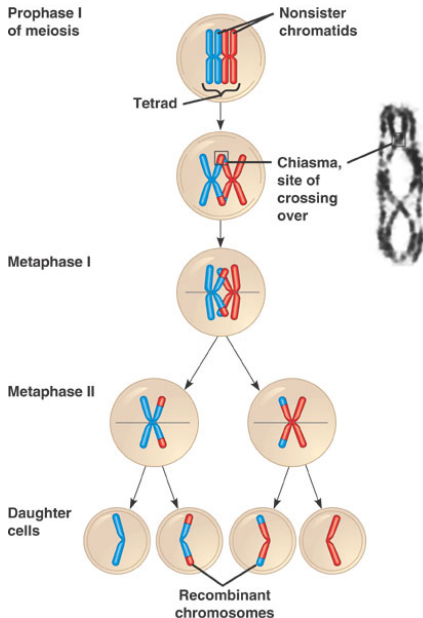
- ▶ Crossing Over
- ▶ DNA Recombination
- ▶ Genetic Markers
- ▶ Genetic Association Analysis
- ▶ Online Resources

# Random Combinations of Gametes

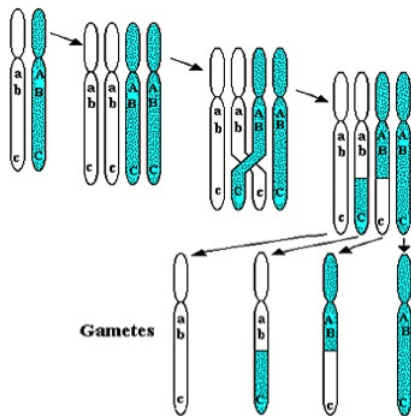
Copyright © The McGraw-Hill Companies, Inc. Permission required for reproduction or display.



# Crossing Over



## DNA Recombination

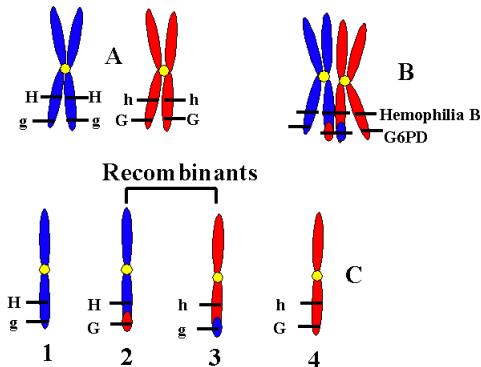


**Crossing-over and recombination during meiosis**

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**

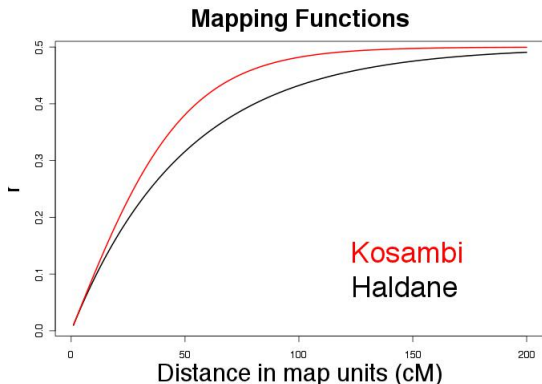
$$r = 1/2 \times \frac{e^{d/25} - 1}{1 + e^{d/25}}$$

**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)	ACGGCAGT <b>CGTCGT</b> CGTCACCGTAT	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)

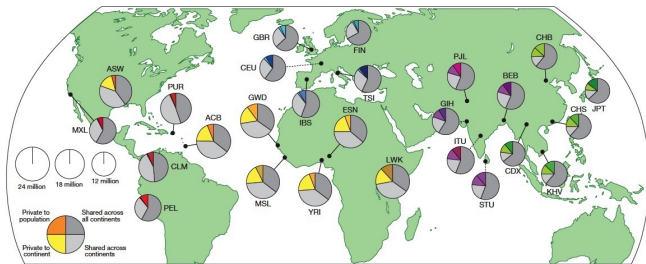
Example: rs189107123

GAGGAGAACG[C/G]AACTCCGCCG

Reference allele: C

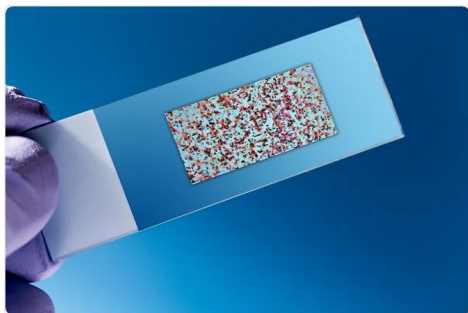
Minor allele: G (frequency 0.03 in Europeans)

# 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.

## Measuring known genetic variation: genotyping

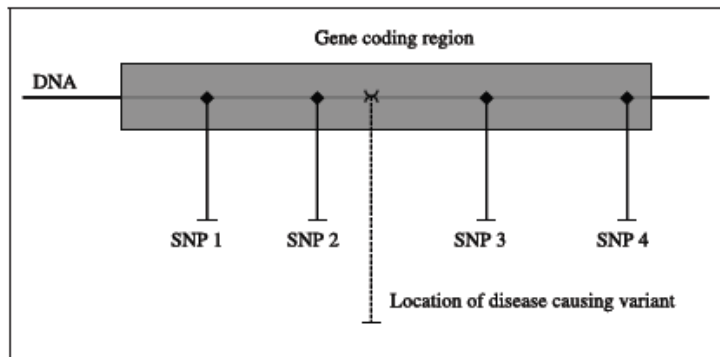


- Most genetic variants in an individual are recurrent in the population. Once they've been discovered/catalogued, a **common array** can be built for measuring them
- DNA microarrays were the key technological advance of the 1990s.
- We will cover how to analyze these image data generated from microarray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

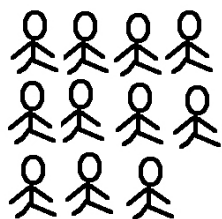
Image credit: sciencephoto/ Shutterstock.com

# Gene Association Analysis

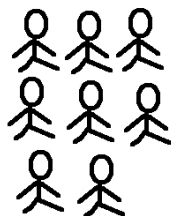
## SNP markers



## 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)

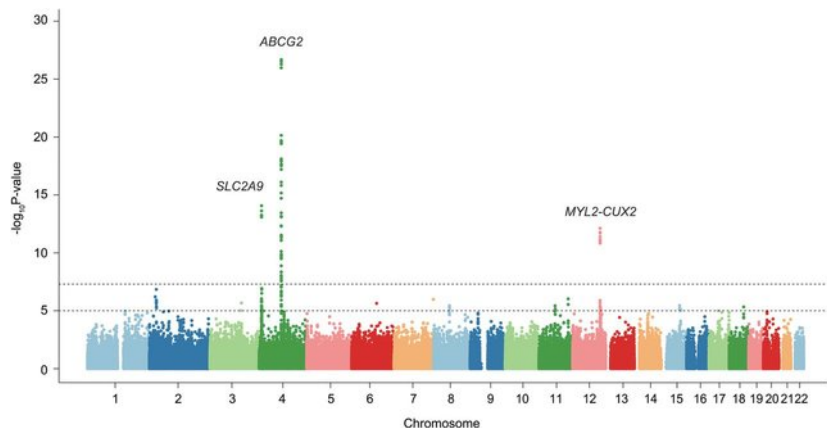


Figure: GWAS analysis of gout

# Online resources: genome browser and public data repositories

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

UCSC Genome Browser Gateway

Genomes Genome Browser Tools Mirrors Downloads My Data Projects Help About Us

**Browse/Select Species** **Find Position**

POPULAR SPECIES

Search through thousands of genome browsers  
Enter species, common name or assembly ID

Position/Search Term  
**PTEN**

Human Assembly  
Dec. 2013 (GRCh38/hg38)

PTEN Homo sapiens (phosphatase and tensin homolog) [PTEN], transcript variant 1, mRNA, show RefSeq, NC\_000314.6  
PTENP1 Homo sapiens (phosphatase and tensin homolog) pseudogene 1 (PTENP1), non-coding RNA, show RefSeq, NC\_029917.1  
PTENP1-AS1 (PTENP1) antisense RNA (show RefSeq, PTENP1-AS1)

UCSC Genome Browser assembly ID: hg38  
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38 p14 (GCA\_000001405.29)  
Assembly date: Dec. 2013 initial release; June 2022 patch release 14  
Assembly accession: GCA\_000001405.29  
NCBI Genome ID: 51 (Homo sapiens (human))  
NCBI Assembly ID: GCF\_000001405.40 (GRCh38.p14.GCA\_000001405.29)  
BioProject ID: PRJNA31237

Search the assembly:

- By position or search term: Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. More information, including sample queries.
- By gene name: Type a gene name into the "search term" box, choose your gene from the drop-down list, then press "submit" to go directly to the assembly location associated with that gene. More information.
- By track type: Click the "track search" button to find Genome Browser tracks that match specific selection criteria. More information.

Download sequence and annotation data:

- Using rsync (recommended)
- Using HTTP
- Using FTP
- Data use conditions and restrictions
- Acknowledgments

Assembly Details



# Online resources: genome browser and public data repositories

UCSC Genome Browser on Human (GRCh38/hg38)

multi-region chr10\_KQ090021v1\_fix:79,262-182,163 102,902 bp. [gene, chromosome range, search terms, help pages, see [go](#) [examine](#) [Patch sequence](#) ▲]

chr10\_KQ090021v1\_fix

Scale 50 kb hg38

chr10\_KQ090021v1\_fix: 90,000 100,000 110,000 120,000 130,000 140,000 150,000 160,000 170,000 180,000

Reference Assembly Fix Patch Sequence Alignments

Reference Assembly Alternate Haplotype Sequence Alignments  
GENCODE V43 (1 items filtered out)

PTEN  
ENSG00000283055 H4

RefSeq genes from NCBI

OMIM Alleles  
OMIM Gene Phenotypes - Dark Green Can Be Disease-causing

Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)  
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types

ENCODE cCREs

Layered H3K27Ac  
H3K27Ac Mark (Often Found Near Regulatory Elements) on 7 cell lines from ENCODE

Cons 100 Verts  
100 vertebrates Basewise Conservation by PhyloP

Multiz Alignments of 100 Vertebrates  
Rhesus  
Mouse  
Dog  
Elephant  
Chicken  
X\_tropicalis  
Zebrafish

Short Genetic Variants from dbSNP release 155  
Repeating Elements by RepeatMasker

RepeatMasker

move start < 2.0 > move end < 2.0 >

track search hide all add custom tracks configure reverse resize expand all refresh

Mapping and Sequencing

Base Position dense	Fix Patches pack	All Haplotypes pack	Centromeres hide	Chromosome Bands hide	Clone Ends hide	Exome Probesets hide
FISH Clones hide	Gap hide	GC Percent hide	GRC Contigs hide	GRC Incident hide	Hg19 Diff hide	LiftOver & RelMap hide

chr10\_KQ090021v1\_fix

## 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 ...