Norwegian Institute of Public Health

UNIVERSITY OF BERGEN

Gentle Introduction To Genetic Epidemiology

— Lecture 2 —

Anil Jugessur

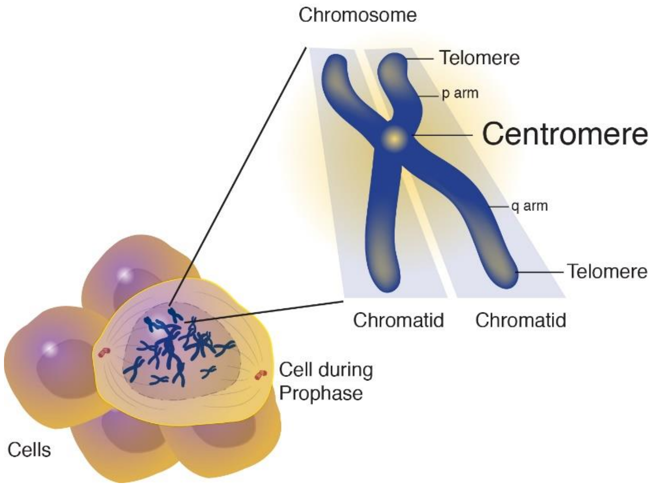
Senior scientist, Norwegian Institute of Public Health, NIPH, Oslo

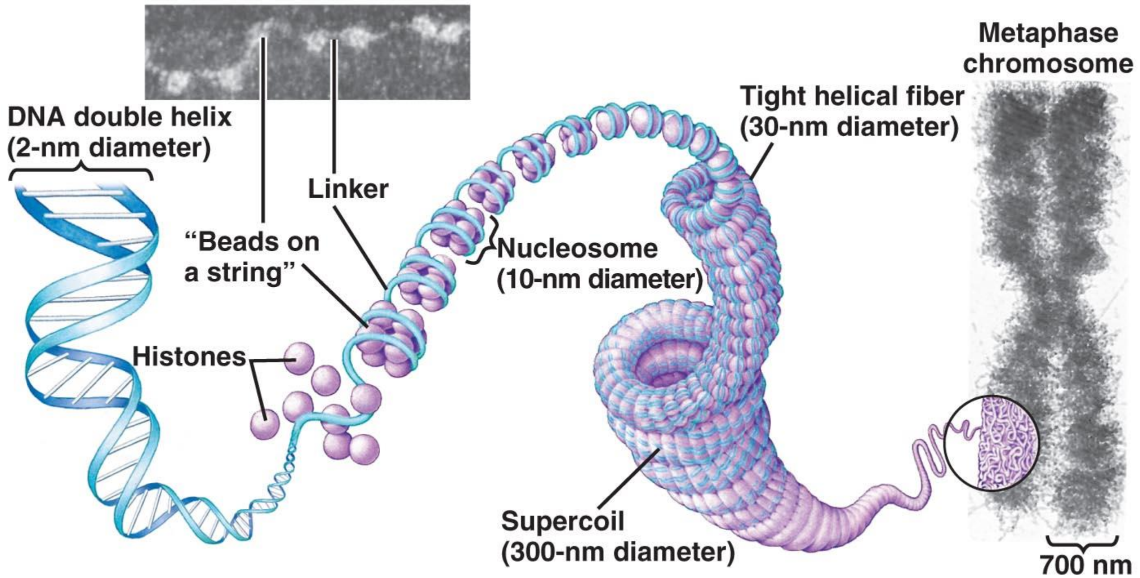
NORBIS GENSTAT-B course, 30th of May – 3rd of June, 2022

Lecture Outline

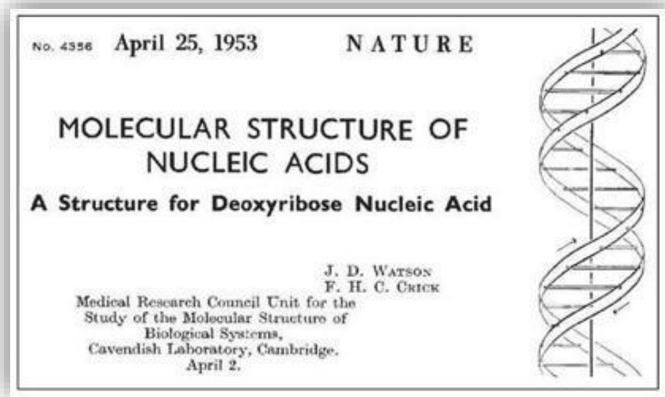
- ○ DNA, Exome, 1000 GP
- ○ Use of SNPs as genetic markers
- ○ Linkage disequilibrium and haplotypes
- ○ Population stratification

DNA: The Molecule of Life Itself – Ch. 1





Copyright © 2009 Pearson Education, Inc.



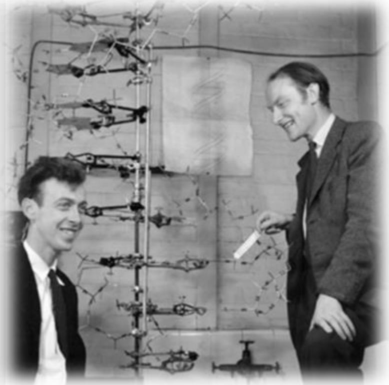
No. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

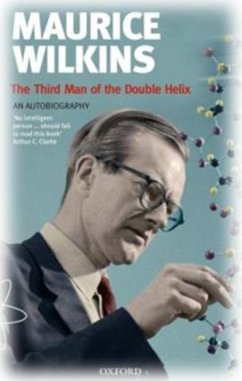
A Structure for Deoxyribose Nucleic Acid

J. D. Watson
F. H. C. Crick

Medical Research Council Unit for the
Study of the Molecular Structure of
Biological Systems,
Cavendish Laboratory, Cambridge.
April 2.



James Watson & Francis Crick

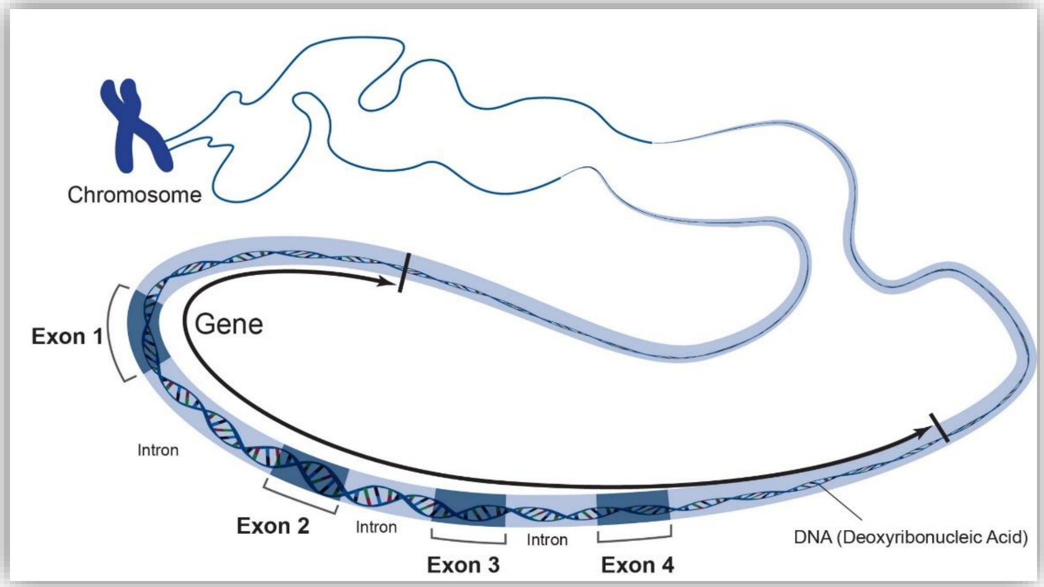


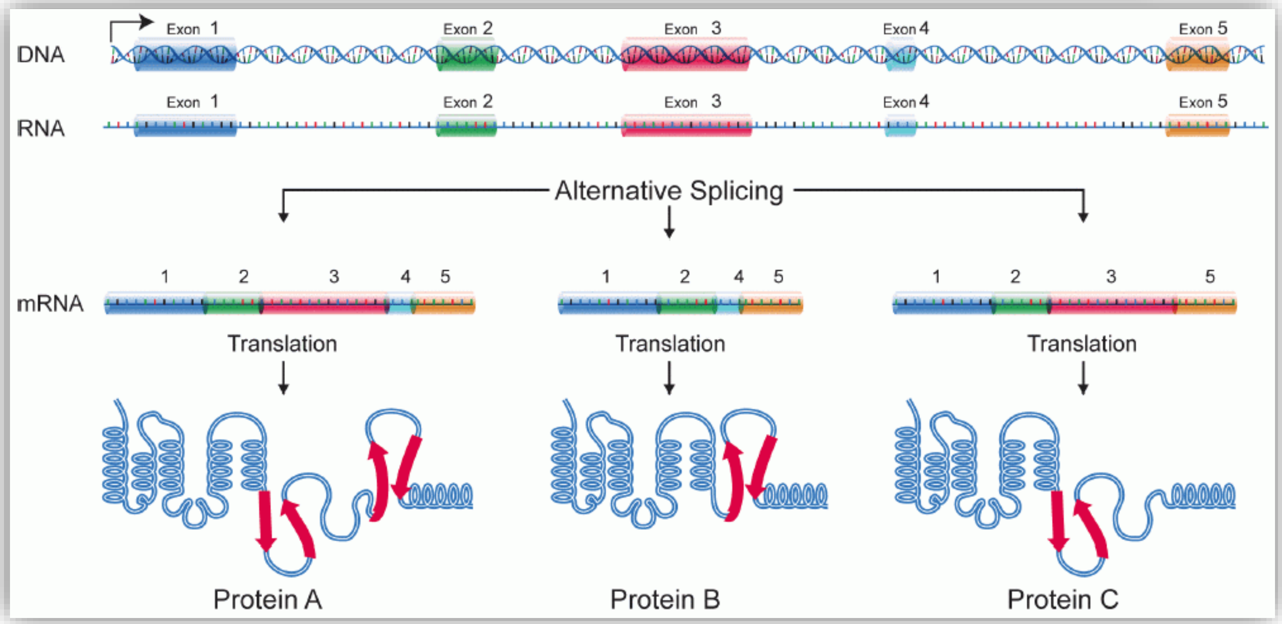
Maurice Wilkins

Rosalind Franklin
& Raymond Gosling

Source: Book «The Human Genome», 3rd Ed., by Richards & Hawley; and https://www.genome.gov/dmd/index.cfm?node=Photos/Graphics
https://www.genome.gov/sites/default/files/tg/en/illustration/chromatid.jpg

The «Exome»





Source: https://en.wikipedia.org/wiki/Alternative_splicing and https://www.genome.gov/dmd/index.cfm?node=Photos/Graphics

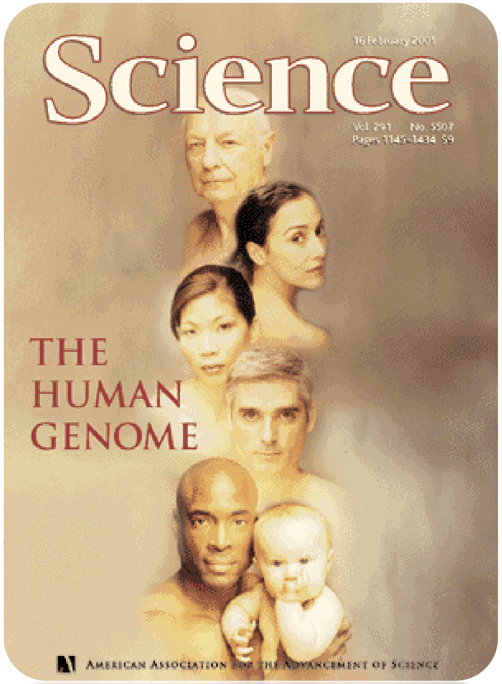
The Human Genome Project (HUGO)

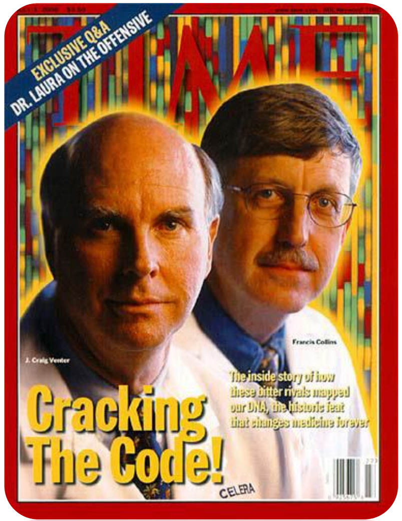
— Sequencing ~3 billion nucleotides —

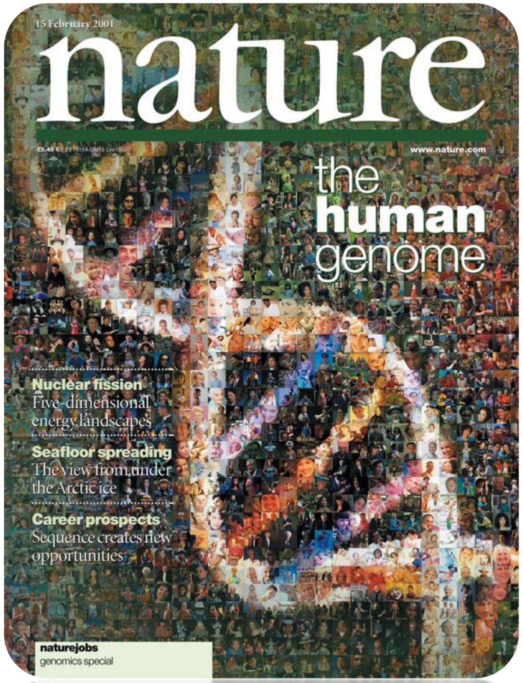
Celera Genomics (private)

1990-2003

The Public HGP




TIME
EXCLUSIVE Q&A
DR. LAURA ON THE OFFENSIVE
J. Craig Venter
Francis Collins
The inside story of how these bitter rivals mapped our DNA, the historic feat that changes medicine forever
Cracking The Code!
CELERA



The public project had a price tag of 2.7 billion USD in FY 2001!

HUGO Led To Two Other Grand Initiatives…

- **The HapMap Project**
  www.hapmap.org
- Officially started around Oct 2002
- 1,301 individuals from 14 different populations (HapMap phase III).
- Ended in June 16, 2016

- **The 1000 Genomes Project**
- www.1000genomes.org
- 7-yr project (2008-2015)
- The overall aim was to sequence 2,500 individuals from 26 populations.
- Massive amount of genomic data
  - Raw data ~180 Tb or 40,000 DVDs!

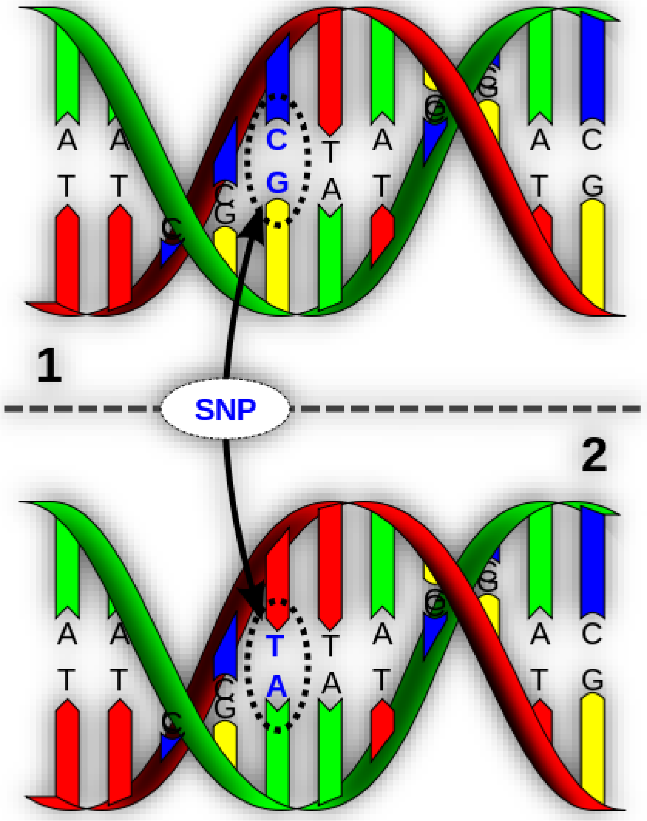
References: 1) An integrated map of structural variation in 2,504 human genomes *Nature 526, 75–81 (01 October 2015);*
2) A global reference for human genetic variation *Nature 526, 68–74 (01 October 2015)*

USING SNPS AS GENETIC MARKERS

Single

Nucleotide

Polymorphism



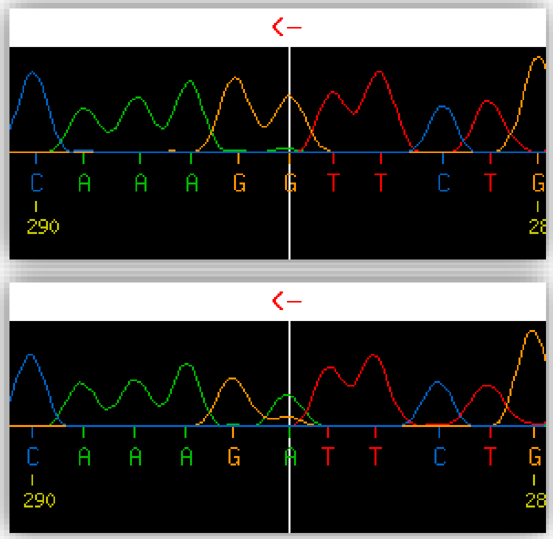
Source: http://en.wikipedia.org/

Variation In DNA Sequence – Ch. 3

- Any two copies of the human genome differ at about one in every thousand bp
  - Any two humans are approxmately 99.9% identical in their DNA sequences.
    - Nucleotide diversity is approximately 0,1 %

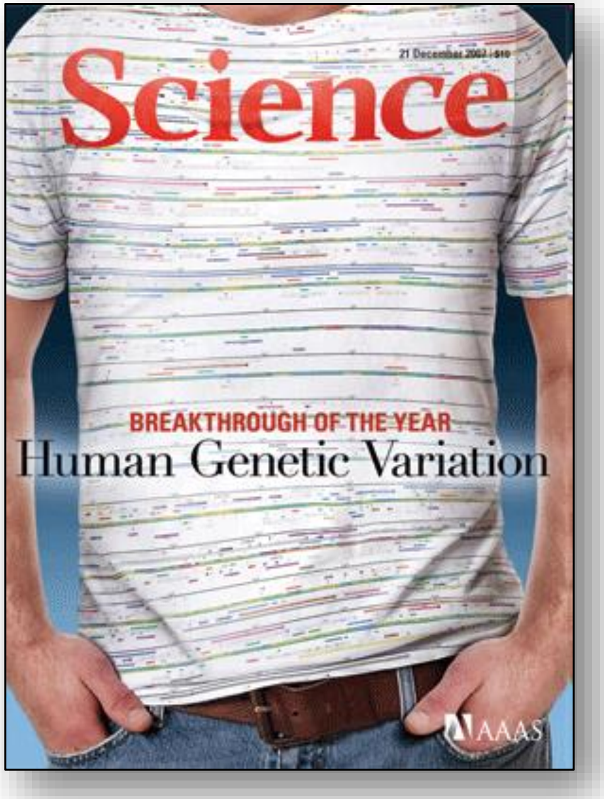
- SNPs are the most common type of genetic variation
  - >10 million SNPs with a frequency >1% in the human genome
    - Explain ~90 % of the total genetic variation.
- Due to their sheer abundance and genome-wide coverage, SNPs can be used as markers
  - Can be hard to test all 10 million SNPs!
    - Can take advantage of how SNPs and other genetic variants are organized on the chromosomes.

SNP G>A



5'...C-A-A-A-G-[**G/A**]-T-T-C-T-G... 3

Breakthrough Of The Year 2007



A T-shirt showing the annotations in a gene-sequence map of chromosome 1

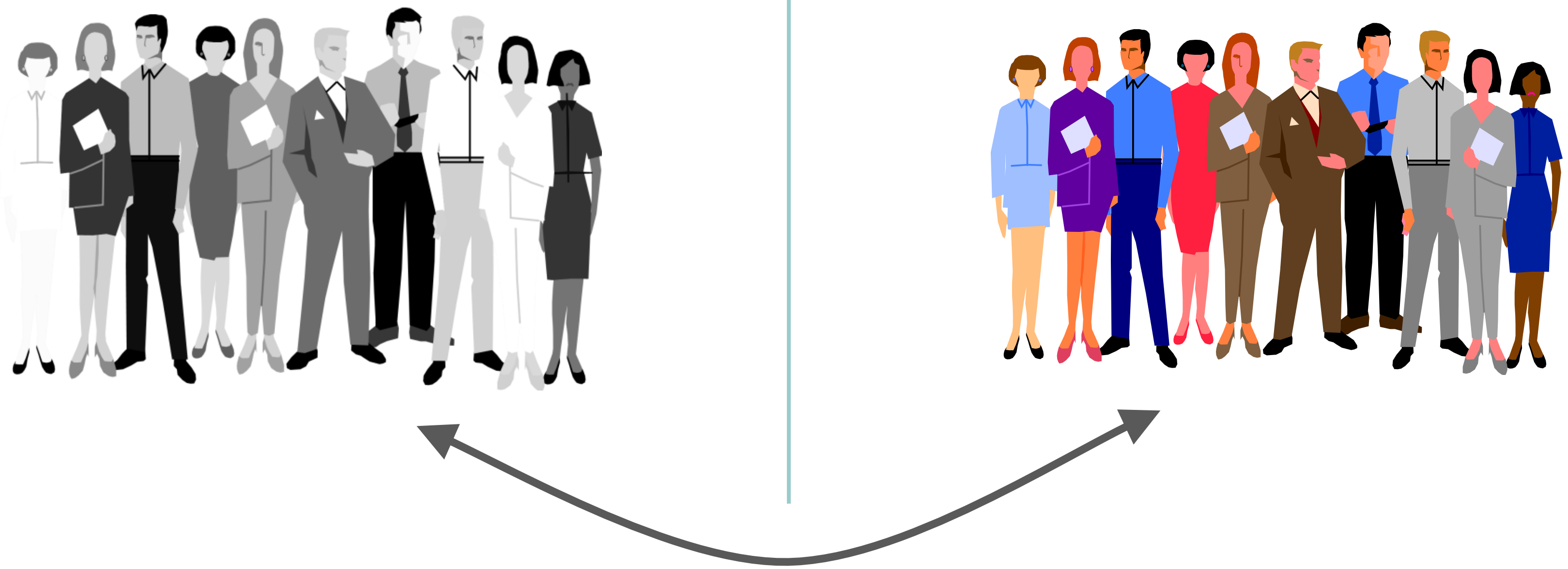
The DNA sequence across individuals differs a lot more than previously thought!

Using SNPs As Genetic Markers – Ch. 10-11

– The Case-Control Design –

“Case” (diseased) population

“Control” (healthy) population

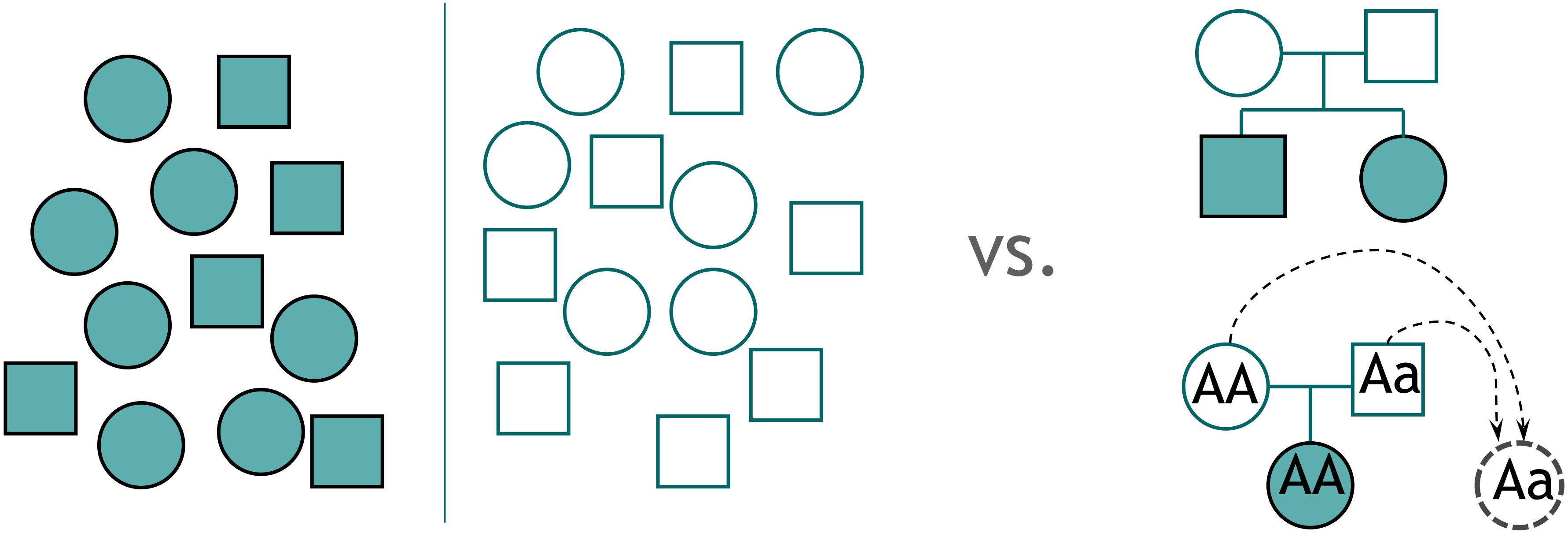


- Compare allele frequencies between cases and controls.
  - **Rationale?** ⇒ If an allele is more common among cases than controls, that SNP can be used as a marker to locate/identify the disease gene.

Case-Control and Nuclear Families

- With case-control data: Compare marker allele frequencies between an unrelated case and control population

- With nuclear family data: Use the non-transmitted parental alleles as control alleles.
  - Test for deviations from the expected 50% Mendelian transmission of an allele from parents to offspring.



Overview

DNA, Exome, 1000 GP

Use of SNPs as genetic markers

○ Linkage disequilibrium and haplotypes

Population stratification

Linkage Disequilibrium (LD)

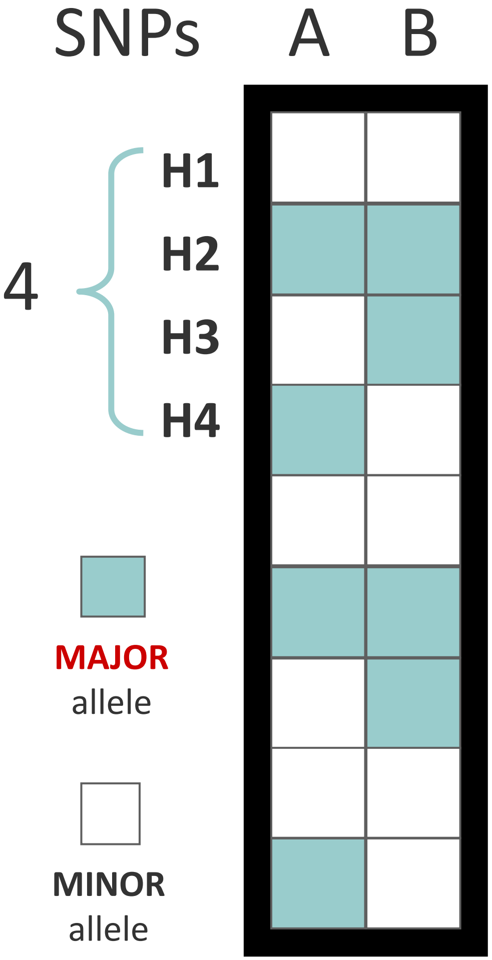
- Non-random association of specific alleles at two loci
  - We inherit chunks of DNA at a time.
- How do LD studies compare with linkage studies?
  - Linkage studies focus on finding disease markers using pedigree data – *families*
  - LD studies consider larger segments of the *population at large*, effectively tracking down ancestral haplotypes.
  - In populations where there is a high degree of inbreeding, linkage and linkage disequilibrium techniques will tend to converge.
- Rationale behind using LD in gene-mapping:
  - By detecting LD between nearby markers and the disease locus, we can narrow down the genetic interval around a disease locus («fine-mapping»).

LD IS A COMPLEX PHENOMENON…

- At the moment of creation, a newly-created allele is surrounded by a series of alleles and a unique haplotype is established.
    - Complete LD exists between the new allele and each of the nearby polymorphisms
    - The new allele is 100% predictive of the alleles nearby.
        - An allele at one SNP can be used as surrogate for an allele at another SNP.

- LD will gradually decay as a result of many processes
    - Recombination may change the pattern of LD.
    - Natural selection for or against certain sequences may drive alleles at adjacent loci to much higher or lower frequencies.
    - Population-specific demographic history, such as bottlenecks, admixture, inbreeding, migration, immigration, and assortative mating will also affect the regional distribution of LD.

LINKAGE *EQUILIBRIUM*

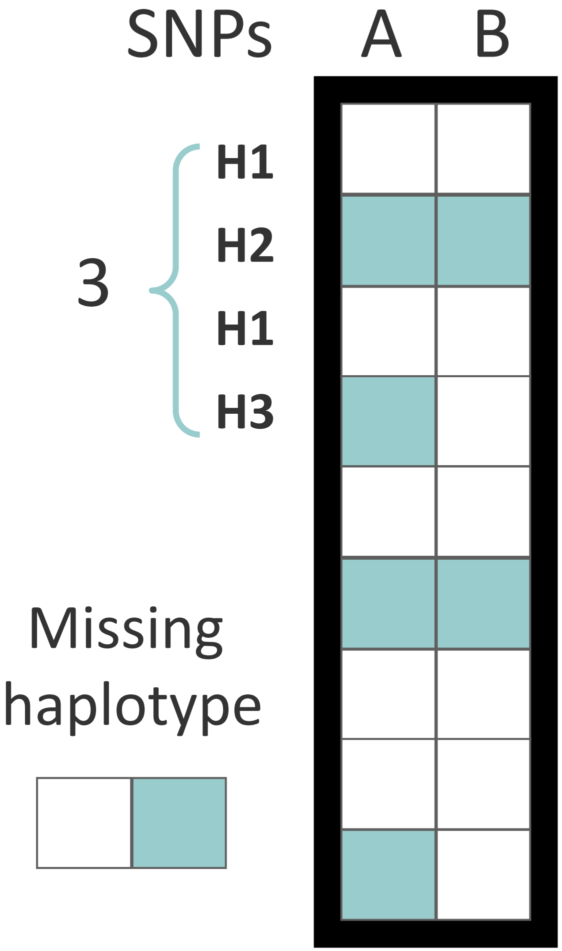


- "Linkage *Equilibrium*": alleles at the two loci are not correlated (completely independent).

- For the two SNPs A and B, there are $2^2 = 4$ possible haplotypes (H1, H2, H3, H4).
  - All 4 haplotypes H1-H4 are observed!
  - Observed haplotype frequency for a given haplotype is the simple product of the individual allele frequencies at the two loci.
    - As expected for a pair of independent events.

Adapted from: http://pgaedu.net/PublicPGAEDU/GtD_Presentations/nickerson.ppt

Strong LD



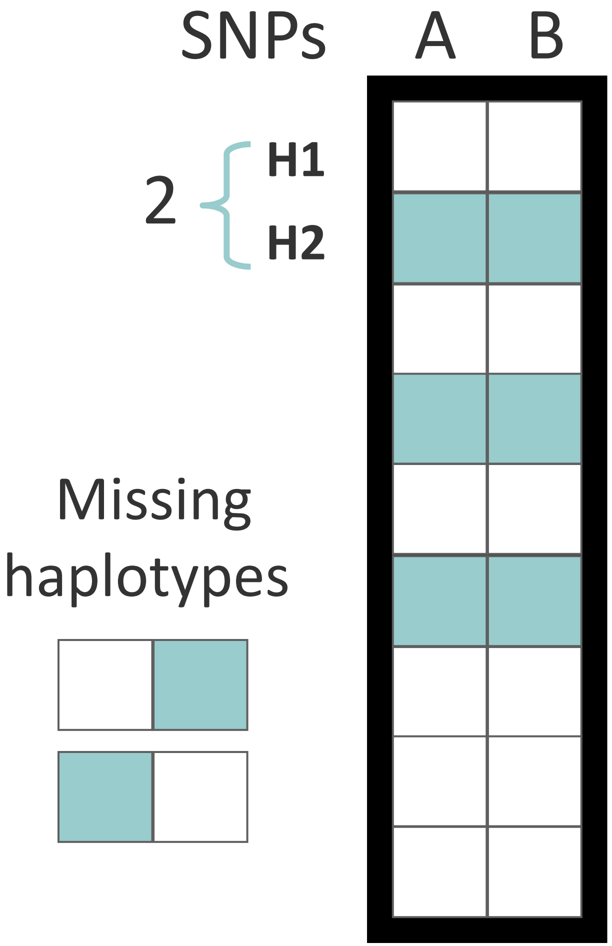
- Allelic association is strong, but not perfect.
  - Only 3 of the possible 4 haplotypes are observed
  - Recombination has not had enough time to create the missing haplotype (H4).
  - Frequency of a haplotype can no longer be predicted by the simple product of the individual frequencies of markers comprising the haplotype.
  - Allelic association is strong, but the genotypes are not perfectly correlated.

D´ = 1

$r^2 < 1$ ←

Adapted from: http://pgaedu.net/PublicPGAEDU/GtD_Presentations/nickerson.ppt

Perfect LD



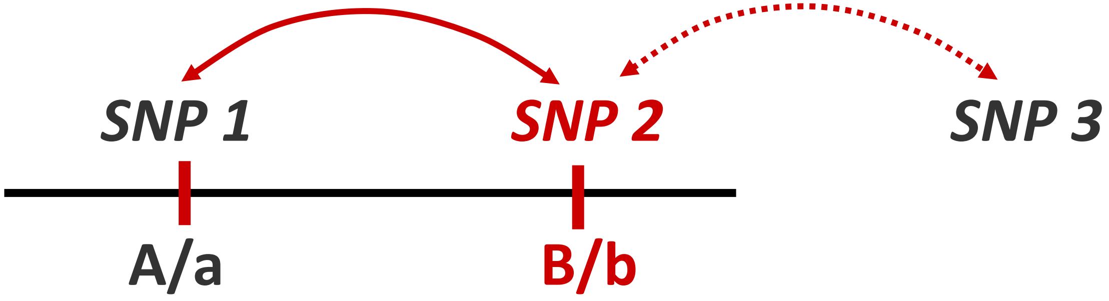
- Allelic association is as strong as possible.
  - Only 2 out of the 4 possible haplotypes are observed.
  - Frequency of the major or minor allele at both SNPs is identical.
  - No detected recombination between SNPs – two missing haplotypes
- Equal allele frequencies mean that:
  - Genotypes are 100% correlated.
    - SNP A predicts SNP B perfectly!

D´ = 1

r^2 = 1

Adapted from: http://pgaedu.net/PublicPGAEDU/GtD_Presentations/nickerson.ppt

LD Parameters D' and R^2



- LD can be assessed using parameter D = (O – E) = P(AB)-P(A)*P(B)
  - Lewontin's D' (absolute value of D) = D ÷ D_{max}

    where D_{max} is the lesser of these products: P(A)*P(b) or P(a)*P(B)
- r^2 value = $[P(AB)-P(A)*P(B)]^2$ ÷ [P(A)*P(b)*P(a)*P(B)]
- D' is less sensitive to allele frequency differences than r^2.
  - r^2 is a better measure for how well one SNP substitutes another SNP.
  - r^2 is a more useful measure for LD-mapping for power & sample size...
    - r^2 is inversely proportional to the sample size needed to find the same association using a substitute marker.
    - So, to find the same association using a 3rd SNP, simply increase the sample size by $1/r^2$ to achieve the same power!

Some Properties of *D'*

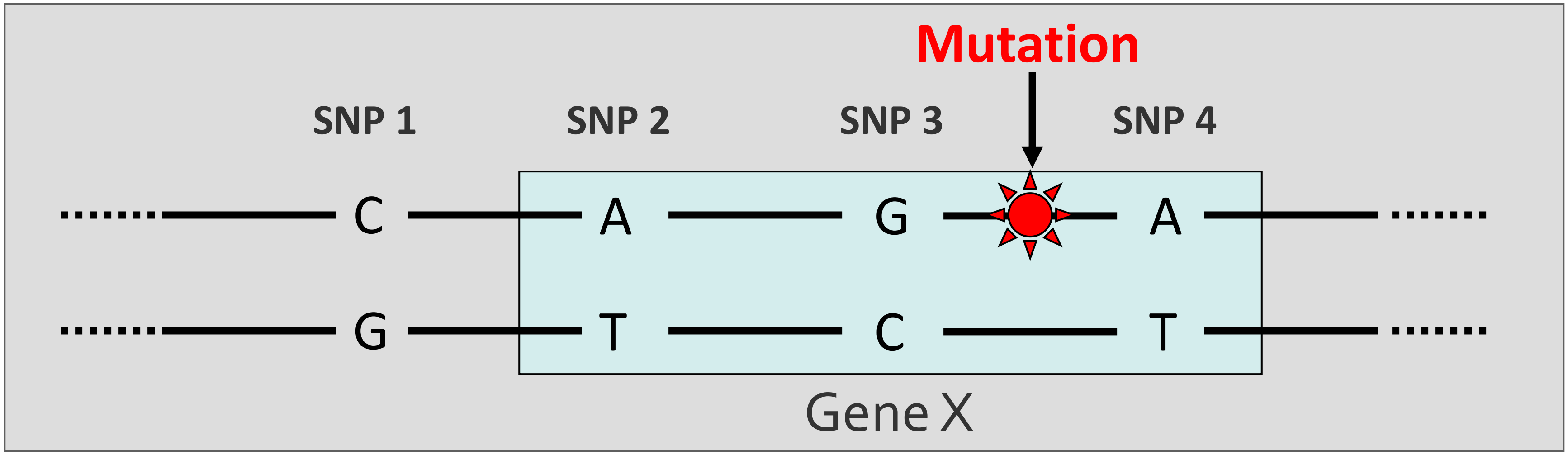
- D’ is scaled to remove effects of allele frequency differences.
  - Is thus less sensitive to allele frequency differences than r^2
  - For small sample sizes, D’ is biased *upwards* (towards 1.0)
    - Perfect LD (D’ = 1.0) may occur by chance.

- D’ does not perform well with low-frequency markers compared to common markers.
  - Complete LD (D’ = 1.0) may occur by chance.
  - Best to exclude markers of exceedingly low MAFs (<0.01).

- But D’ is a better measure of historical recombination than r^2
  - When defining blocks of LD, it is preferable to use a map based on D’ values.

HAPLOTYPES

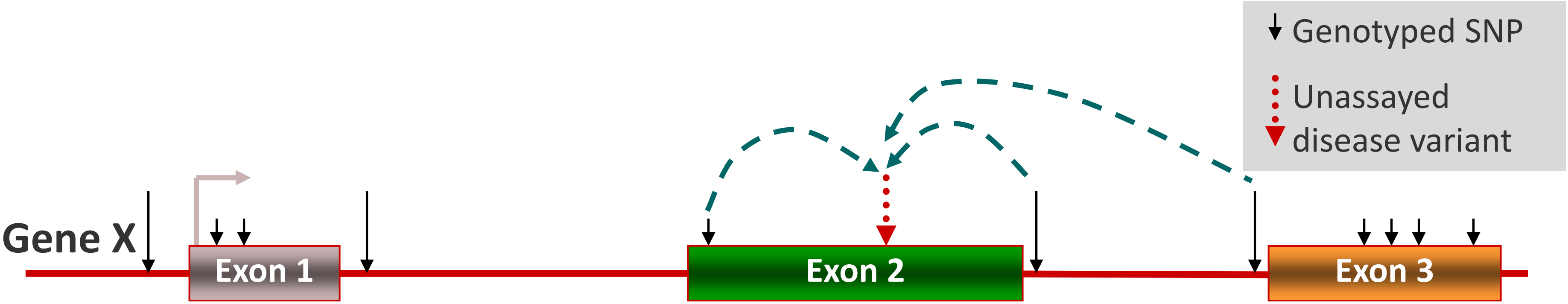
- A haplotype is a specific pattern of alleles on one chromosome.
- A mutation occurs in a specific haplotype.



- After multiple generations, recombinational events will break up the haplotype carrying the mutation.
  - Only the closest markers will maintain the strength of association.
  - The strength of LD between the SNPs is said to «decay/erode» with time.
- Gradually, the SNP alleles will be in linkage *equilibrium*
  - The observed haplotype frequency will be equal to the product of the individual frequencies – the expected frequency for two independent events.

WHY STUDY HAPLOTYPES?

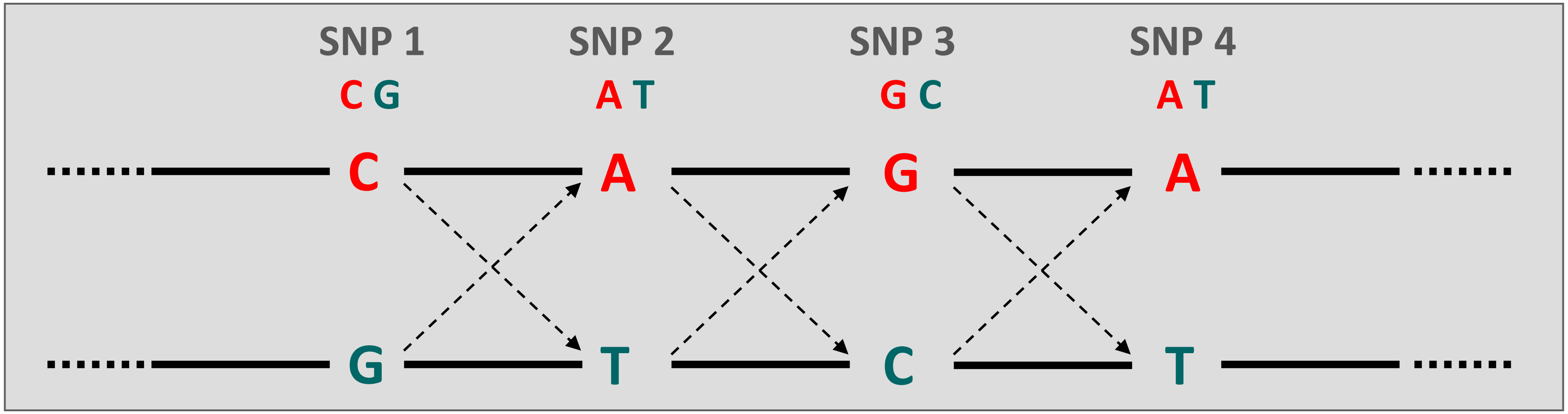
- Close correlation between alleles at one SNP and alleles at a nearby SNP within a gene because of LD.
  - Alleles not transmitted one at a time, independent of their neighbors, but rather as haplotypes.
- More information can be gained from using haplotypes.
  - Whereas a SNP has only 2 alleles, there are multiple different haplotype combinations.
  - Haplotypes can be surrogates for potentially unidentified or yet unassayed SNPs.



- More statistical power for association analyses using haplotypes.
- Use of haplotypes reduces the number of tests to be carried out.
  - Higher odds of being heterozygous for haplotypes than heterozygous for SNPs.
  - ⟹ Larger number of informative (heterozygous) families to analyze.
  - ⟹ More statistical power for analysis with smaller sample size!

Problems With «Phase»

- «Phase» (and therefore haplotypes) is usually unknown.
  - Haplotypes have to be reconstructed from empirical data
  - Only the status of each individual marker is known.



- Which haplotype/phase do we have here?
  - Is it **C-A-G-A** and thus **G-T-C-T**?, **G-A-C-A**?, **C-T-G-T**?, or perhaps **G-C-G-C**???

- For *K* bi-allelic markers, there are 2^k possible individual haplotypes.
  - E.g., for SNP1 A>T & SNP2 C>G, we have $2^2 = 4$ haplotypes (A-C, A-G, T-C & T-G)

Practical Problems In Haplotype Analysis

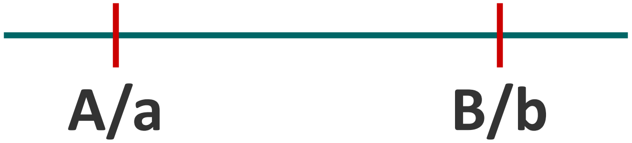
- For *L* number of SNPs:
  - No. of possible haplotypes, $K = 2^L$
  - No. of possible triad combinations: K^4 or 2^{4L}

- What does this mean?
  - Even with only a few SNPs, we end up with a daunting no. of haplotypes ⇒ Clearly impossible to implement in a statistical model.
  - Many cells will have no counts because many of the triad combinations are not even observed in real data!
  - Too many parameters to estimate if model is not simplified ⇒ Extensive time/computer memory usage in calculations.

- Some simplification of the model is still possible
  - Thanks to LD between SNPs, the number of observed haplotypes is substantially fewer than what's theoretically possible.

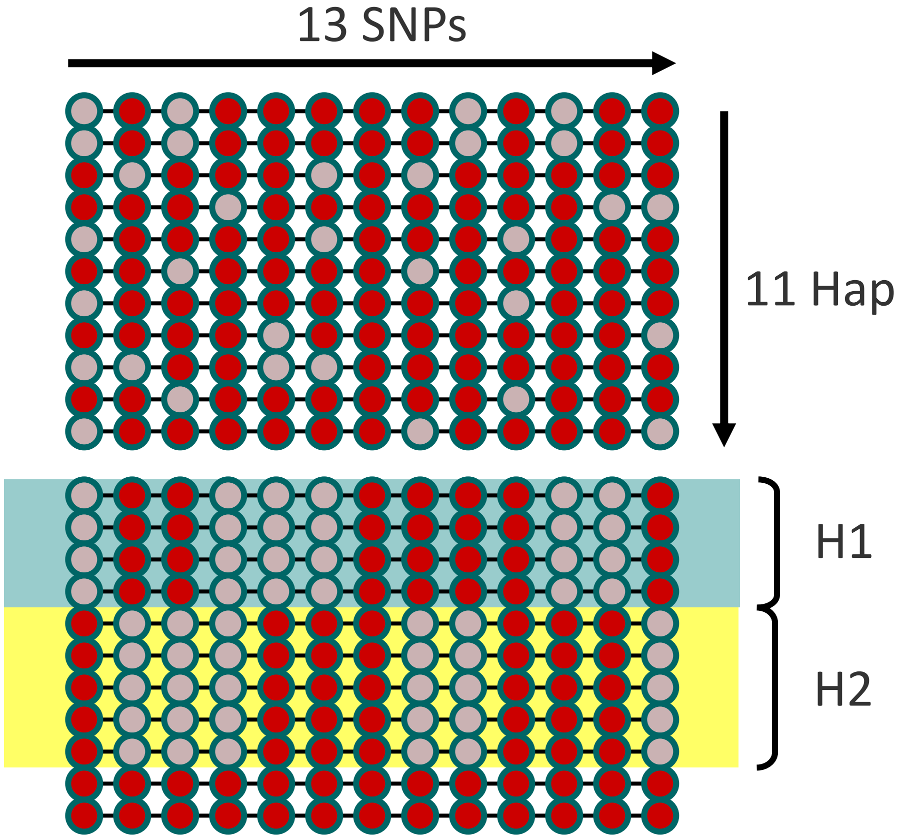
Haplotype Block

With *n* SNPs → 2^n possible haplotypes



2 SNPs = 2^2 haplotypes

AB Ab aB ab

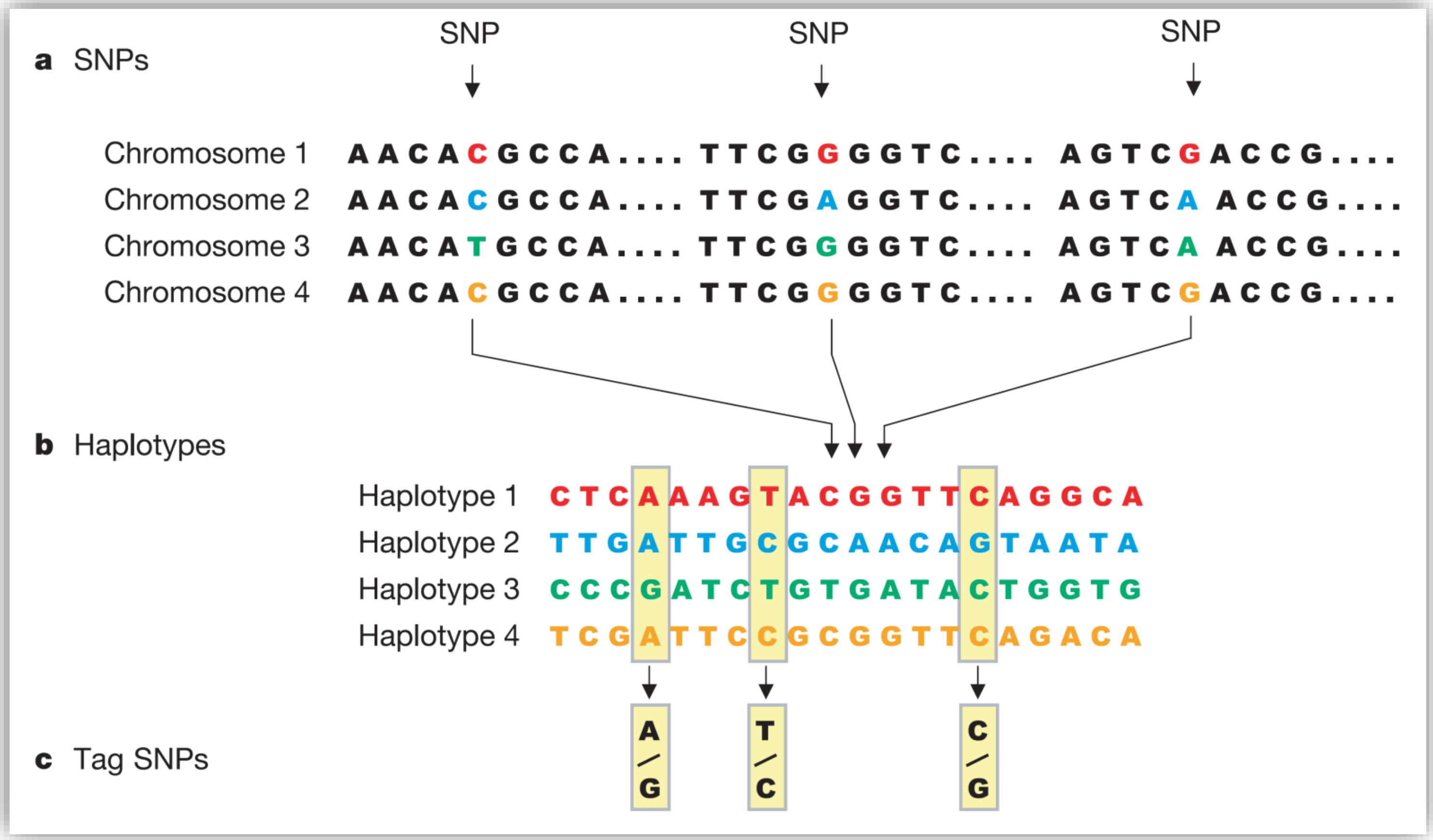


This pattern reflects the **"random assortment of alleles"** at different sites

But most chromosomes will carry one or a few common haplotypes

↓ **"haplotype diversity"**

Concept Of «Haplotype-Tagging SNPs»

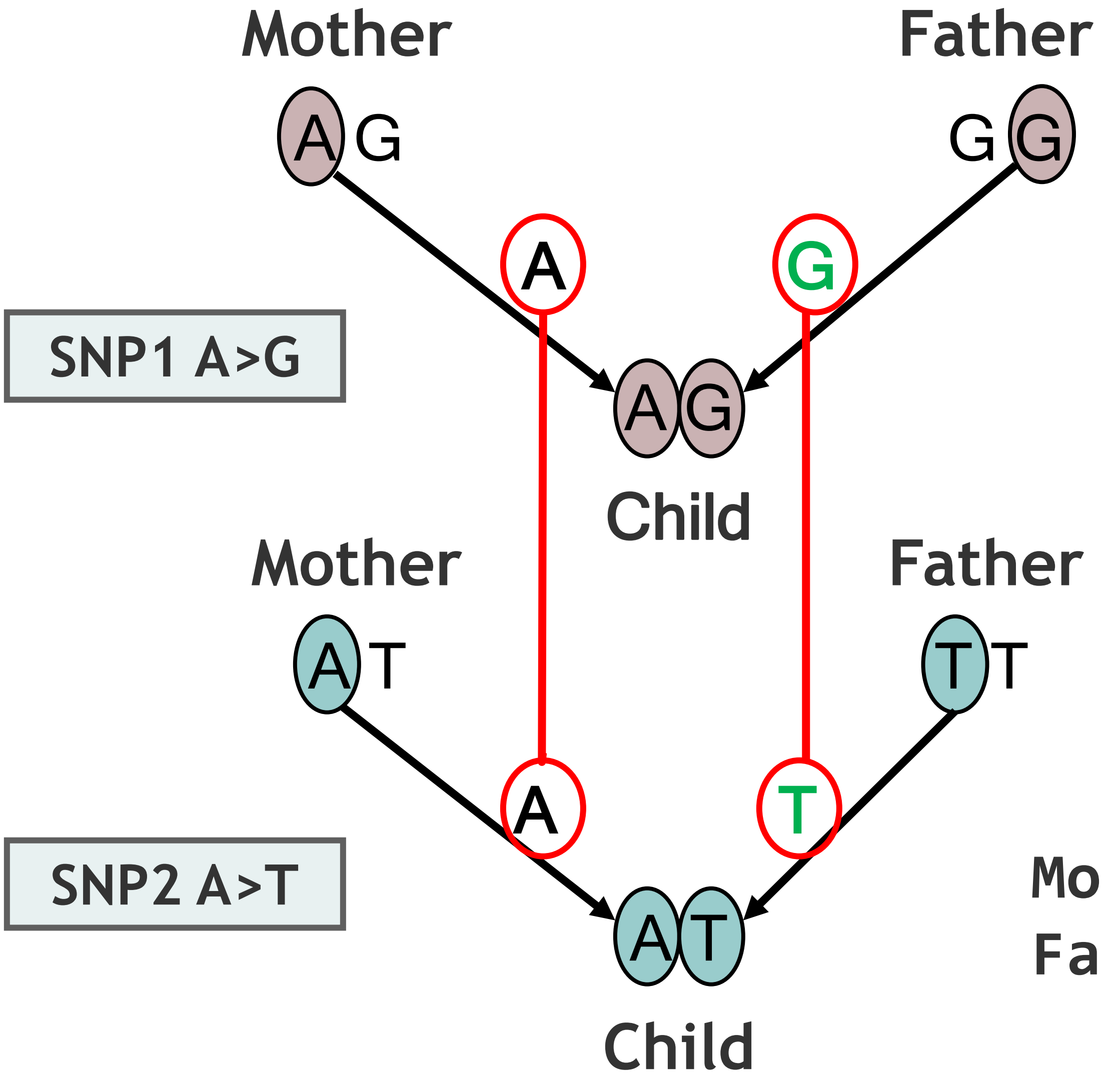

a SNPs
SNP
SNP
SNP
Chromosome 1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
Chromosome 2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
Chromosome 3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
Chromosome 4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
b Haplotypes
Haplotype 1 CTCAAAGTACGGTTCAGGCA
Haplotype 2 TTGATTGCGCAACAGTAATA
Haplotype 3 CCCGATCTGTGATACTGGTG
Haplotype 4 TCGATTCCGCGGTTCAGACA
c Tag SNPs
A/G
T/C
C/G

- ❑ Genotyping just these 3 htSNPs out of the 20 SNPs identifies all 4 haplotypes.

  ⇒ If a chromosome has the pattern A-T-C at these 3 tags, this matches the pattern determined for haplotype 1, etc.

Source: The International HapMap Consortium. NATURE | Vol 426 | 18/25 December 2003

Triads are Useful for Haplotype Inference



Mother

A G

Father

G G

A

G

SNP1 A>G

A G

Child

Mother

A T

Father

T T

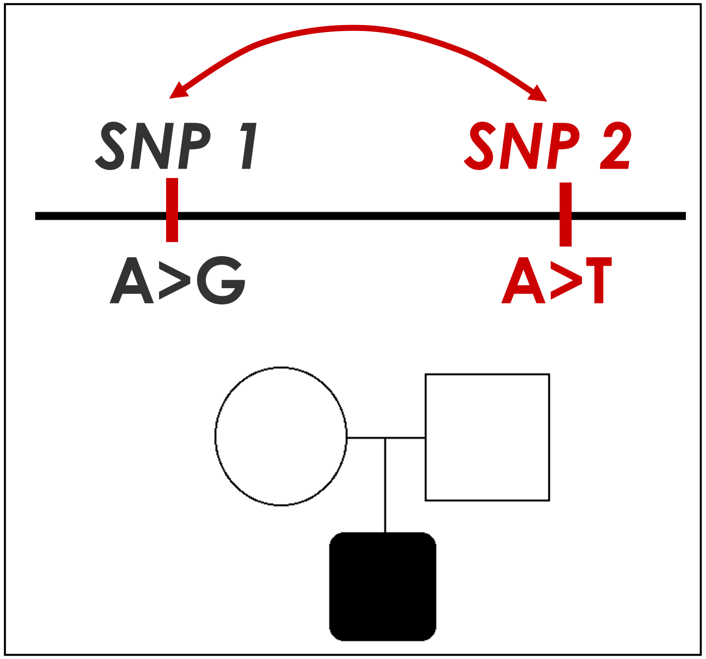
A

T

SNP2 A>T

A T

Child



SNP 1

SNP 2

A>G

A>T

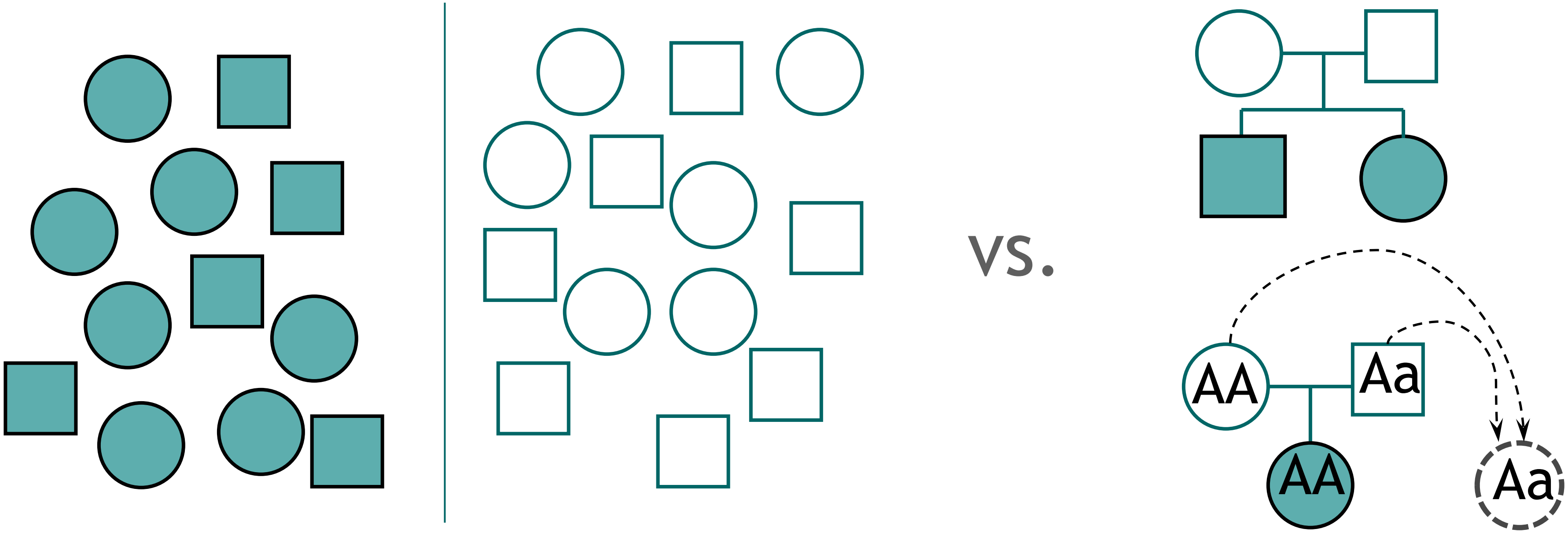
Mother: -A-A- haplotype

Father: -G-T- haplotype

Genetic Association Studies

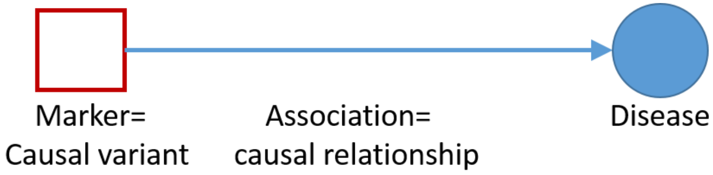
- With case-control data: Compare marker allele frequencies between an unrelated case and control population

- With nuclear family data: Use the non-transmitted parental alleles as control alleles.
    - Test for deviations from the expected 50% Mendelian transmission of an allele from parents to offspring.

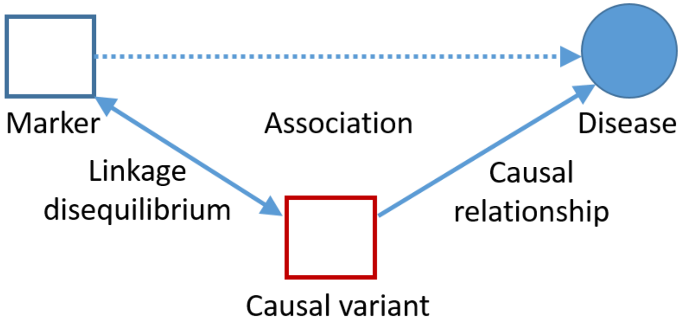


Reasons For An Observed Genetic Association

- The marker itself is a functional variant (i.e., the association is causal):
  - Marker/ Causal variant → Disease



- The marker is in LD with a causal variant:
  - Marker ⇔ Causal variant → Disease

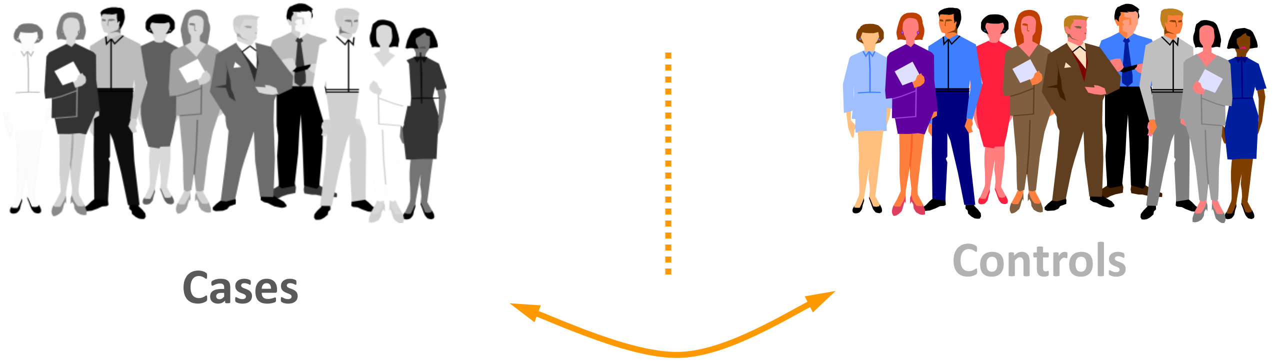


- The association is due to confounding by population stratification.
  - Marker ⇔ Population stratification ⇔ Causal variant → Disease

Source: http://www-gene.cimr.cam.ac.uk/clayton/

POPULATION STRATIFICATION

- For population stratification to affect a genetic association study, two conditions must be met :
  - Both disease prevalence and allele frequency differences must exist between cases and controls.



- Consequence? ⇒ “Spurious” association
  - Differences in allele frequency between cases and controls will be due to systematic differences in other factors (e.g., ancestry) rather than a genuine association of the allele with disease.

Source: Cardon & Palmer 2003. *Lancet* Vol 361; Feb 15; 598-604

How To Deal With Stratification Effects?

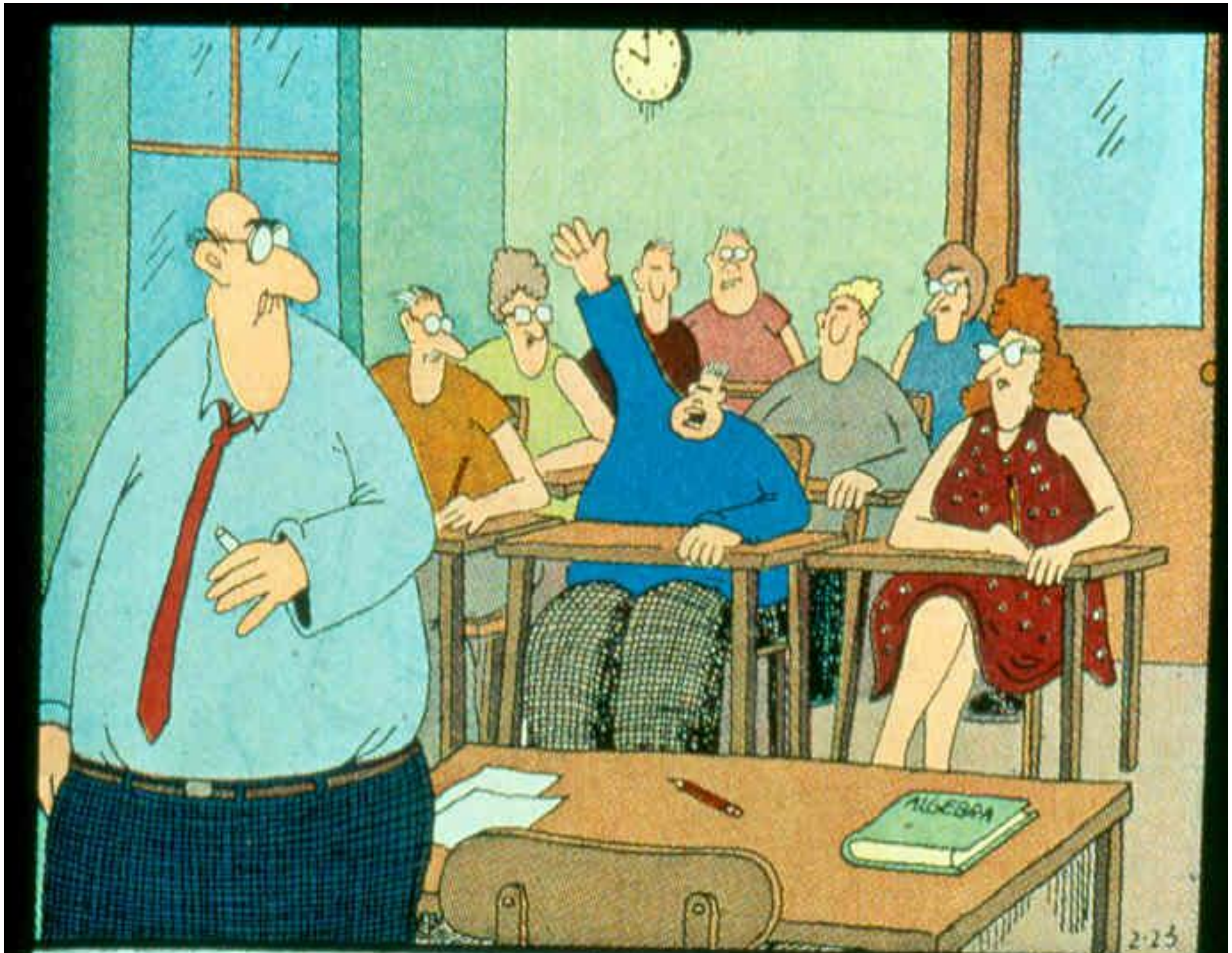
- Carefully match cases and controls by e.g. ancestry and geographic origin.
- Use alternative study designs, such as family-based designs.
- Population stratification often reflected in substantial deviations in HWE.
    - Genotype a few unlinked genetic markers to see whether there are substantial deviations from HWE.
- Use "genomic controls" to control for ancestry.
- Use PCA analysis to identify ethnic outliers.
- Use the software ***STRUCTURE*** to identify individuals with different ancestries and use this information to adjust ancestry as a covariate in the association analysis (***fastSTRUCTURE*** **for large SNP datasets**).
    - Basis for "Admixture Mapping"
        - If population stratification can be measured through structure assessment, test for association **within** strata.

[†] **Reference:** **STRUCTURE:** JK. Pritchard *et al.* (2000). *Genetics* **155**, 945-959; https://web.stanford.edu/group/pritchardlab/structure.html

Overview

- DNA, Exome, 1000 GP
- Use of SNPs as genetic markers
- Linkage disequilibrium and haplotypes
- Population stratification

Questions?



"Mr. Osborne, may I be excused? My brain is full."