Statistics for Human Genetics and Molecular Biology
Lecture 2: Introduction to Population Genetics

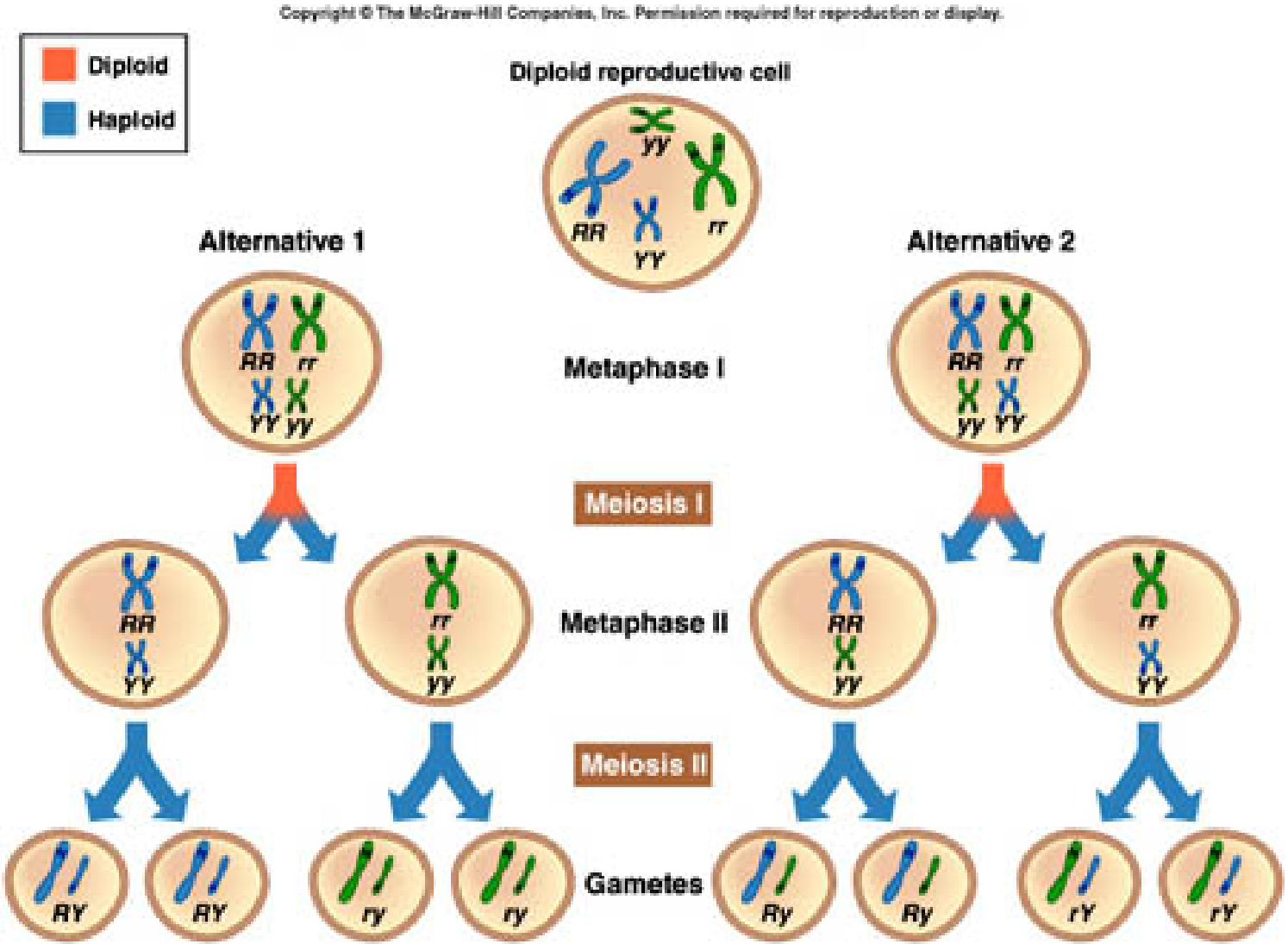
Dr. Yen-Yi Ho (yho@umn.edu)

Sep 11, 2015

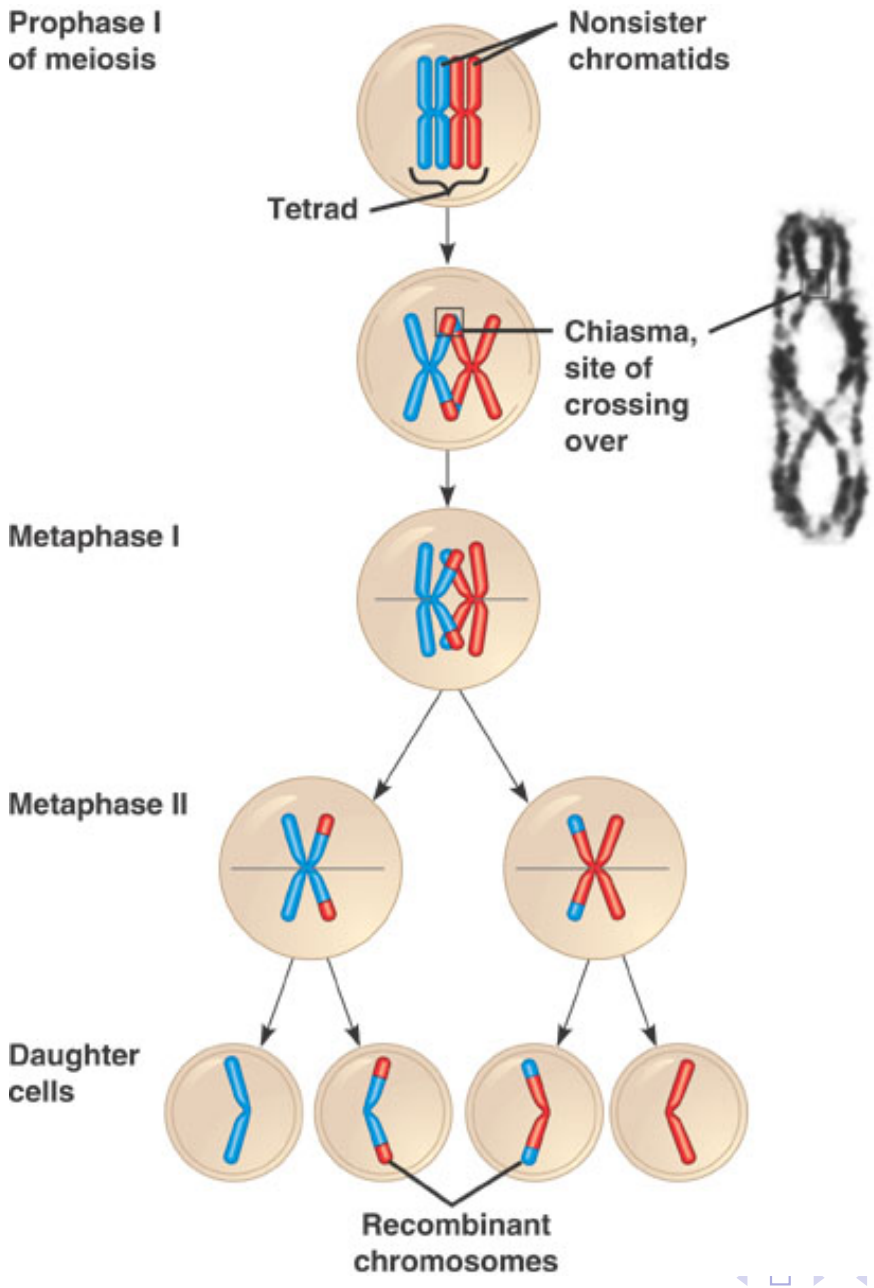
Objectives of Lecture 2

- Review basic terminology of population genetics
  - Crossing Over
  - DNA Recombination
  - Genetic Markers
  - Genetic Association Analysis
- Genetic Data
- Introduction to R

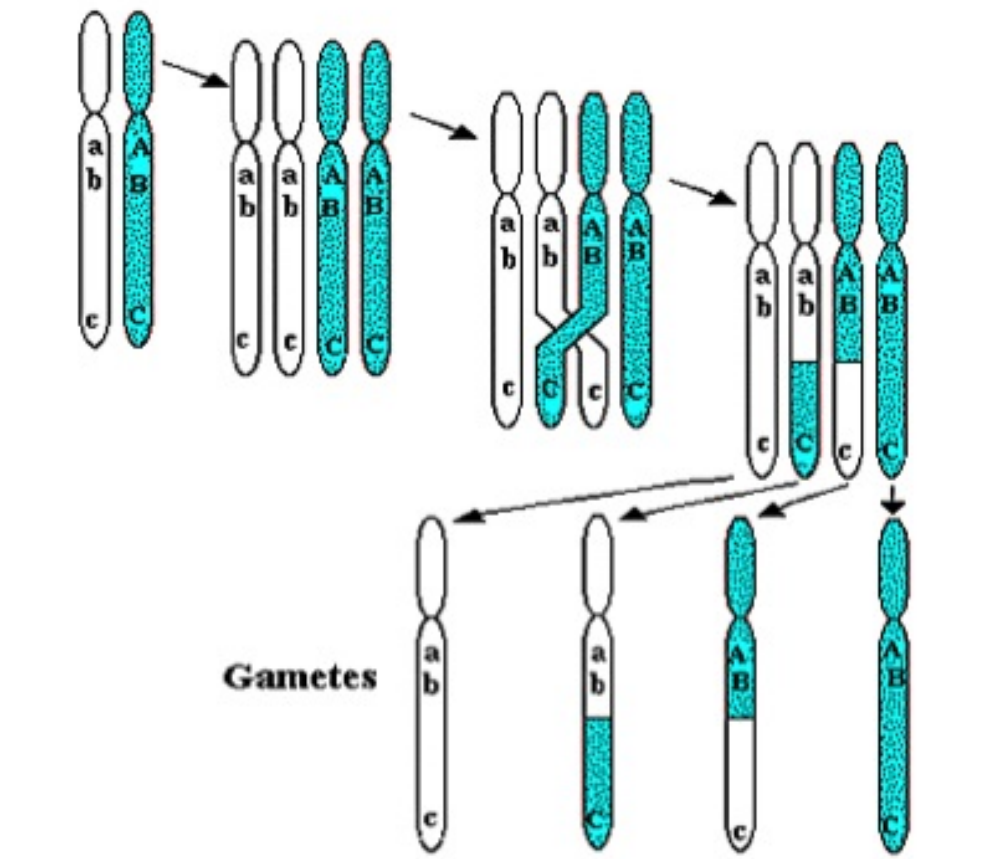
Random Combinations of Gametes



Crossing Over



DNA Recombination

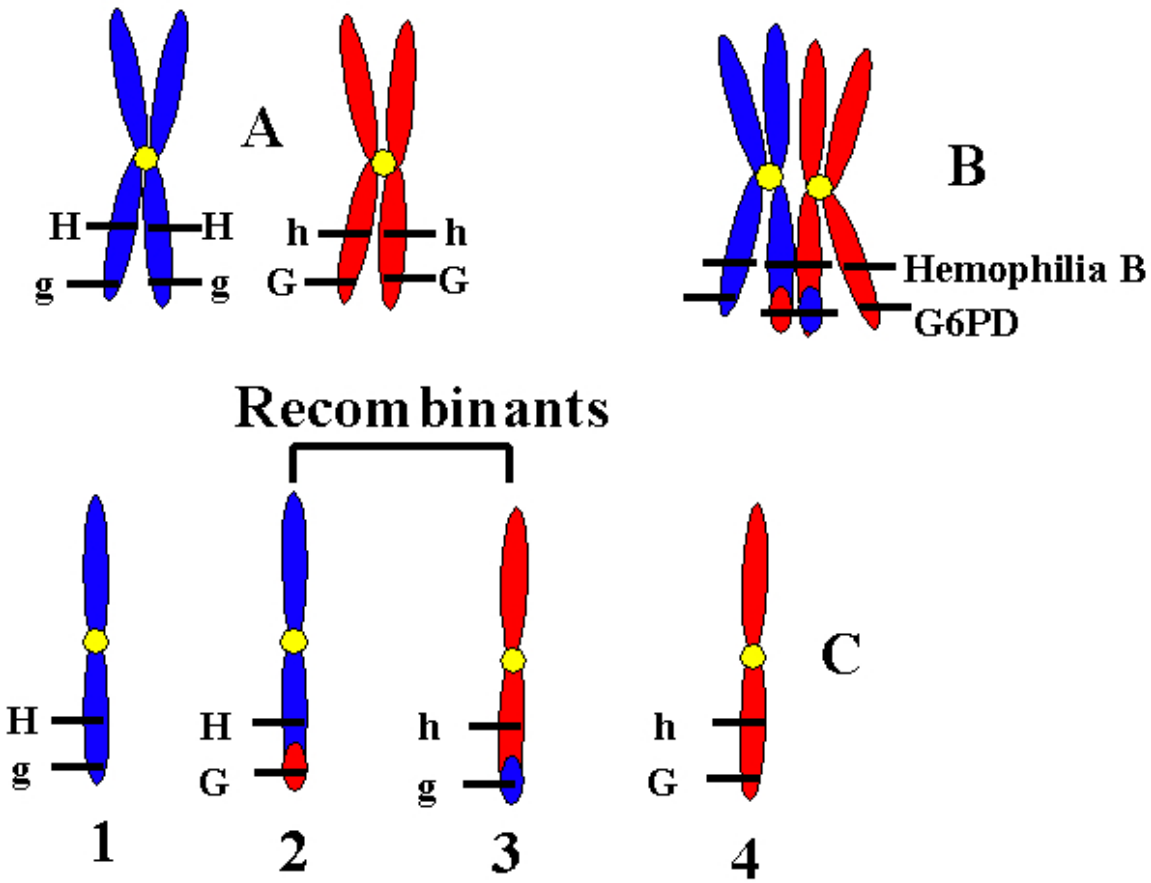


Crossing-over and recombination during meiosis

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.
- 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.

Kosambi

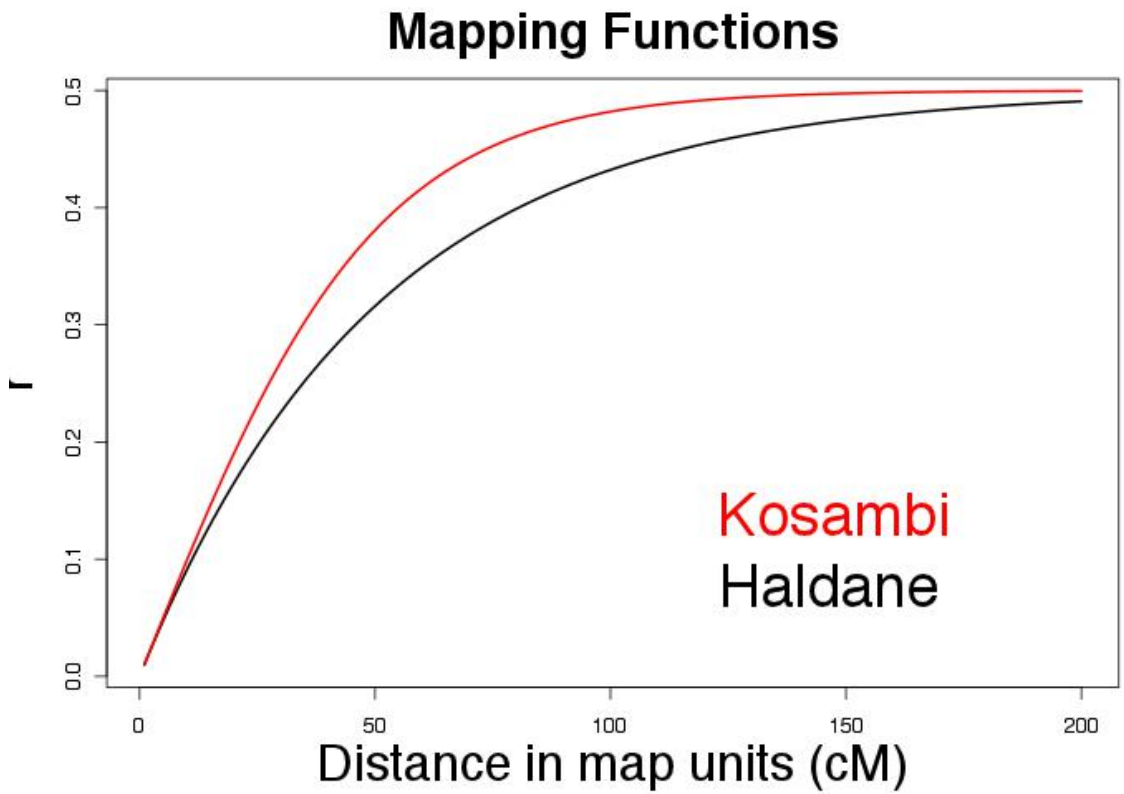
$r = 1/2 \times \frac{e^{d/25}-1}{1+e^{d/25}}$

Haldane

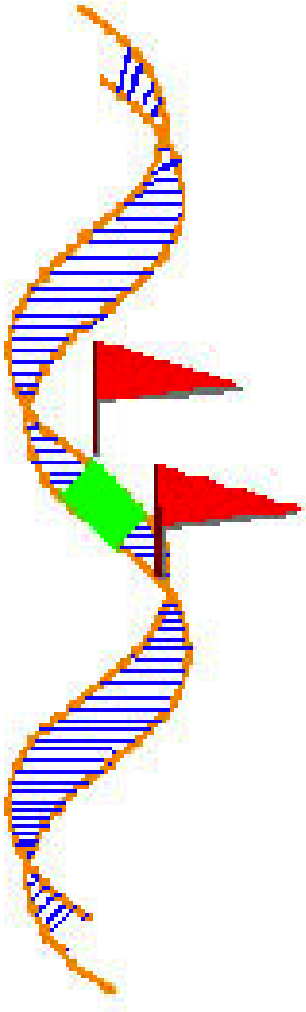
$r = 1/2 \times (1 - e^{-d/50})$

d: map units (cM)

1cM $= 1\% \approx 10^6$ base pairs.



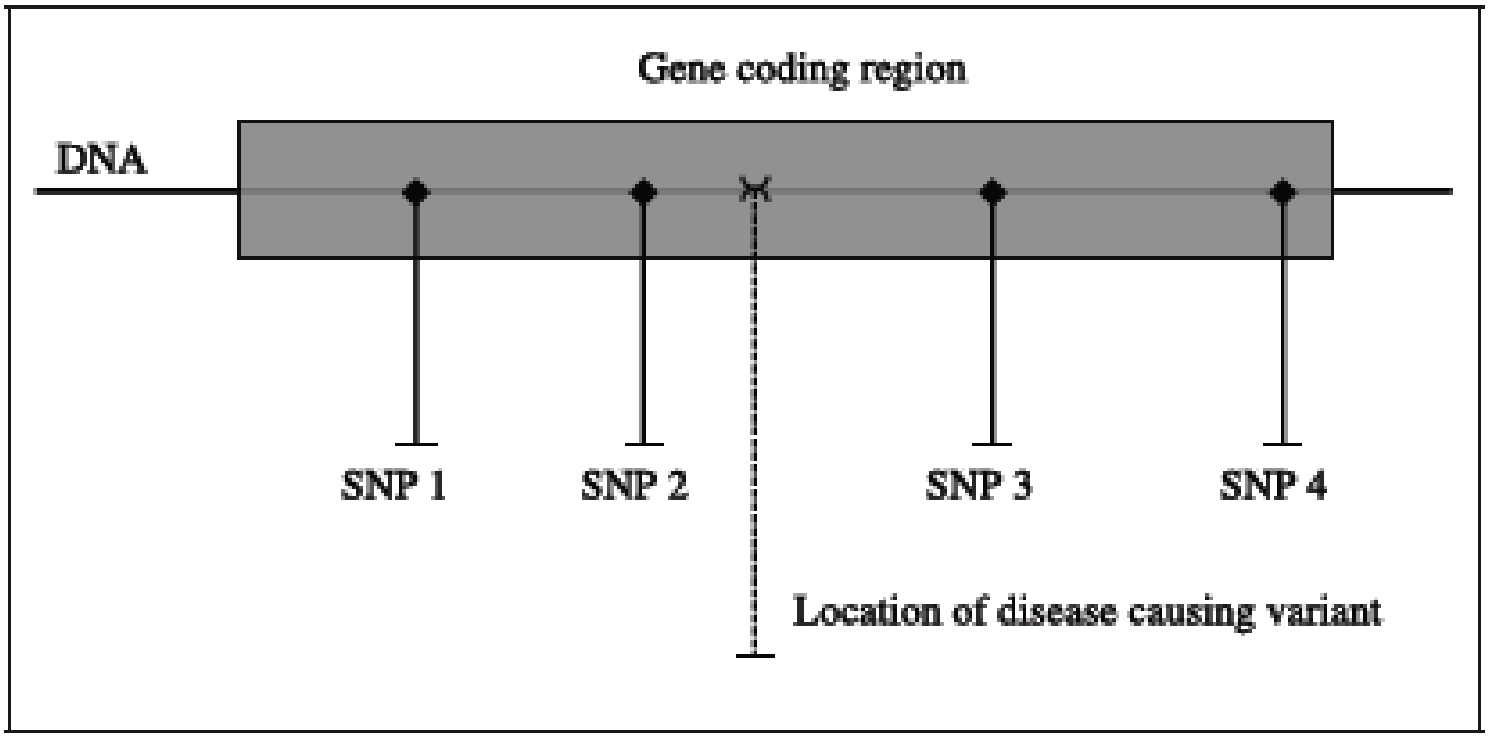
Genetic Markers



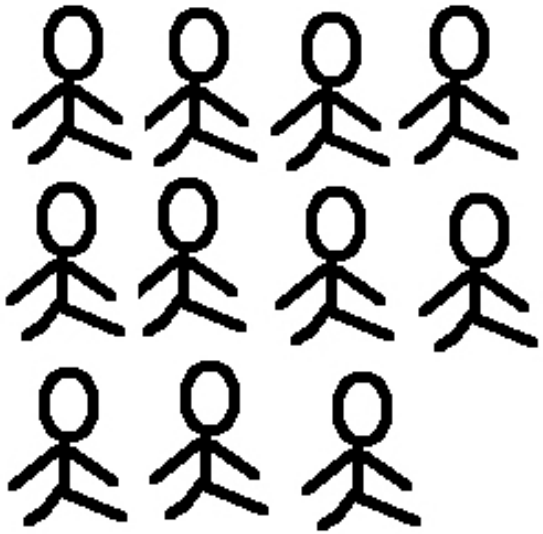
A genetic marker is a DNA sequence with a known physical location on a chromosome.

Gene Association Analysis

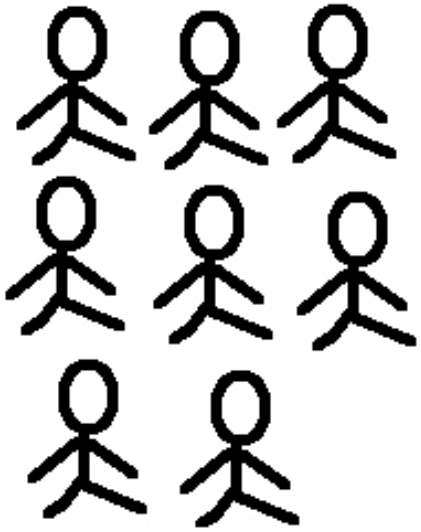
SNP markers



Gene Association Analysis



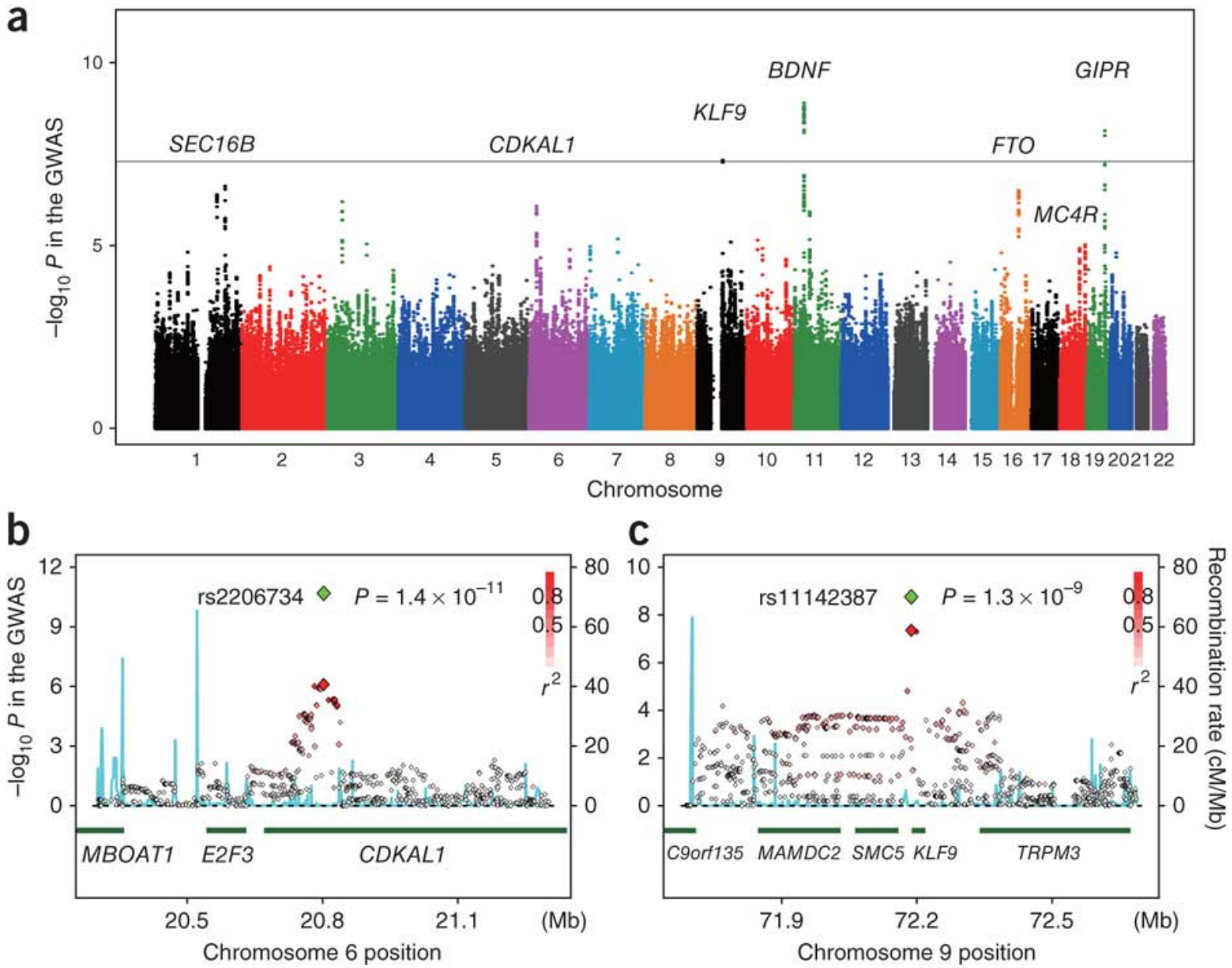
552 Type I diabetes cases



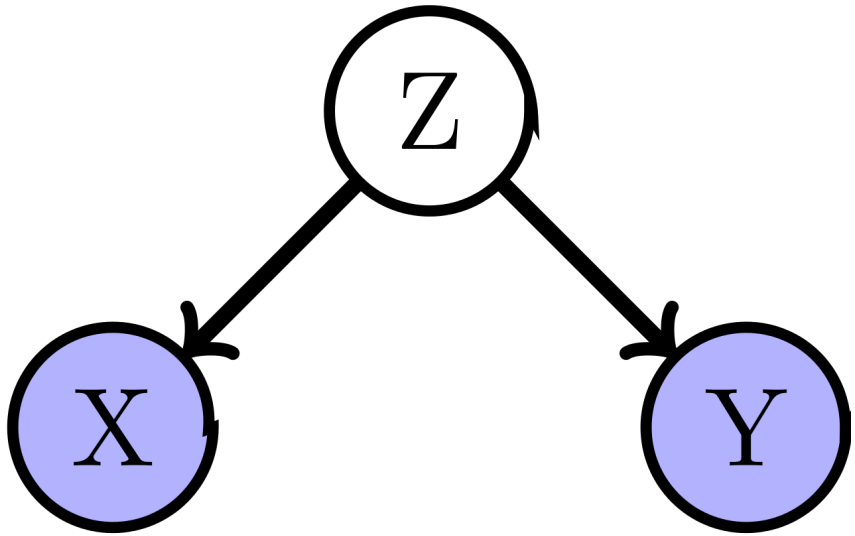
395 non-Type I diabetes controls

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

Genome-Wide Association Analysis (GWAS)



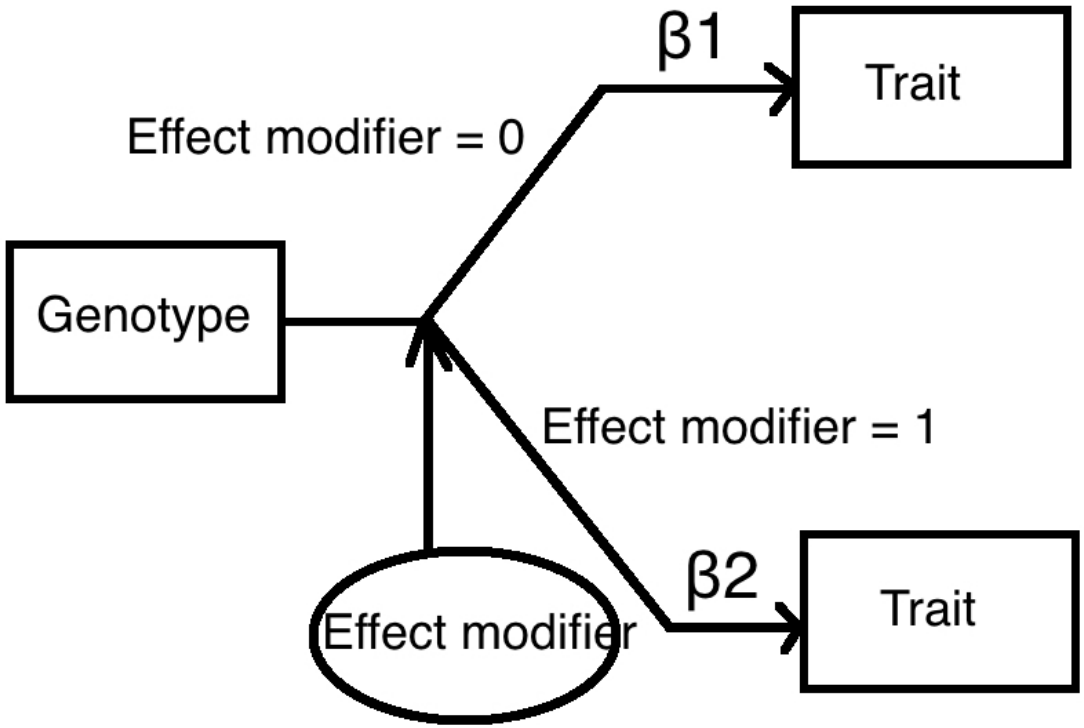
Confounding



- Associated with X
- Independently associated with Y
- Not in the causal pathway between X and Y
- ex: ice cream consumption and drowning death

Effect Modification

Effect of a predictor variable on the outcome depends on the level of another variable (interaction).



Genetic Data Structure

- Genetic information/Genotype
- Trait
- Individual-specific covariates

Genetic Data Used throughout this Course

FAMuSS Study: The Functional SNPs Associated with Muscle Size and Strength

- n=1397
- 225 SNP markers
- individual-specific covariates: gender, age, race,
- traits: muscle strength, BMI, ...

	id	acdc_rs1501299	actn3_r577x	actn3_rs540874
1	FA-1801	CA	CC	GG
2	FA-1802	CA	CT	GA
3	FA-1803	CA	CT	GA
4	FA-1804	CC	CT	GA
5	FA-1805	CA	CC	GG

The Human Genome Diversity Project (HGDP)

- n=1064
- 4 SNP markers on *AKT1* gene
- individual-specific covariates: gender, population, geographical area, ...

	ID	Gender	Population	Geographic.area	AKT1.C0756A
1	HGDP00980	F	Biaka Pygmies	Central Africa	CA
2	HGDP01406	M	Bantu	Central Africa	CA
3	HGDP01266	M	Mozabite	Northern Africa	AA
4	HGDP01006	F	Karitiana	South America	AA
5	HGDP01220	M	Daur	China	AA

The Virco Data

- n=1066 viral isolates
- 99 sequence information within the protease region of the viral genome
- drug-specific fold-resistance

	SeqID	APV.Fold	IDV.Fold	P10	P11	P12
1	3852	7.50	14.20	I	-	-
2	3865	3.00	13.50	I	-	-
3	7430	3.30	16.70	I	-	-
4	7459		3.00	I	-	-
5	7460		7.00	-	-	-

The ALL Dataset

- Microarrays data with 12,625 gene expression probes (features) from 128 individuals with acute lymphoblastic leukemia (ALL).
- individual specific covariates: gender, age, tumor type and stage, translocation mutation

	01005	01010	03002	04006	04007
1000_at	7.60	7.48	7.57	7.38	7.91
1001_at	5.05	4.93	4.80	4.92	4.84
1002_f_at	3.90	4.21	3.89	4.21	3.42
1003_s_at	5.90	6.17	5.86	6.12	5.69
1004_at	5.93	5.91	5.89	6.17	5.62

R Topics Outline

- Get Started
- R as a calculator
- Vectors
- Matrices, Arrays, Factors, List, Data Frame
- Import/Export Data
- R Graphics
- Random number generating
- Writing R function
- for loops
- rep, seq, which, match

R: Pros and Cons

Pros

+ Free
+ Available for all major platforms
+ Powerful graphics
+ Comprehensive
+ Well-designed programming language (object-oriented)
+ Unlimited extensibility
+ Widely used by statisticians
+ Increasingly used for genomic analyses (Bioconductor)

Cons

- No dedicated support
- Complex Syntax
- Not point-and-click
- No warranty

Get Started

- Installation
  - google R $\rightarrow$ The R project for Statistical Computing
  - R64 bits (large datasets) vs. R32 bits
- Getting help with R
  - At the command prompt, type, for example ?read.table or help(read.table)
  - At the command prompt, type, for example, help.search("read") or apropos("read").
  - Within R, use the menu bar: Help: R help.
  - Quititng R. q()
  - How to save work space

R Resources

John Verzani's SimpleR notes
R Reference Card
CRAN (Document/Manuals)

Note: To run some of the example in John Verzani's notes, run:
> install.packages("UsingR")
> library(UsingR)

R as a calculator

```
> 3 + 2
[1]5
> 7/2
[1]3.5
> 3 * 5
[1]15
> 2 ^ 3
[1]8
> 7%% 3 ## answer 1, modulo reduction
> log(1 : 4)
> log 2(1 : 4)
> log(1 : 4, base = 3)
> exp(1)
[1]2.718282
> abs(-3)
> sqrt(3)
> sin(0.5)
```

Vectors

Vectors contain elements of just 1 type, either numeric, logical, or character.

- Accessing elements in a vector: [ ]. (Very important)
- By logical conditions.

```
> x<- 3
> x # print x
>x<-4
>x
### Creating simple vectors
> x<- c(1,3.5,-28.4,10) #numerical vector
> y<-c("cat","dog","mouse","monkey") #character
> z<-c(TRUE,TRUE,TRUE,FALSE,FALSE) #logical vector
> x<-1:10
> seq(1, 10, by=1)
> seq(3,9, by=3)
> rep(2,10)
> log(seq(1,2,by=0.1))
>x <- c(1,5,10,NA,15)
> sum(x)
> sum(x,na.rm=TRUE)
> prod(x,na.rm=TRUE)
> mean(x,na.rm=TRUE)
```

Accessing Elements in a Vector

```
>y <- c(8,32,15,-7, 2,19)
> length(y)
>y[3:5] ##position in vector as positive integer
> y[-c(1,5,6)] ## exclude:  use negative integers
>y< 15
> y[y<15]
> which(y==32)
> x <- 1:10
> match(y, x)
> colors <- c(``red", ``blue", ``pink", ``yellow")
> which(colors == ``yellow")
> x < −c(1,5,10,NA,15)
> which(is.na(x))
> which(!is.na(x))
```

Factors

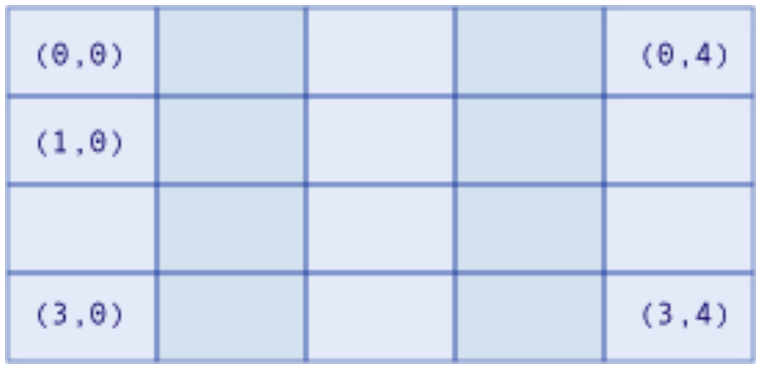
Factors: vectors with levels. Handy for regression modeling.

Example:

```
> colors <- c(1, 1, 2, 3)
> colors <-
factor(colors, label=c("red","green","blue"))
> table(colors)
colors
red green blue
2       1     1
```

Matrices

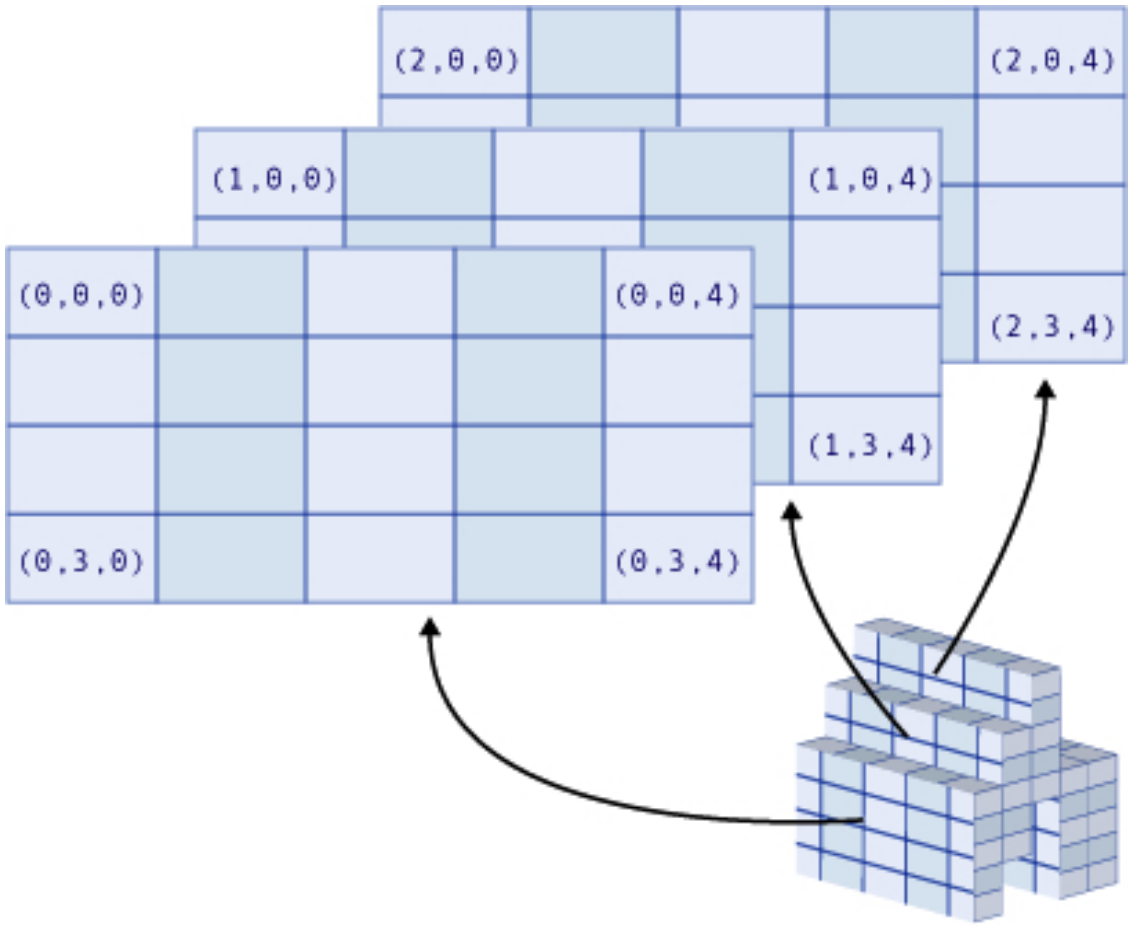
- Dimension: Row by column.
- Accessing elements in a matrix: [ ].



```
> help(cbind)
>y <- c(8,32,15,-7, 2,19)
> x <- 1:6
> mat<- cbind(x,y)
> help(rbind)
> dim(mat) ## check dimension
> ncol(mat) ## the number of columns of a matrix
> nrow(mat) ## the number of rows of a matrix
> mat[2,3] # the value in the 2nd row and the 3rd column
> mat[1:3,] ## the first three row of mat
> mat[,2] ## the 2nd column of mat
> mat[-1,] ## exclude the first row
> newmat<-matrix(1:9, nrow=3) ## create new matrix
> newmat
> m<-matrix(1:9, nrow=3, byrow=T) ## fill row first
> colnames(m) <- c("a", "b", "c") ## label column name
> rownames(m) <- c("r1", "r2", "r3")
> vect<-as.vector(newmat)
```

Arrays

- Dimension: Row by column by height.



```
> myarray<-array(1:64, dim=c(4,4,4))
> myarray
> myarray[1,2,3]
```

Data Frames

Data Frame: like matrices, but each column can be a different data type.

```
>str(mydata)
'data.frame': 10 obs. of 3 variables:
 $ y : num 24.2 26.6 23.9 23.6 23.6 ...
 $ x1: num 3.02 2.43 3.35 3.86 3.7
 $ x2: Factor w/ 2 levels "F","M": 2 2 2 2 2 2 1 1 1 1
```

Lists

List: a bag contains different things (vectors, matrices, data frames,)

- Accessing components in a list: [[ ]].
- Accessing to elements within components.

Lists

```
> x <- list(one=c(18:36),two=c("AK","AL","AZ"),
            three=c(T,T,F,T),four=matrix(1:12,3,4))
> str(x)
List of 4
$ one :  int [1:19] 18 19 20 21 22 23 24 25 26 27
$ two :  chr [1:3] "AK" "AL" "AZ"
$ three:  logi [1:4] TRUE TRUE FALSE TRUE
$ four :  int [1:3, 1:4] 1 2 3 4 5 6 7 8 9 10 ...
> x[[1]]
> x$one
> y <- unlist(x)
```

For loops

```
for(i in 1:100){
    d <- Sys.time()
    print(paste("Now is", d, sep=" "))
    print(i*i)
}
```

Next Lecture

- R
  - Get Started
  - R as a calculator
  - Vectors
  - Matrices, Arrays, Factors, List, Data Frame
  - Import/Export Data
  - for loops
  - R Graphics
  - Random number generating
  - Writing R function
  - rep, seq, which, match