Introduction to linkage and association

Dmitry Prokopenko

dprokopenko@mgh.harvard.edu

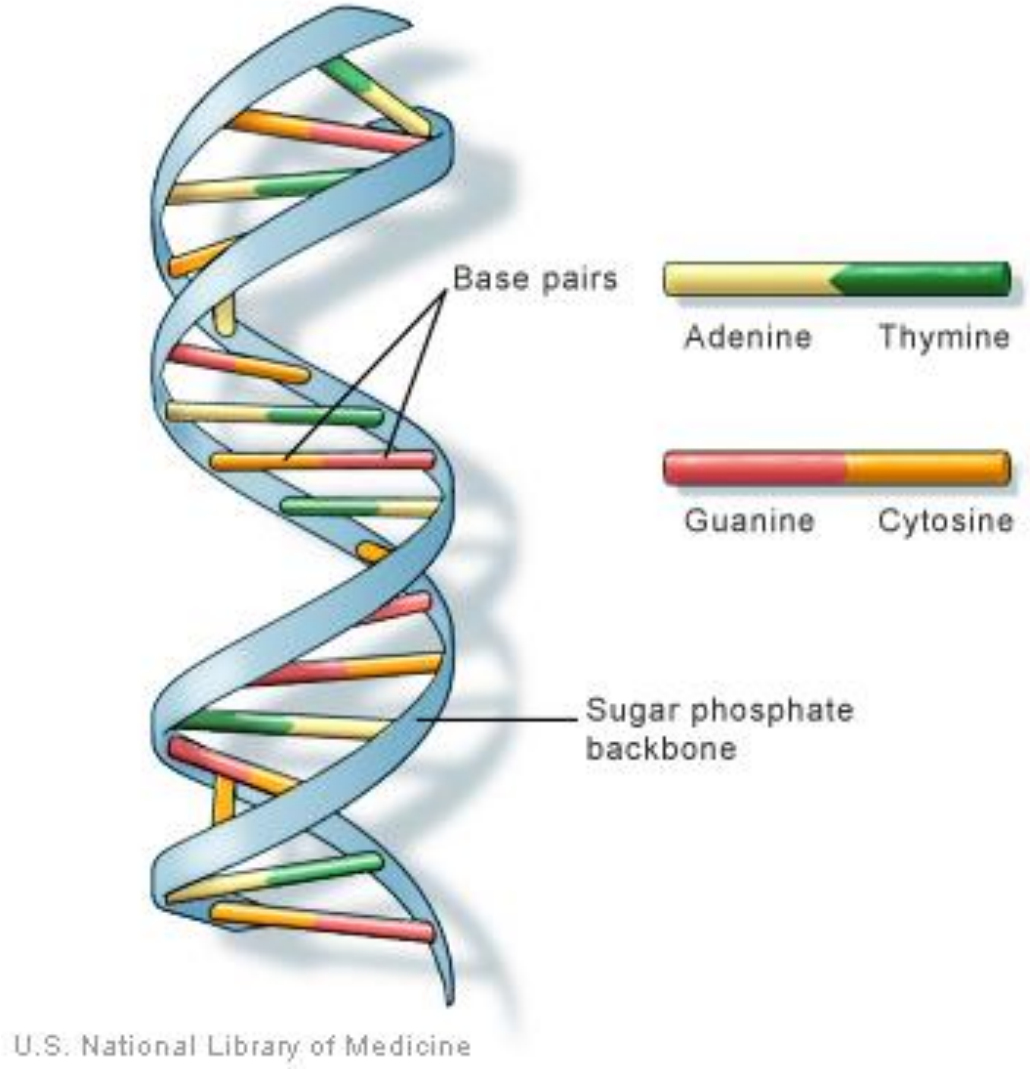
- Mon Apr 1, 2019 Introduction to linkage and association
- Wed Apr 3, 2019 Genome-Wide and Family-Based Association Studies

Contents

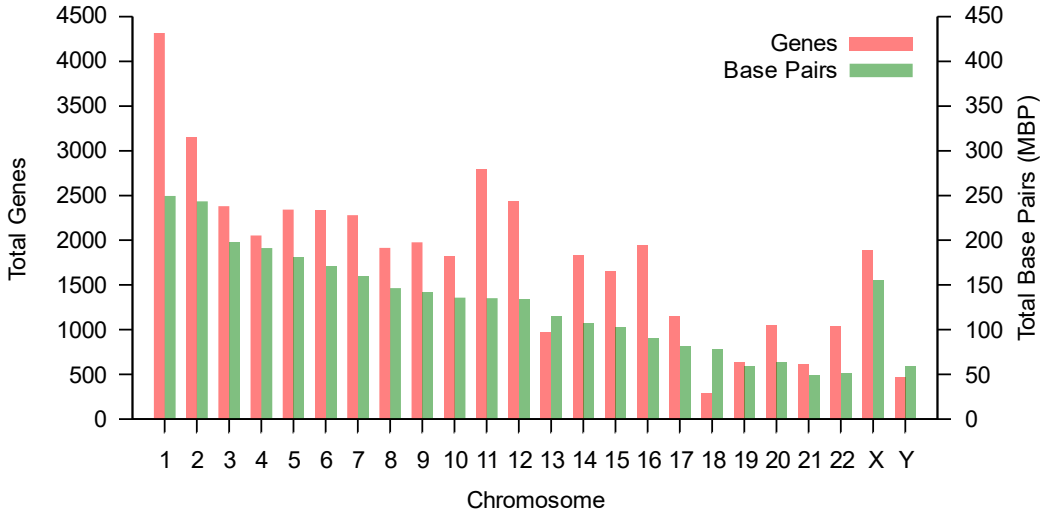
- Definitions
- Genetic data overview
- What is linkage
- Linkage methods
- What is association
- LD and other concepts
- Linkage versus association

DNA

- Deoxyribonucleic acid – double stranded helix molecule
- Organized into long structures – chromosomes
- Every chromosome carries 2 copies of information
- Length of unwrapped human DNA ~ 6ft (1.8 m) per cell



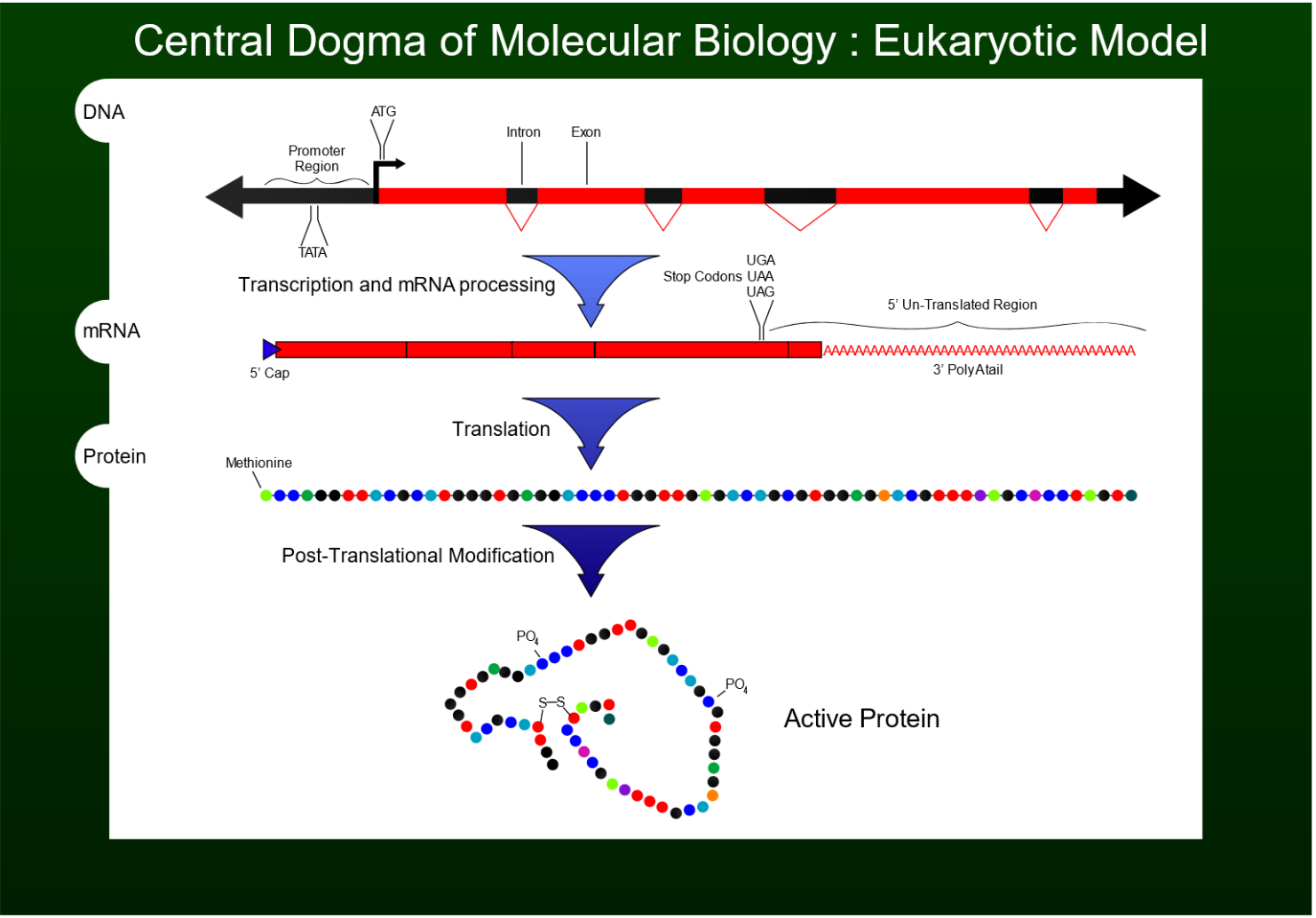
Human Genome



- 2 copies of 22 autosomes + 2 sex chromosomes
- ~3,200 Mb (Mega-basepairs) = 3,200,000,000 basepairs
- ~20,000 protein-coding genes

https://commons.wikimedia.org/w/index.php?curid=18594472

Central dogma of molecular biology



- DNA->RNA->Aminoacid->Protein
- Coding exons ~ 1 % of human genome
- 99.9% of human genomes – same
- ~ 12M mutations (SNPs) with a frequency > 1% in population

https://commons.wikimedia.org/w/index.php?curid=32026515

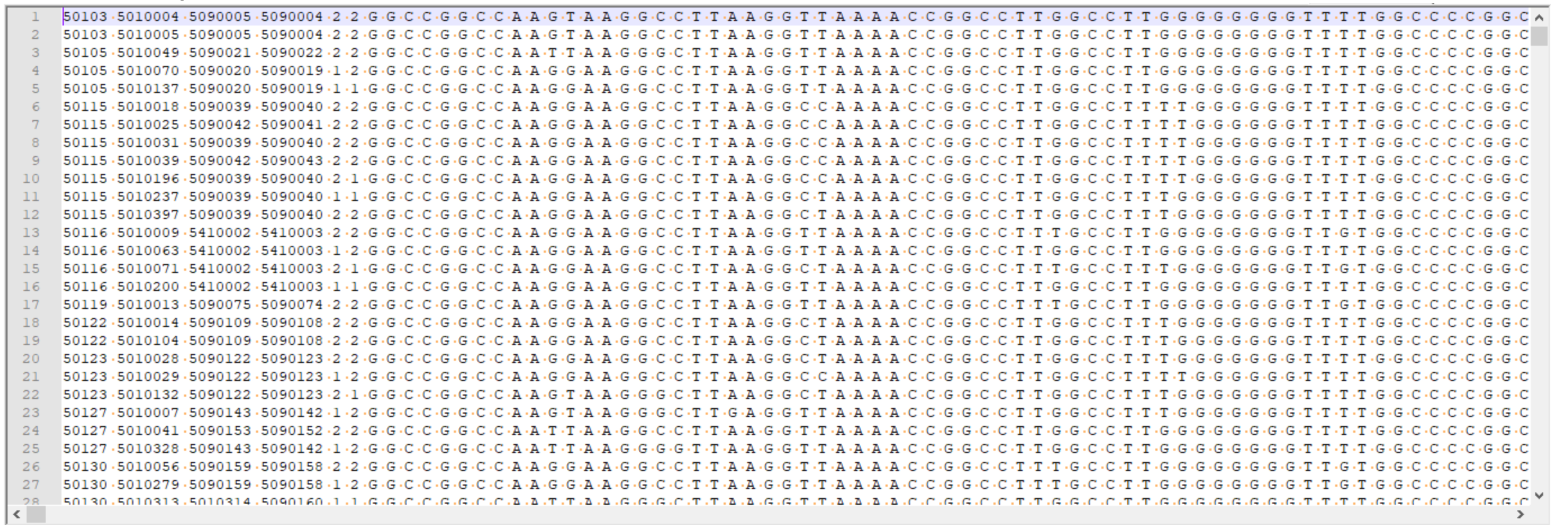
Genetic data representation

.fasta

.sam/.bam

```
@HD	VN:1.0	SO:coordinate
@SQ	SN:chr20	LN:64444167
@PG	ID:TopHat	VN:2.0.14	CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-realign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714	16	chr20	190930	3	100M	*	0	0	CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCTC	BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@	AS:i:-15	XM:i:3	XO:i:0	XG:i:0	MD:Z:55C20C13A9	NM:i:3	NH:i:2	CC:Z:=	CP:i:55352714	HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961	16	chr20	193953	50	100M	*	0	0	TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGAG	DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC	AS:i:-16	XM:i:3	XO:i:0	XG:i:0	MD:Z:60G16T18T3	NM:i:3	NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030	16	chr20	270877	50	100M	*	0	0	GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAAACCAC	DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB	AS:i:-11	XM:i:2	XO:i:0	XG:i:0	MD:Z:0A85G13	NM:i:2	NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699	0	chr20	271218	50	50M4700N50M	*	0	0	GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

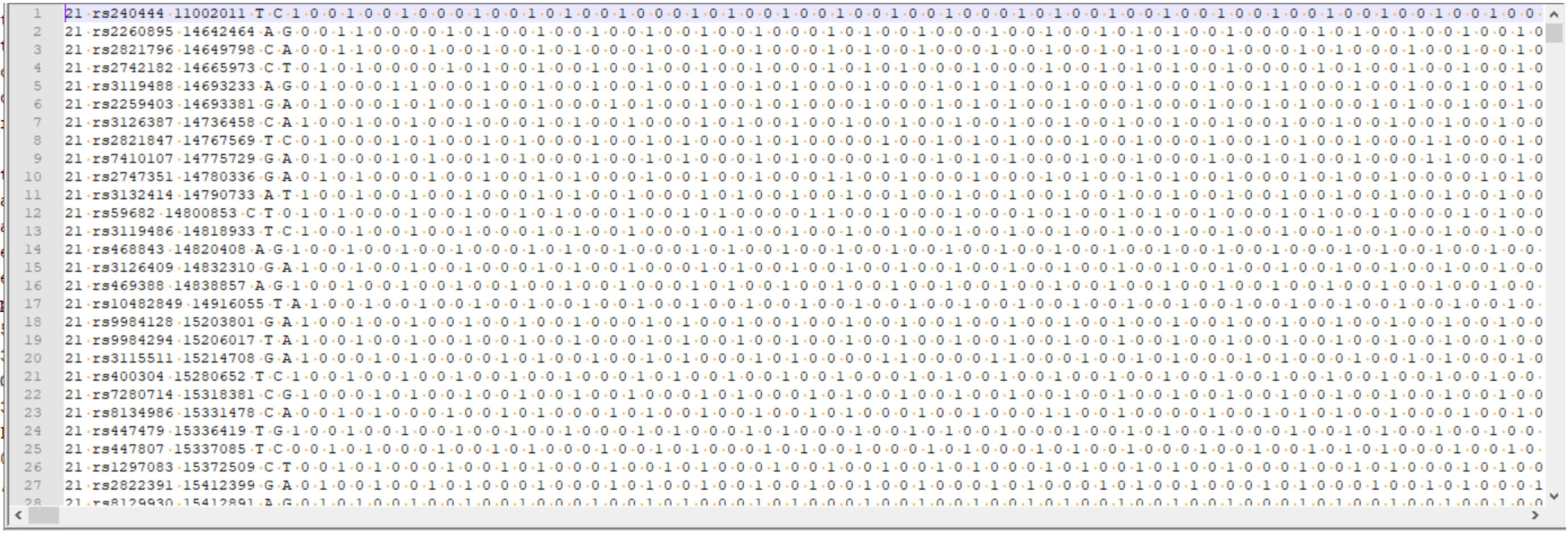
.ped/.bed



.vcf/.bcf

```
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=FAO,Number=A,Type=Integer,Description="Flow Evaluator Alternate allele observation count">
##FORMAT=<ID=FDP,Number=1,Type=Integer,Description="Flow Evaluator Read Depth">
##FORMAT=<ID=FRO,Number=1,Type=Integer,Description="Flow Evaluator Reference allele observation count">
##FORMAT=<ID=FSAF,Number=A,Type=Integer,Description="Flow Evaluator Alternat
##FORMAT=<ID=FSAR,Number=A,Type=Integer,Description="Flow Evaluator Alternat
##FORMAT=<ID=FSRF,Number=1,Type=Integer,Description="Flow Evaluator referen
##FORMAT=<ID=FSRR,Number=1,Type=Integer,Description="Flow Evaluator referen
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality, the Phi
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observat
##FORMAT=<ID=SAF,Number=A,Type=Integer,Description="Alternate allele observa
##FORMAT=<ID=SAR,Number=A,Type=Integer,Description="Alternate allele observa
##FORMAT=<ID=SRF,Number=1,Type=Integer,Description="Number of reference obse
##FORMAT=<ID=SRR,Number=1,Type=Integer,Description="Number of reference obse
#CHROM  POS     ID  REF ALT QUAL     FILTER  INFO     FORMAT  Sample1 Sample2 Samp
chr1    2488153 .   A   G   4476.14 PASS    AC=4;AF=1.00;AN=4;DP=648;FDP=19
chr1    2491258 .   C   G   2611.42 PASS    AC=2;AF=0.500;AN=4;AO=146;DP=13
chr1    6528100 .   GGCCCCT GGCCCTC 10278.10    PASS    AC=2;AF=1.00;AN=2;A
chr1    6528468 .   C   T   1859.16 PASS    AC=2;AF=0.500;AN=4;AO=120;DP=89
chr1    6529188 .   C   T   11263.97    PASS    AC=2;AF=0.500;AN=4;AO=606;D
chr1    6529443 .   A   G   5283.78 PASS    AC=2;AF=0.500;AN=4;AO=331;DP=22
chr1    6529747 .   A   AT  1631.35 PASS    AC=1;AF=0.500;AN=2;AO=10;DP=478
```

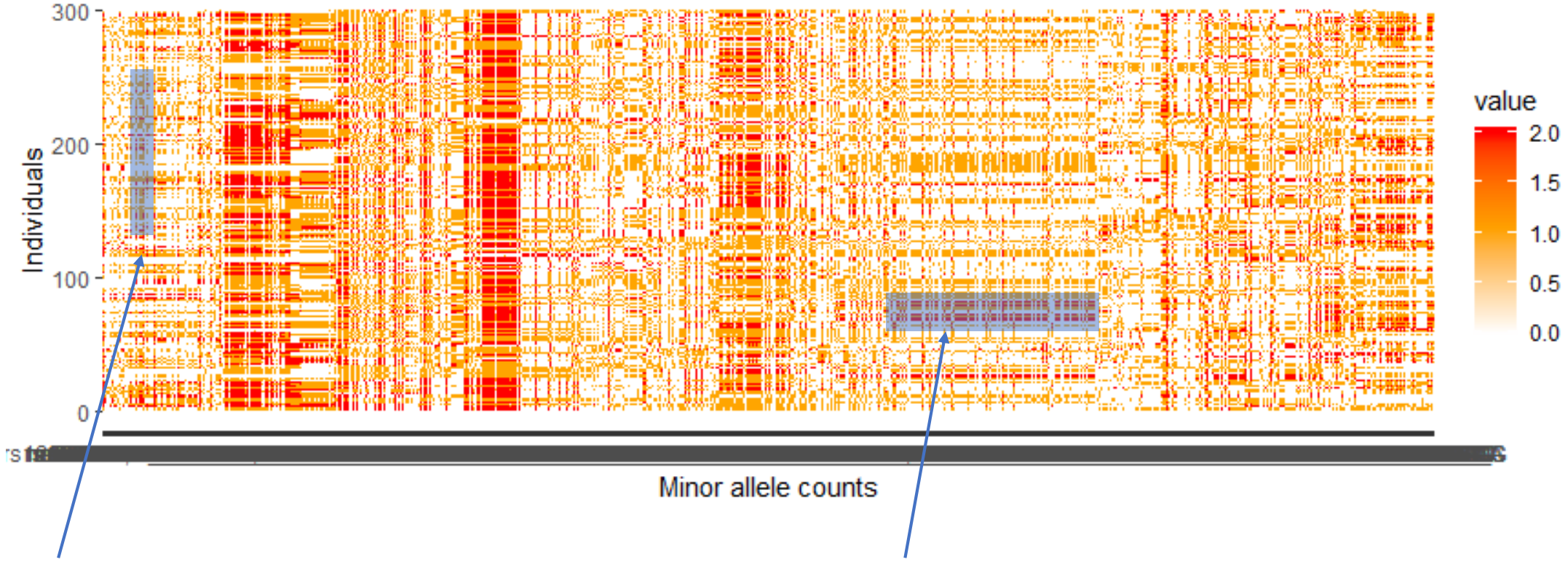
gen/.bgen



Some definitions

- A *locus L* is a well-defined position along a chromosome.
- single nucleotide variant (SNV) – variation in a single nucleotide
- In the population, different variants can exist at a locus. These variants are called *alleles*. A locus *L* is described by the set of alleles {A,G}, or {A,G,T}, {G,ACTAA}, {AGT,A}. One allele is fixed as the "reference allele".
- A *genotype* at a locus consists of a pair of alleles, one inherited from the father and one from the mother. A person is said to be *homozygous* at locus *L* if both alleles are the same (i.e., *A/A or G/G*) and is said to be *heterozygous* if the alleles are different (i.e., *A/G*).

Genetic data representation



Correlations due to relatedness

Correlations due to linkage

What is linkage?

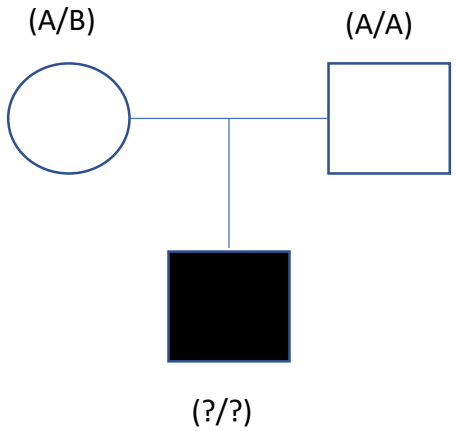
Gregor Mendel

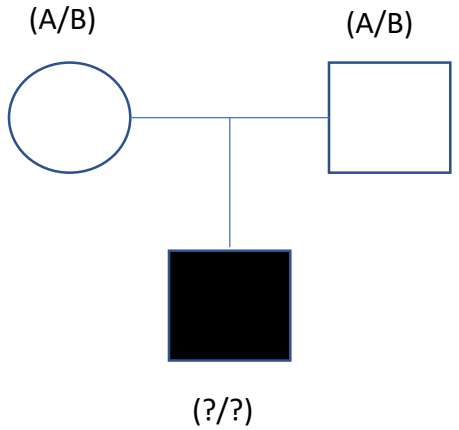
- Moravian Augustinian monk who founded the modern science of genetics
- Derived the principles of Mendelian inheritance (Mendel's laws)
- Used only dichotomous traits and assumed full penetrance, one disease – one locus

Mendel’s first law

- Mendel's First Law (Segregation): One allele of each parent is randomly and independently selected, with probability 1/2 , for transmission to the offspring; the alleles unite randomly to form the offspring's genotype.

Mendel first law





Mendel's first law

Father's Genotype	Mother's Genotype	Offspring's Genotype		
		dd	dD	DD
dd	dd	1	0	0
dd	dD	$\frac{1}{2}$	$\frac{1}{2}$	0
dd	DD	0	1	0
dD	dd	$\frac{1}{2}$	$\frac{1}{2}$	0
dD	dD	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
dD	DD	0	$\frac{1}{2}$	$\frac{1}{2}$
DD	dd	0	1	0
DD	dD	0	$\frac{1}{2}$	$\frac{1}{2}$
DD	DD	0	0	1

Table 2.1: Distribution of offspring's genotype conditional upon parental genotypes

Laird and Lange, 2010 "The fundamentals of modern statistical genetics"

Mendel's second law

- Question: are the genes underlying two traits transmitted independently?
- Mendel's Second Law (Independent Assortment): The alleles underlying two or more different traits are transmitted to offspring independently of each other; the transmission of each trait separately follows the first law of segregation.
- Linkage – lack of independent transmission, when the second law does not hold.

Mendelian diseases versus complex diseases

- Monogenic aka Mendelian diseases
  - Caused by one or few genes (highly penetrant)
  - Rare disease
  - Example: sickle-cell anemia, cystic fibrosis
- Complex disorder
  - Caused by several factors, including multiple genes and environment
  - Variants can increase/decrease risk to a particular disease (variable penetrance)
  - Example: COPD, Alzheimer’s disease, Schizophrenia

Penetrance functions

- P(Y=1|G) – penetrance function
- f_{DD} :=P(1|DD); f_{Dd} :=P(1|Dd); f_{dd} :=P(1|dd)

- Mendel's penetrance functions (complete penetrance):
  - Dominant mode of inheritance: $f_{DD}= f_{Dd}=1$; $f_{dd}=0$
  - Recessive mode of inheritance: $f_{DD}=1$; $f_{Dd}=f_{dd}=0$

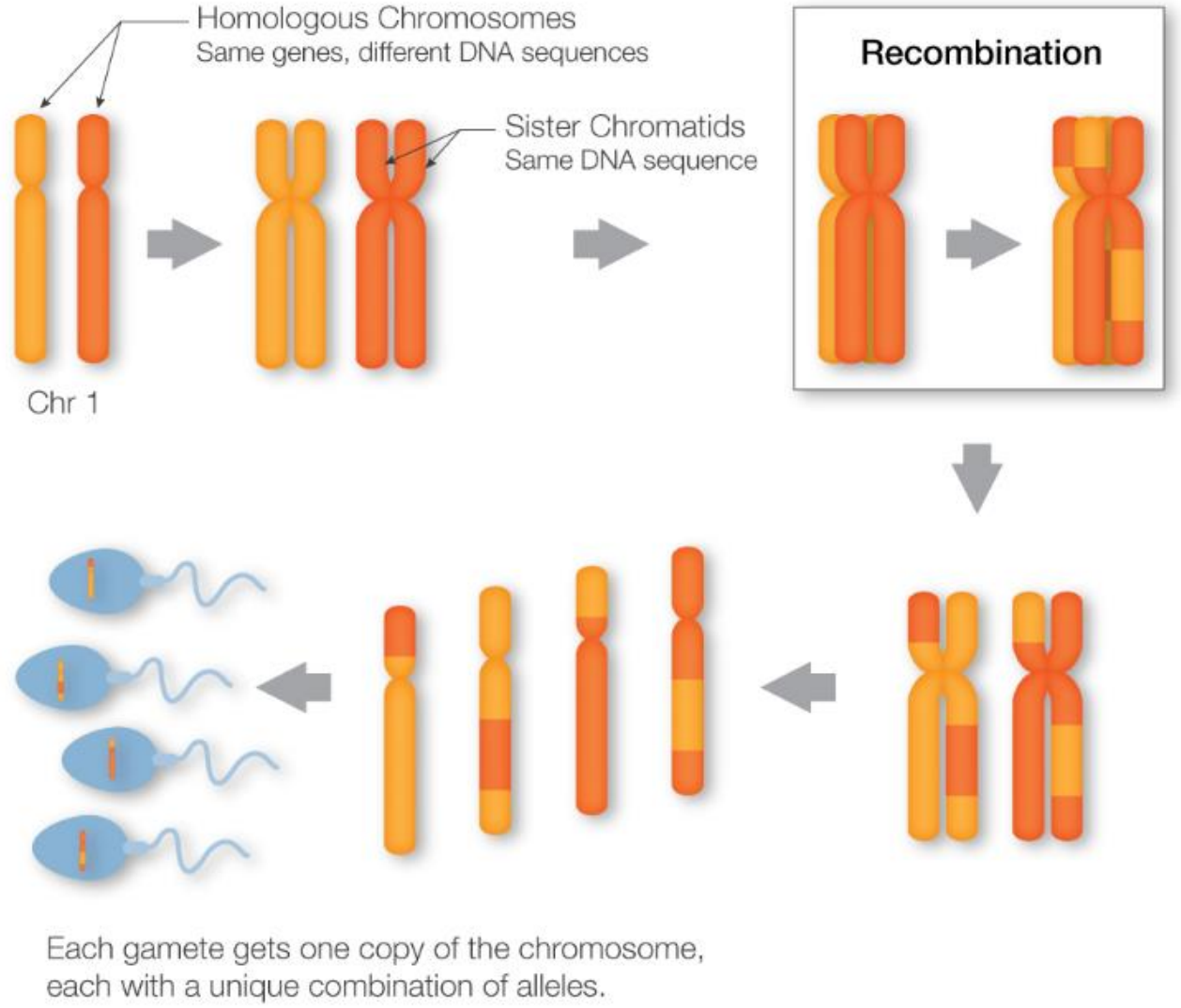
Realistic penetrance for complex diseases

- $P(Y|G)<1$
- For dominant model: $f_{DD}= f_{Dd}<1$; $f_{dd}>0$
- Other models
  - Recessive
  - Additive

Recombination

Genetic recombination happens as a result of the separation of genes that occurs during gamete formation in meiosis

- θ – recombination fraction between two loci
- θ = ½ under independent assortment



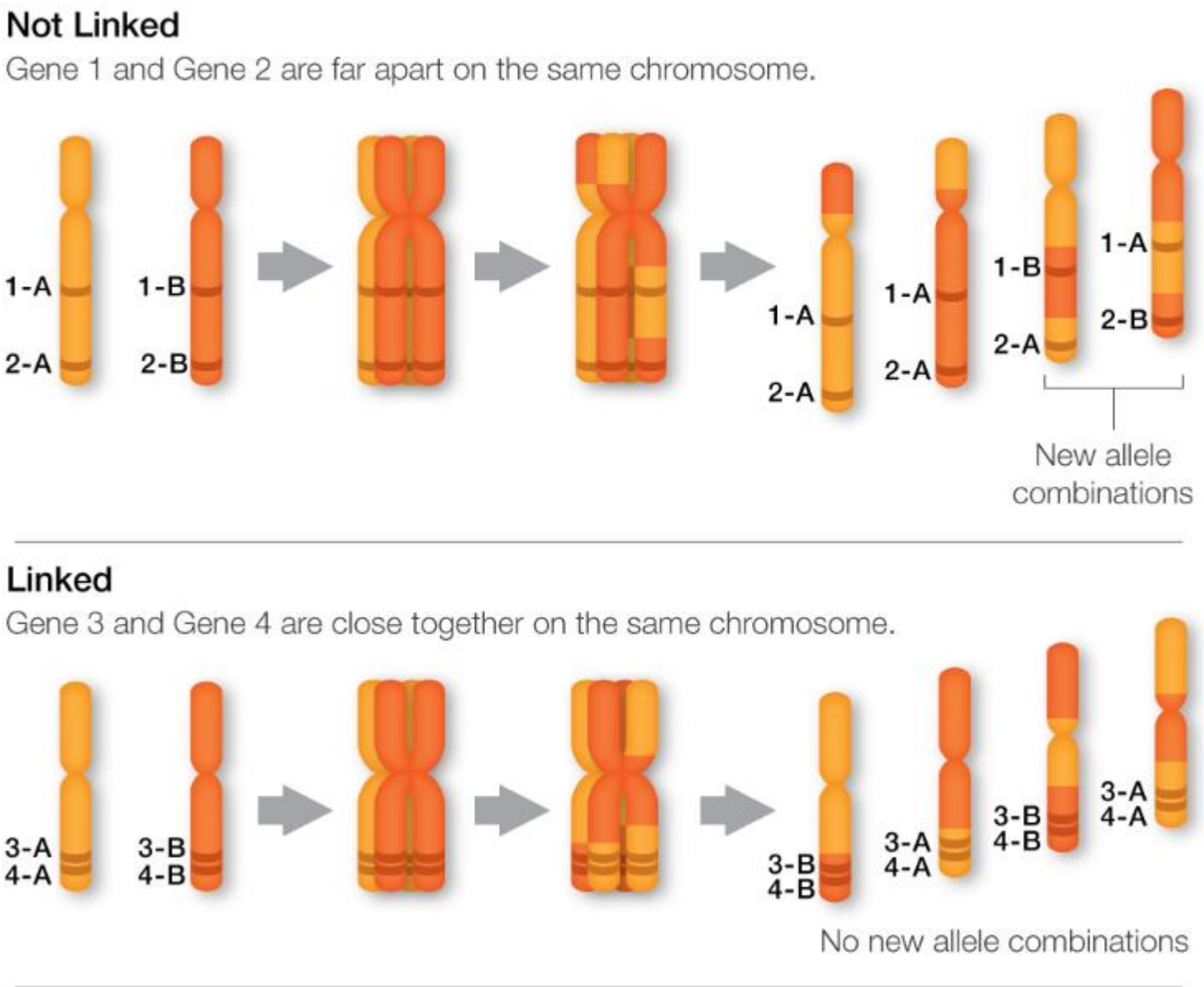
Recombination rearranges chromosomes, generating new allele combinations. While just one homologous chromosome pair is shown above, the same process happens for all of them.

Courtesy of https://learn.genetics.utah.edu/content/pigeons/geneticlinkage/

What is Linkage

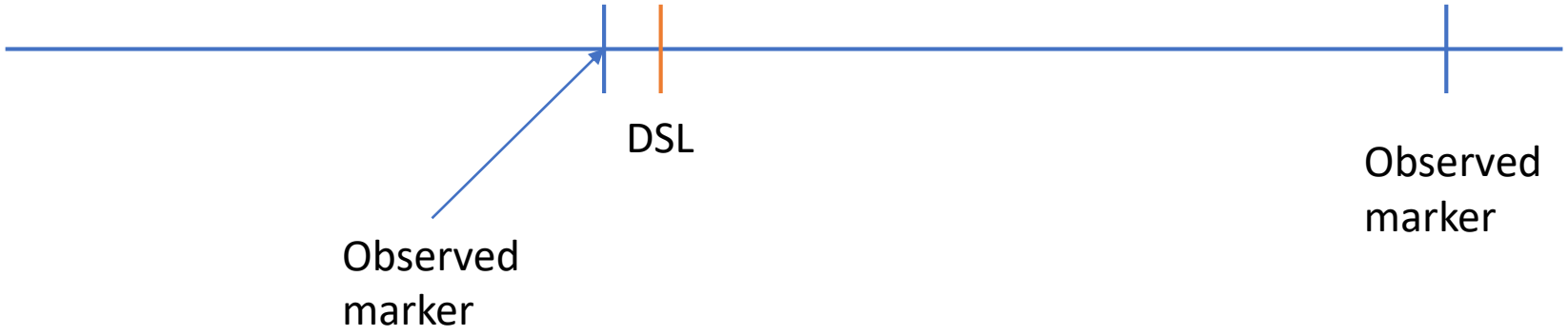
- Genetic linkage is a physical concept.
- Two loci are linked, if they are physically near to each other on the chromosome and hence unlikely to be separated during recombination. $\Theta \neq 1/2$.
- Which markers are perfectly unlinked?

Linkage example



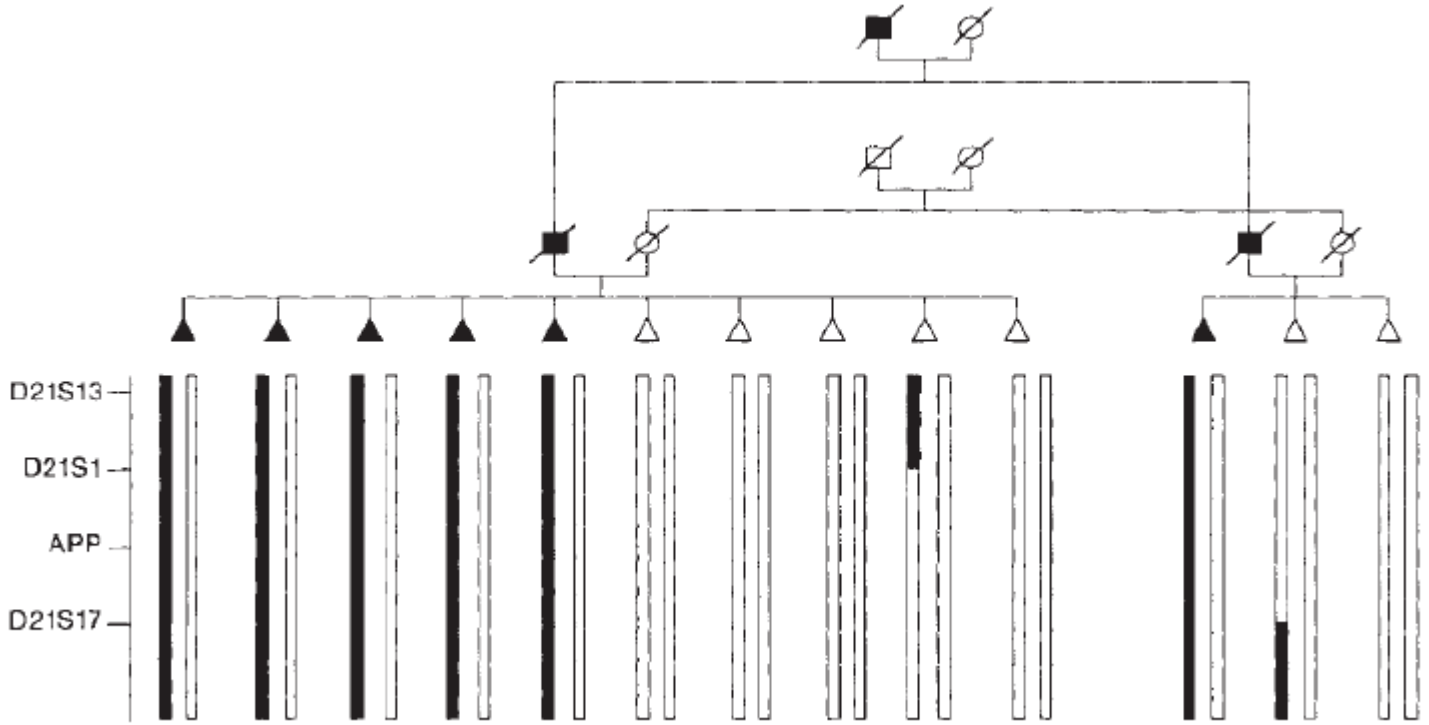
Courtesy of
https://learn.genetics.utah.edu/content/pigeons/geneticlinkage/

Linkage analysis (mapping)



- Approaches that use joint transmission of affection status and alleles at the observed marker locus to localize the DSL (disease sucseptability locus)

Linkage of early onset AD to *APP* gene



Source: Goate et al (1991)

Advantages and disadvantages

- Few genome-wide markers required
- Requires being able to infer relationship between variant at DSL and disease trait. Easy with Mendelian diseases and known mode of inheritance

- - may implicate large regions with many genes
- - Requires family data, which might be difficult to collect

Linkage methods

- Parametric linkage analysis
  - H0: θ = ½, i.e. marker and DSL are unlinked
  - H1: θ <1/2
  - LOD score:

$$Z(\theta) = \log_{10} \frac{\prod_{i=1}^{n} L_i(\theta \mid y_i)}{\prod_{i=1}^{n} L_i(\theta = 1/2 \mid y_i)} = \sum_{i=1}^{n} \log_{10} \frac{L_i(\theta \mid y_i)}{L_i(\theta = 1/2 \mid y_i)}$$

  - Requires specification
    - Disease model
    - Marker parameters (allele frequencies)
- Nonparametric linkage analysis
  - Based on IBS and IBD (allele sharing)

What is association

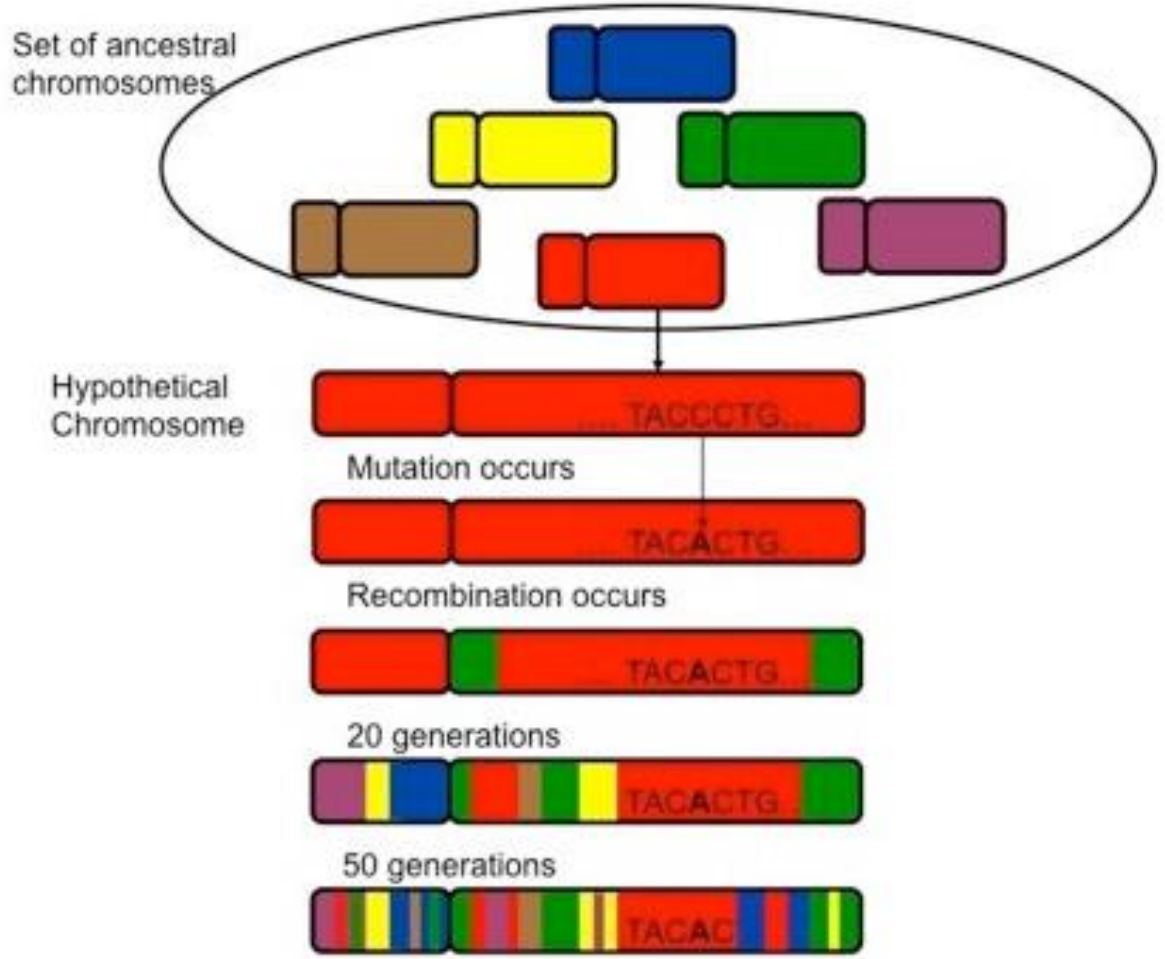
What is association

- Allelic association (linkage disequilibrium)
  - Allelic values at DSL and marker are associated in the population
  - Population-based concept

- Association (broad term)
  - Testing statistical independence between a variant and a trait of interest.
  - Can use population-based (unrelated) or family-based designs
  - GWAS – genome-wide association study
  - WGAS – whole genome association study

LD

- Non-random sharing of combinations of genetic variants within a **population** due to the history of recombination, mutation, and selection in a genomic region
- Linkage disequilibrium – population concept, exists when two loci are close (up to 500kb)
- There are regions of long range LD – conserved genetic regions
- 3 measures of LD



G. McVean: Handbook of Statistical Genetics
Picture from Dick et al. 2011

LD

- Consider 2 diallelic loci {A,a} and {B,b}. If the occurrence of alleles A and B are independent events, then $pAB = pA \cdot pB$. And the alleles are said to be in linkage equilibrium.

Population Allele Frequencies at Two Loci under Linkage Equilibrium

	B Locus		
A Locus	B	b	Total
A	$p_{AB} = p_A p_B$	$p_{Ab} = p_A p_b$	p_A
a	$p_{aB} = p_a p_B$	$p_{ab} = p_a p_b$	p_a
Column Total	p_B	p_b	

Laird and Lange, 2010 “The fundamentals of modern statistical genetics”

Coefficient D

- Measures the departure from independence
- $D = p_{AB} - p_A p_B$
- D=0 – linkage equilibrium
- Highly sensitive to marginal values (allele frequencies)

Coefficient D’

- “Normalized” D
- $D' = \begin{cases} \frac{D}{D_{max}}, if\ D < 0 \\ \frac{D}{D_{min}}, if\ D > 0 \end{cases}$
- Dmax=$\min(p_A p_b, p_a p_B)$
- Dmin=$-\max(p_A p_B, p_a p_b)$
- Ranges from 0 to 1

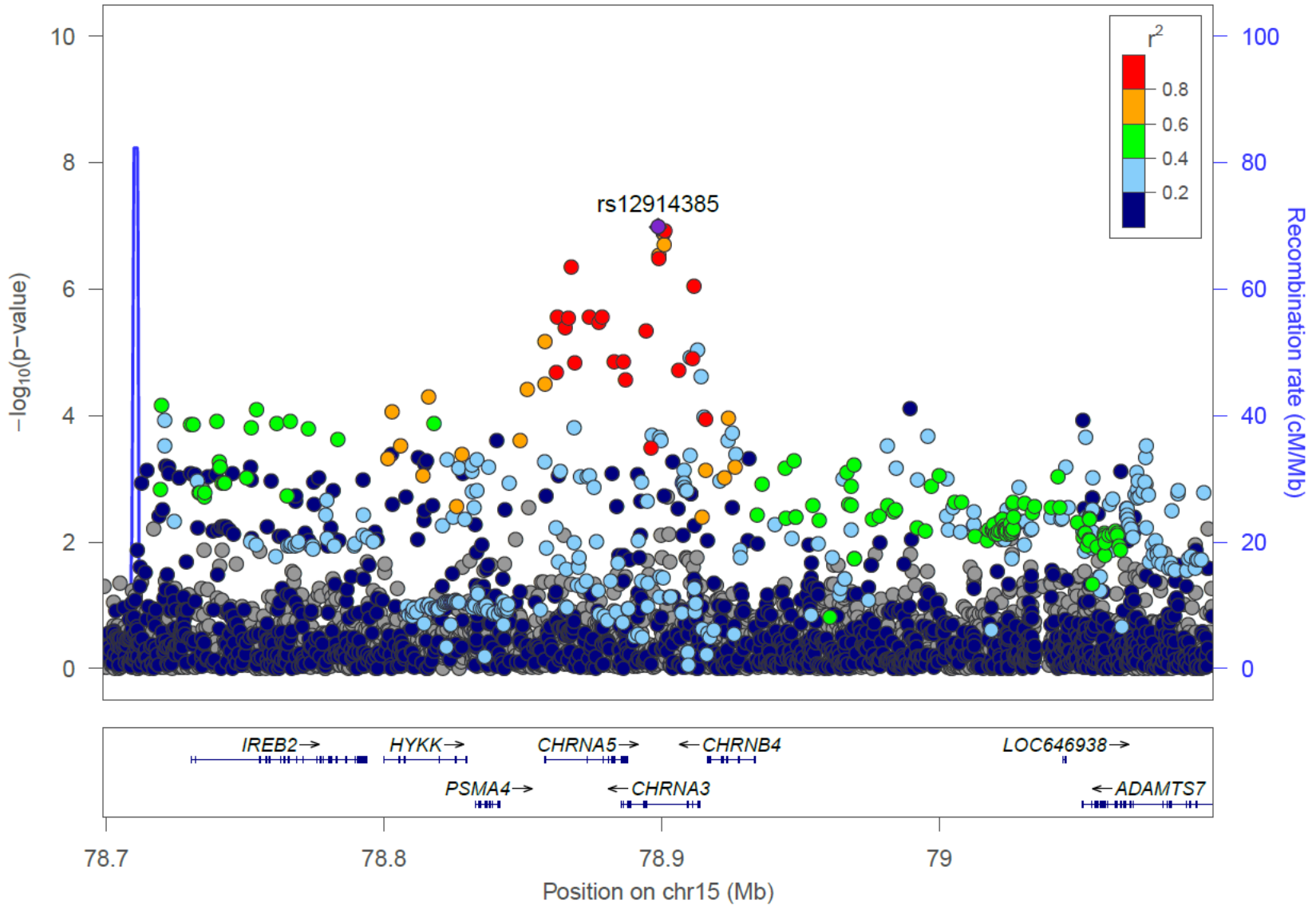
R or R^2

- Correlation coefficient between 2 markers
- $r = p_{AB} - pApB/\sqrt{p_A p_B p_a p_b} = D/\sqrt{p_A p_B p_a p_b}$

		B Locus		Row
	A Locus	B	b	Total
	A	43	27	70
	a	2	28	30
Column Total		45	55	100

$$D = (43 - 70 * 45/100)/100 = 0.115$$
$$D_{\max} = \min(70 * 55, 30 * 45)/10,000 = 0.135$$
$$D' = 0.115/0.135 = 0.8581$$
$$r = 0.115/\sqrt{0.7 * 0.3 * 0.55 * 0.45} = .5044$$

Laird and Lange, 2010 "The fundamentals of modern statistical genetics"


rs12914385
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
IREB2
HYKK
CHRNA5
CHRNB4
LOC646938
PSMA4
CHRNA3
ADAMTS7
Position on chr15 (Mb)

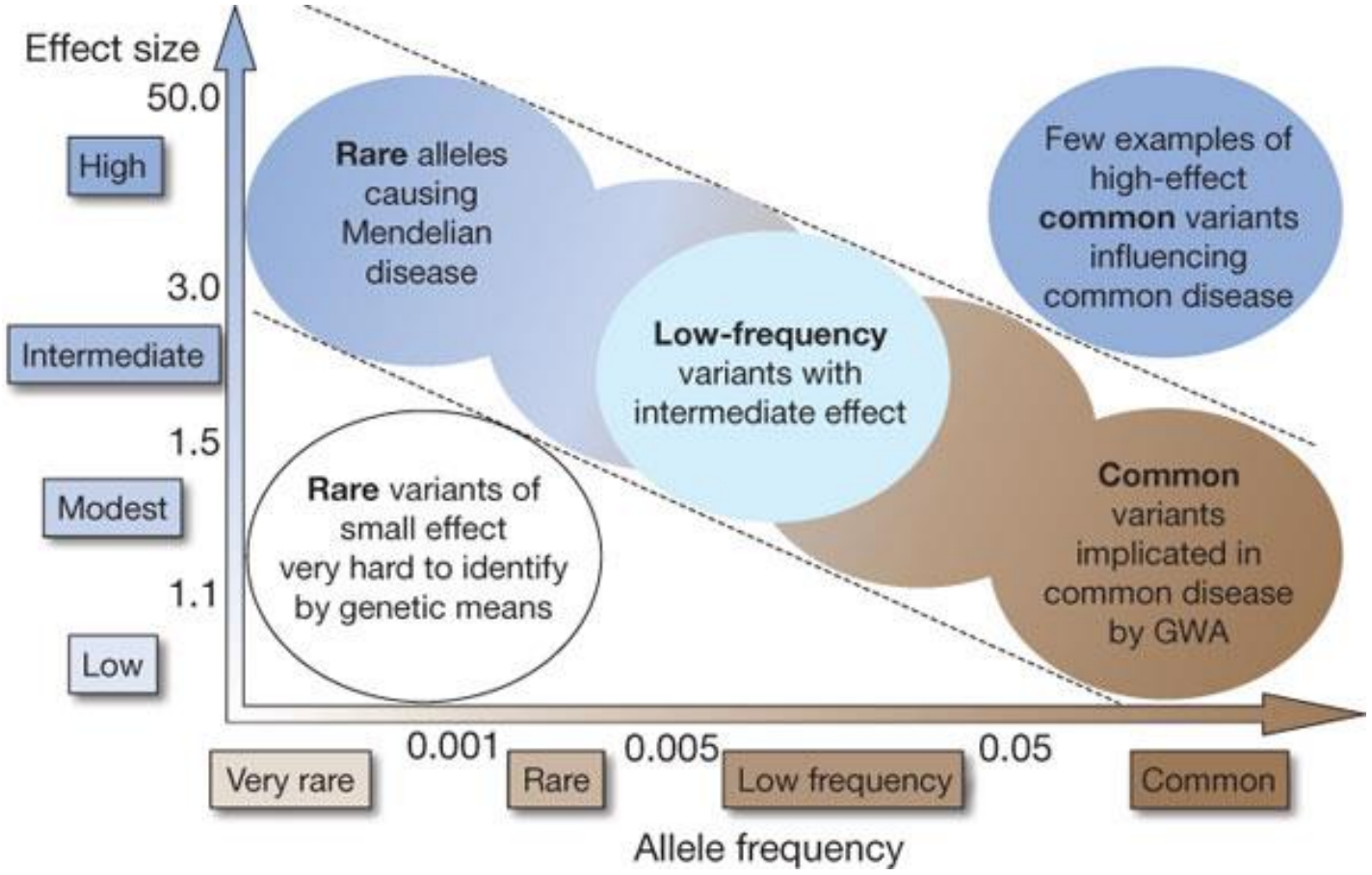
Allele frequencies

- Allele frequency – proportion of chromosomes in population/dataset having the allele of interest.
- {A,a}, n_{AA},n_{Aa},n_{aa}
- $\hat{p} = (n_{Aa} + 2n_{AA})/n$
- $\hat{q} = 1 - \hat{p}$

Example:

	AA	Aa	aa	$\hat{p}$	$\hat{q}$
Dataset 1 (n=1000)	90	420	490	0.3	0.7

Strength of genetic effect versus allele frequency



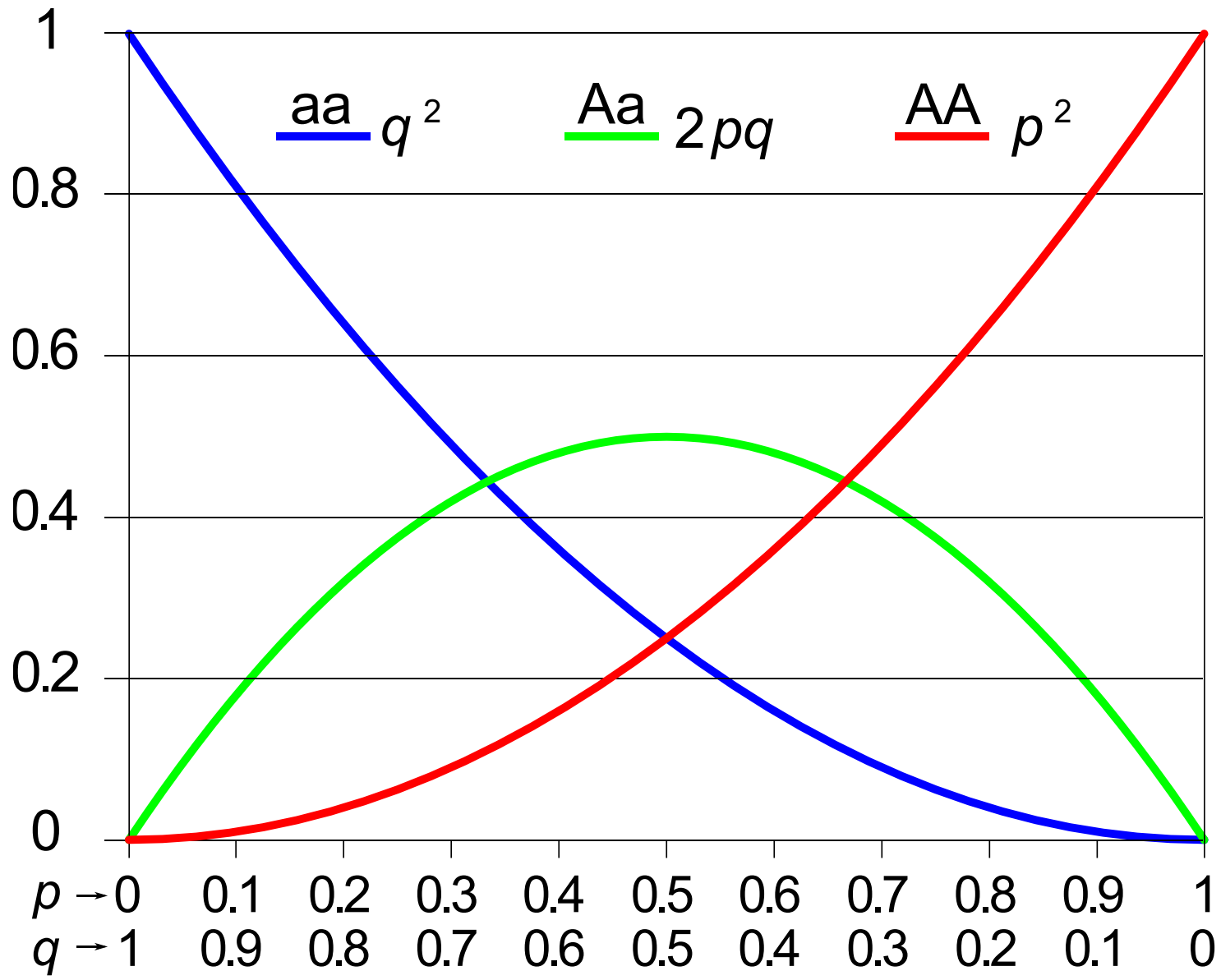
TA Manolio et al. *Nature* 461 (2009)

HWE (Hardy-Weinberg equilibrium)

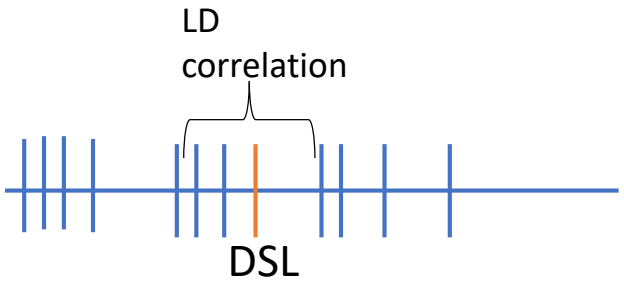
- A population is said to be in HWE if the genotypes satisfy:
- $P(AA) = p^2$;
- $P(Aa) = 2pq$;
- $P(aa) = q^2$;

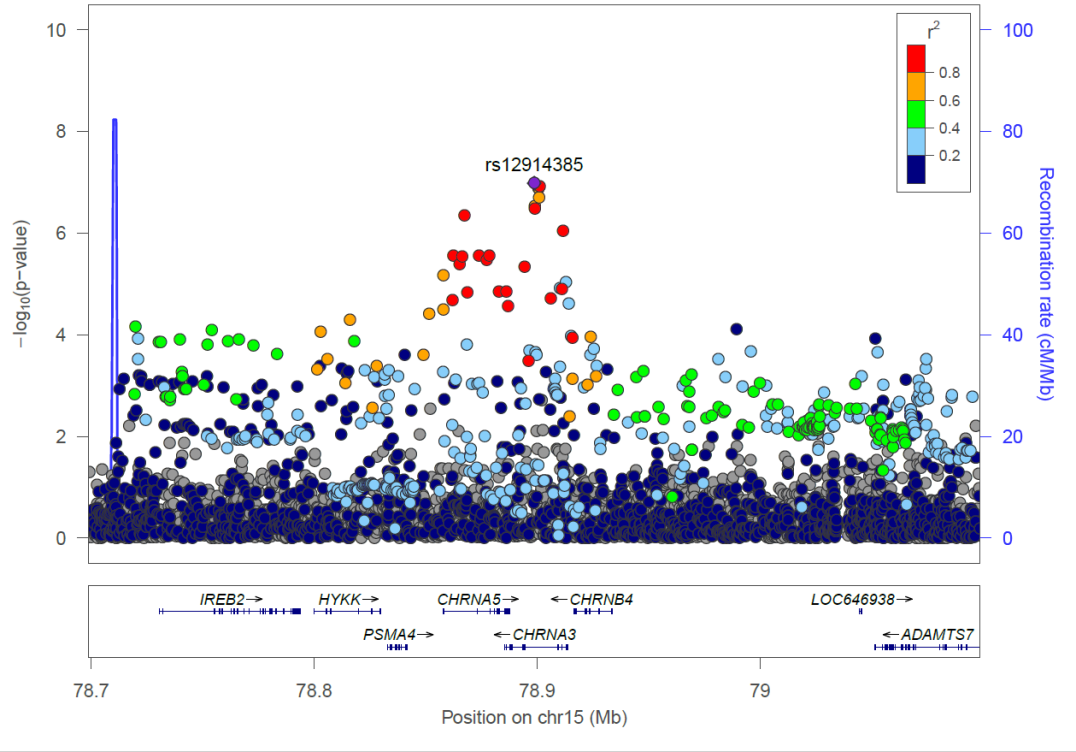
- Implication: Allele frequencies does not change in a population from generation to generation.
- Useful to identify genotyping errors, population structure

HWE



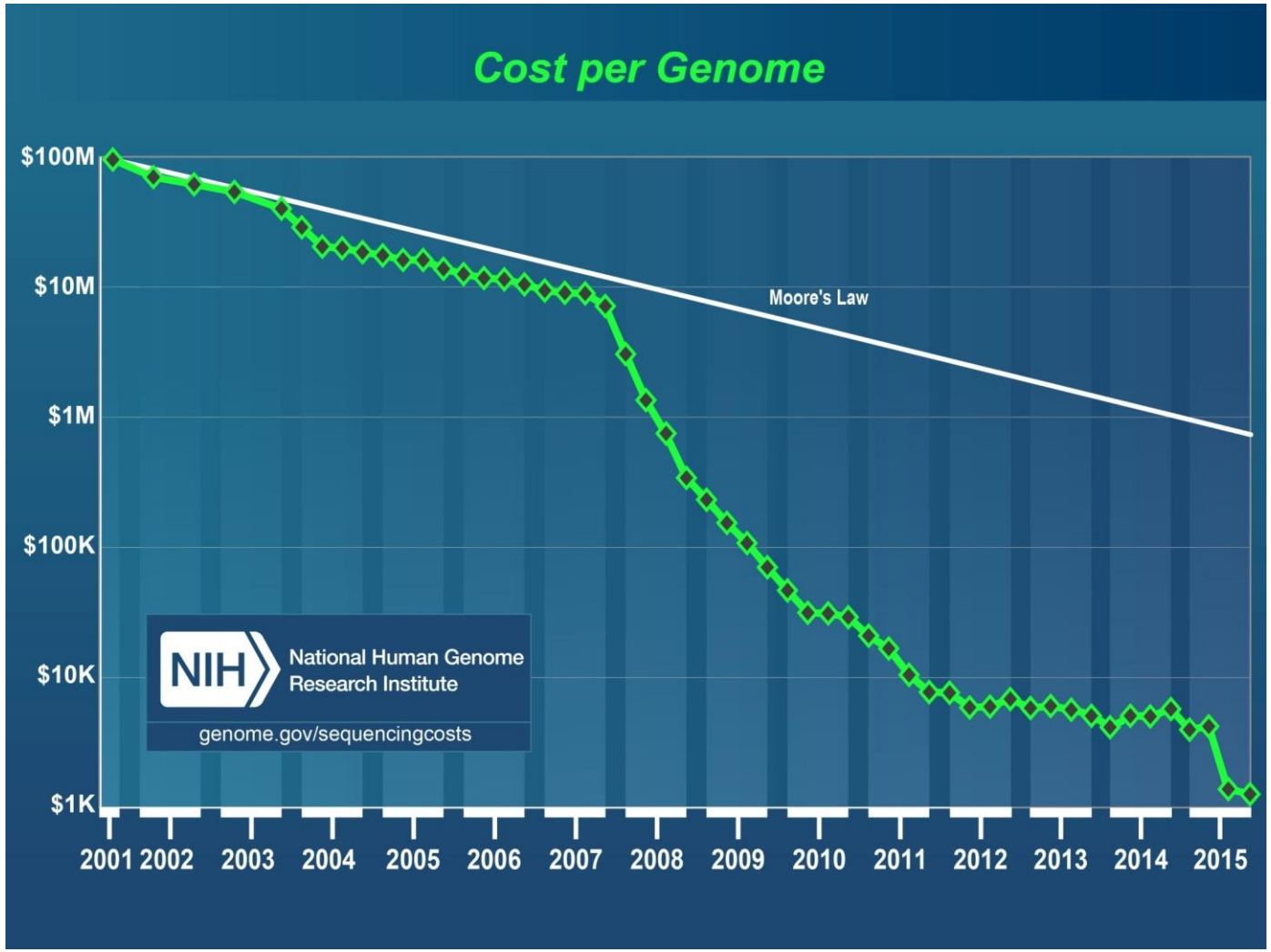
Using LD in association studies





$$\mathbf{Y} \sim \boldsymbol{\beta G} + \boldsymbol{\gamma Z} + \boldsymbol{\varepsilon}$$

Test for association between a phenotype Y and a marker G


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

Advantages and disadvantages of association studies

- Uses LD to localize DSL
- Better coverage through large number of markers
- Appropriate for complex traits

- A large number of individuals required -> different sources of bias
- Large number of markers -> multiple testing problem

Association methods (population-based)

- Linear and logistic regression models

$$\mathbf{Y} \sim \boldsymbol{\beta G} + \boldsymbol{\gamma Z} + \boldsymbol{\varepsilon}$$

  - Covariates include principal components to account for population stratification
- Mixed models (LMM and GLMM)
  - Allows to increase the sample size by including related individuals
  - Also efficiently accounts for population structure

Association methods (family-based)

- Test for both linkage and association
- Robust to population substructure: admixture, stratification, failure of HWE
- Requires genotyped families (parent-child, or siblings)
- TDT test for trio design (affected offspring)
- FBAT - generalization to general phenotypes, general pedigrees, missing parental genotypes, and multiple variants (Lake and Laird 2001, Laird and Lange 2006,…)

Linkage versus association

- Linkage is a physical concept: The two loci are “close’ together on the same chromosome. There is hardly any recombination between disease locus and marker locus
- Association is a statistical / population-based concept: A particular marker allele tends to be present with disease allele in a population and is associated with the phenotype.

Next lecture

- Genome-wide association studies
  - Population-based association studies
  - Family-based association studies
  - Common genetic variant methods
  - Rare genetic variant methods
  - Problems (QC, POPSTRAT)