

Genetic Epidemiology

Ioanna Tzoulaki itzoulaki@bioacademy.gr

Director of Research, Biomedical Research Institute Academy of Athens

**Professor in Chronic Disease Epidemiology, School of Public Health,
Imperial College London**

Course contents

Genetic
Epidemiology

GWAS, gene x
environment,
GRS

Omics,
exposome

Mendelian
randomization

Readings

- Palmer LJ, Burton P, Davey Smith G. An Introduction to Genetic Epidemiology. Health & Society Series, 2011.
 - Consists of series of papers on genetic epidemiology that first appeared in Lancet in 2005
- Claussnitzer, M., Cho, J.H., Collins, R. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020). <https://doi.org/10.1038/s41586-019-1879-7>
- Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet.* 2017 Jul 6;101(1):5-22. doi: 10.1016/j.ajhg.2017.06.005. PMID: 28686856; PMCID: PMC5501872.
- Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: Realizing the promise. *Am J Hum Genet.* 2023 Feb 2;110(2):179-194. doi: 10.1016/j.ajhg.2022.12.011. Epub 2023 Jan 11. PMID: 36634672; PMCID: PMC9943775.

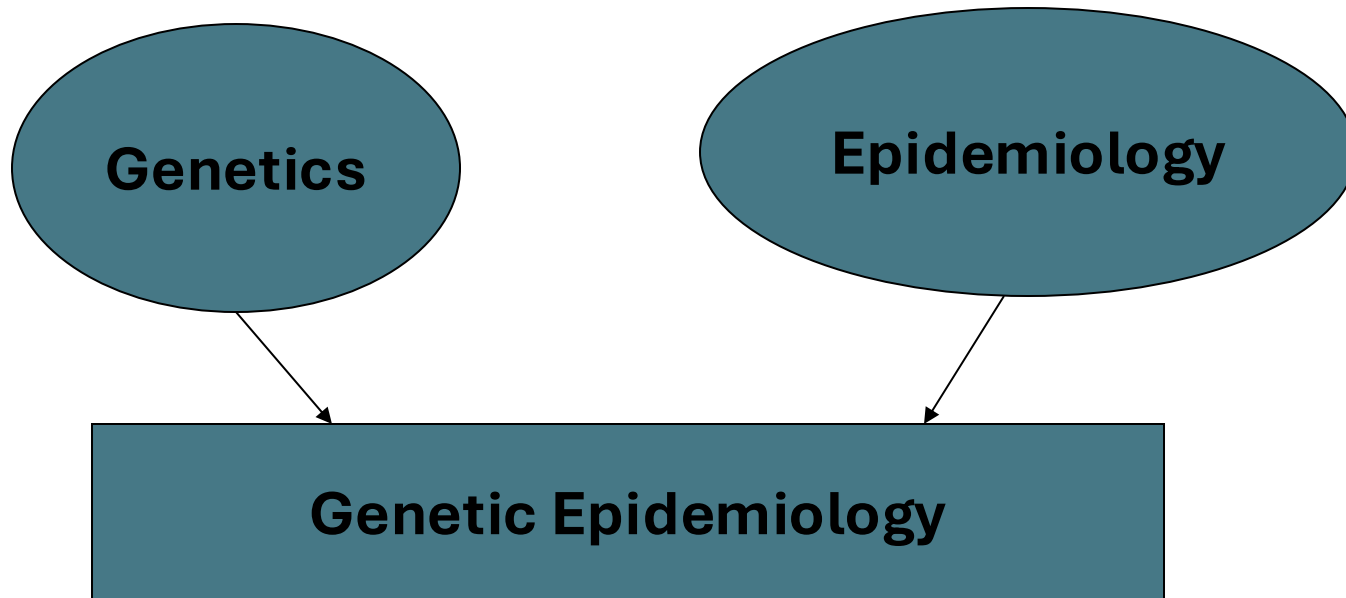
Introduction to Genetic Epidemiology

Outline

- Genetic Epidemiology
- Genome structure and genetic variation
- Genetic and epidemiological study designs
 - Estimation of genetic effects
- Key concepts
 - Heritability
 - Linkage disequilibrium

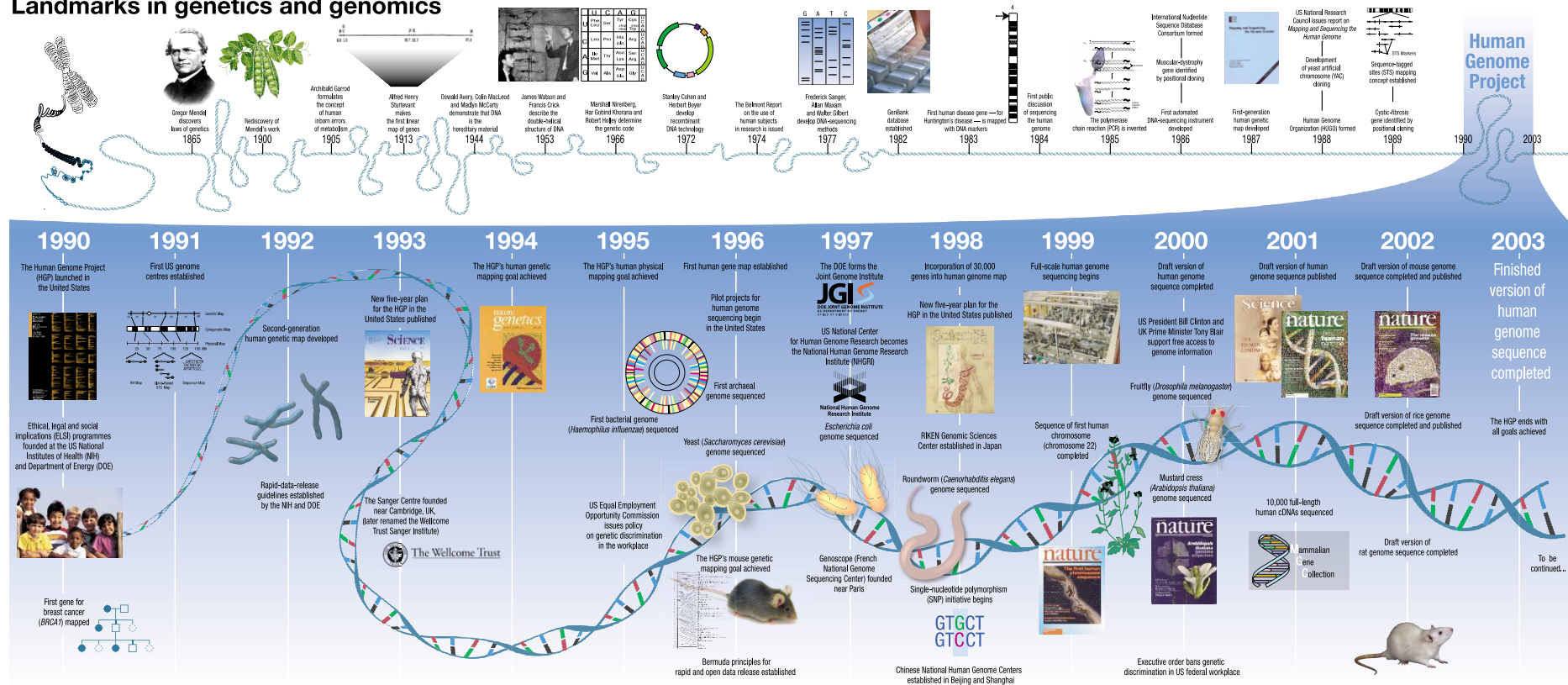
Genetic Epidemiology

A hybrid science focusing on *complex* diseases (where both genetic & environmental factors contribute to etiology of disease)

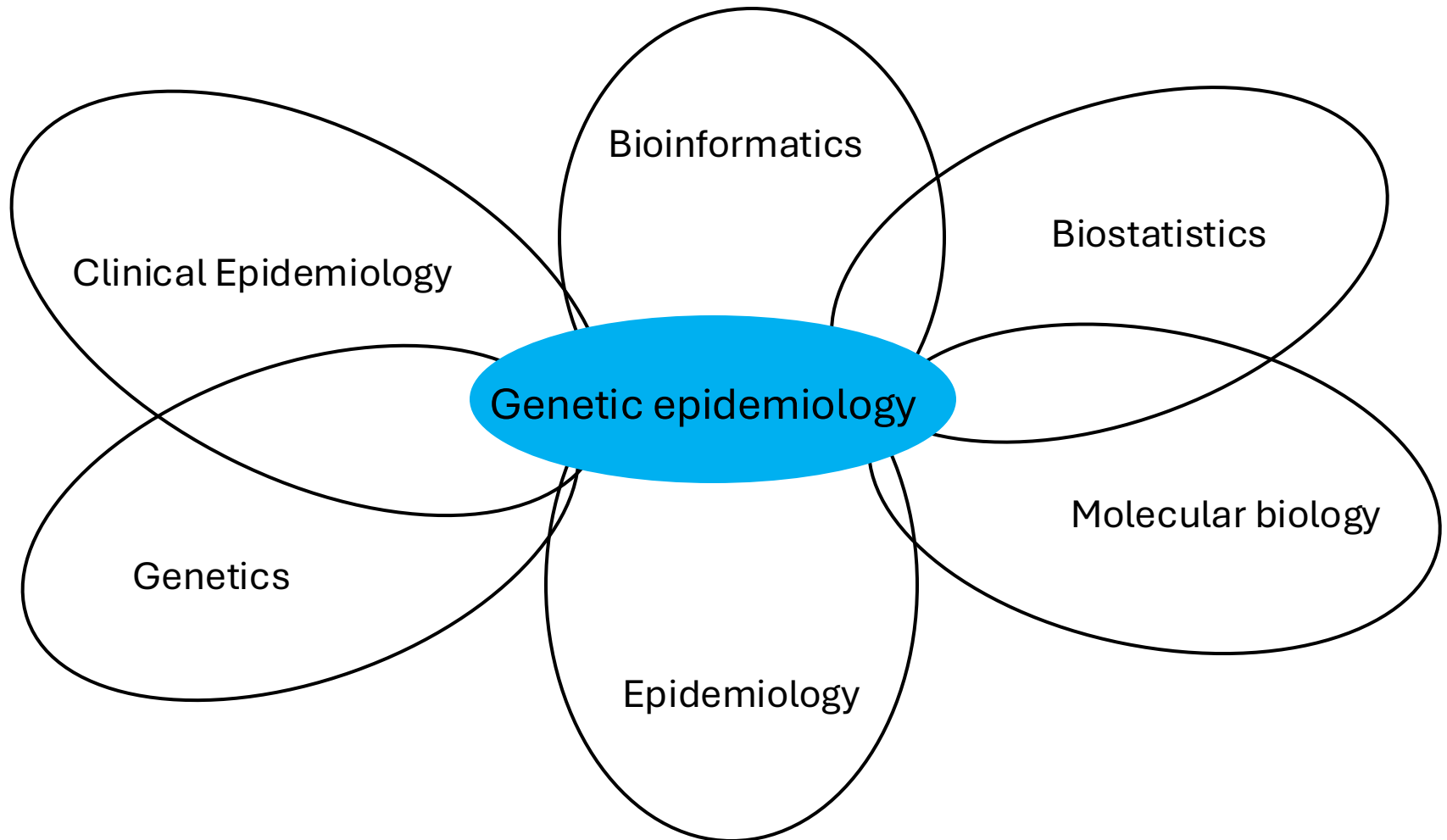


Parent sciences (genetics & epidemiology) share common goals but they differ in their histories & perspectives.

Landmarks in genetics and genomics



Genetic Epidemiology



Genetic Epidemiology

- “A field of science that focuses on the role of genetic factors and their interaction with environmental factors in the occurrence of disease in human populations”

Genetic Epidemiology

A science that deals with etiology, distribution and control of disease in families and with inherited causes of diseases in populations

N Morton

Genetic Epidemiology

- Is based on principles of population genetics
- Utilizes statistical approaches to detect the genetic effects on susceptibility to chronic diseases and quantitative traits
 - Type 2 Diabetes
 - Prostate cancer
 - Obesity or quantitative trait, e.g. BMI

Central questions in Genetic Epidemiology

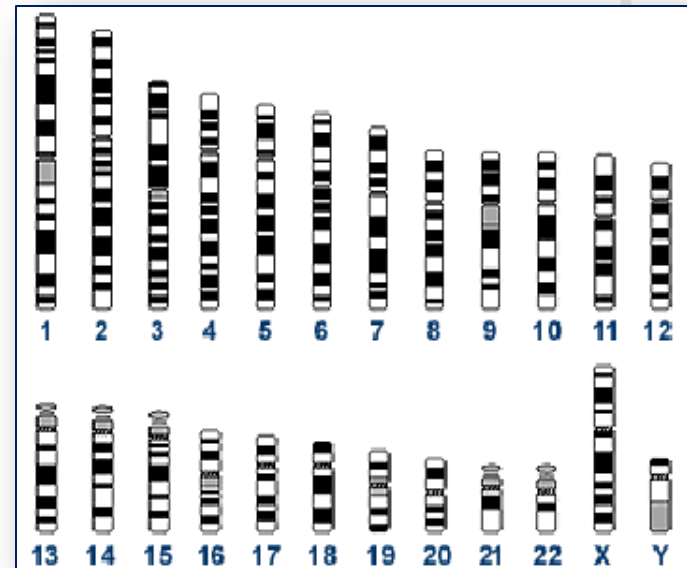
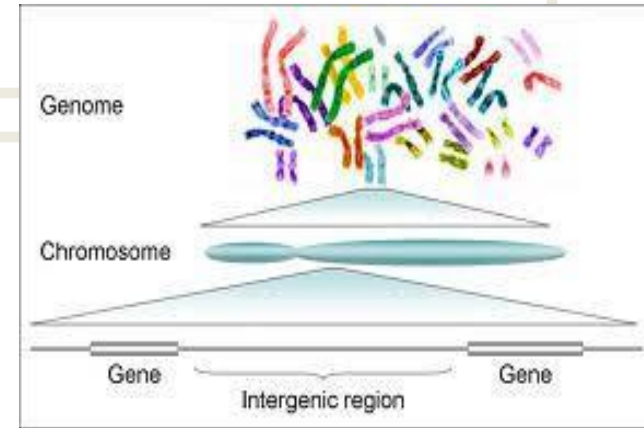
1. Does the trait cluster in families?
2. Can familial clustering be explained by genes or shared environment?
3. What is the best model of inheritance?
4. Can we locate genes for complex diseases/traits?
5. How does the gene control risk of disease?

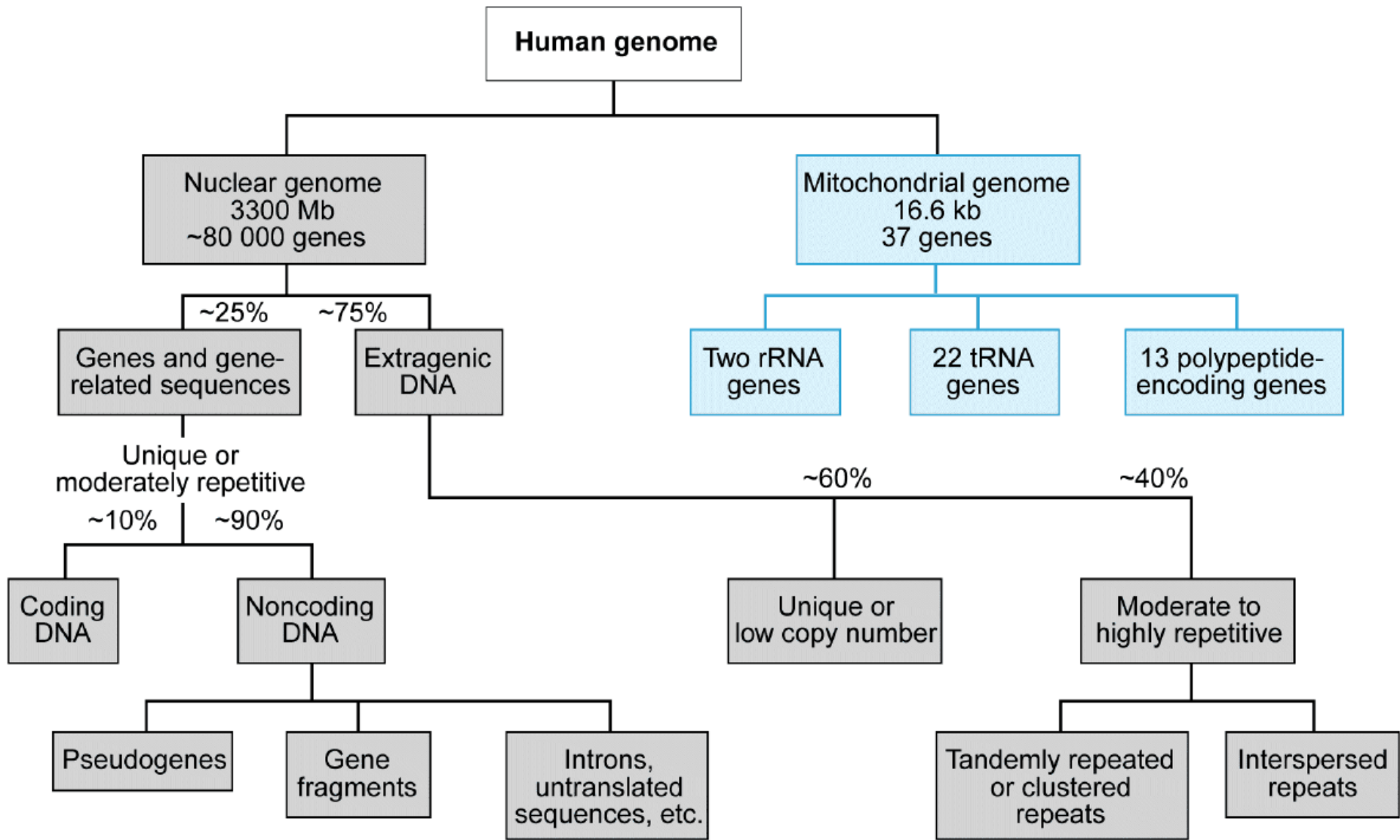
Genome organisation

- The human genome consists of all the DNA present in the cell.
- It can be divided into the nuclear genome (about 3200 Mbp) and the mitochondrial genome (16.6 kb).

Organization of the human genome

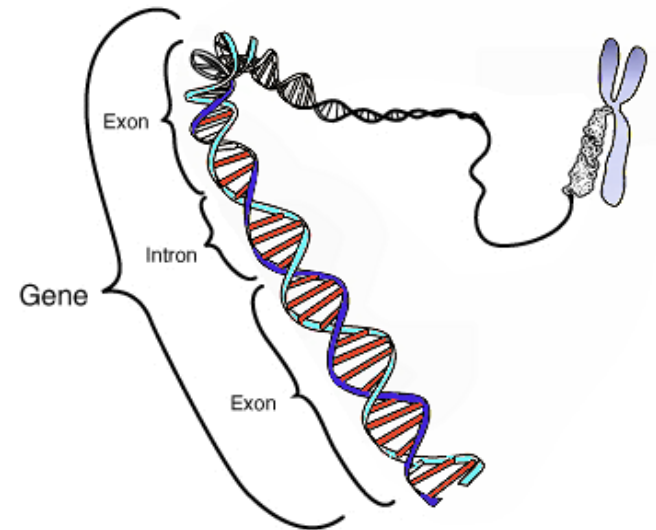
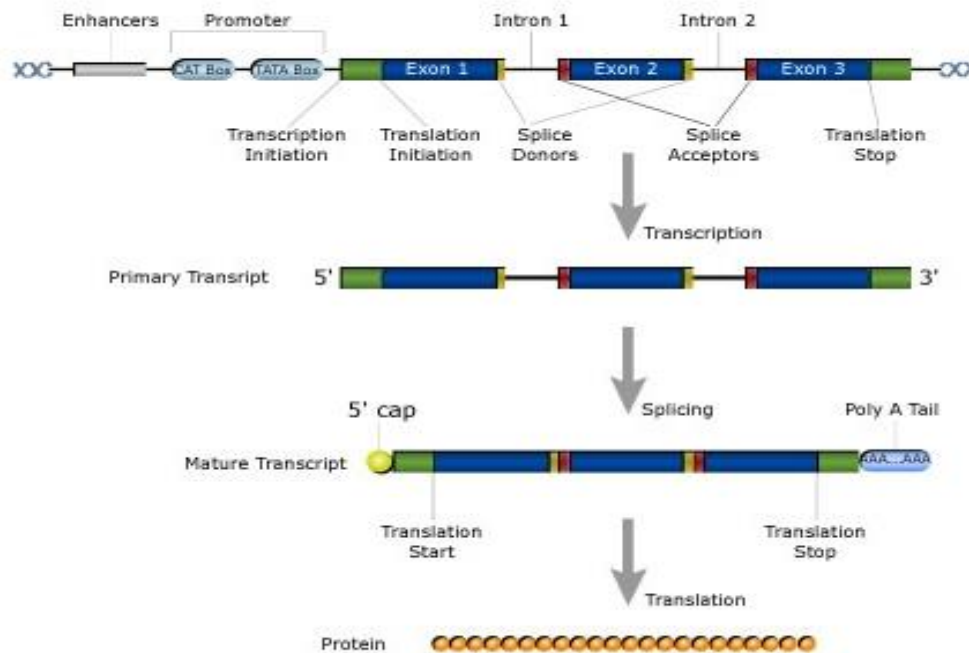
- Nuclear genome
 - 3200 Mb
 - 23 (XX) or 24 (XY) linear chromosomes
 - ~20,000 protein-coding genes
 - 1 gene/30-60kb
 - Only 10% is coding sequence
 - Introns
 - 3% coding
 - Repetitive DNA sequences (45%)
 - Recombination
 - Mendelian inheritance (X + auto, paternal Y)



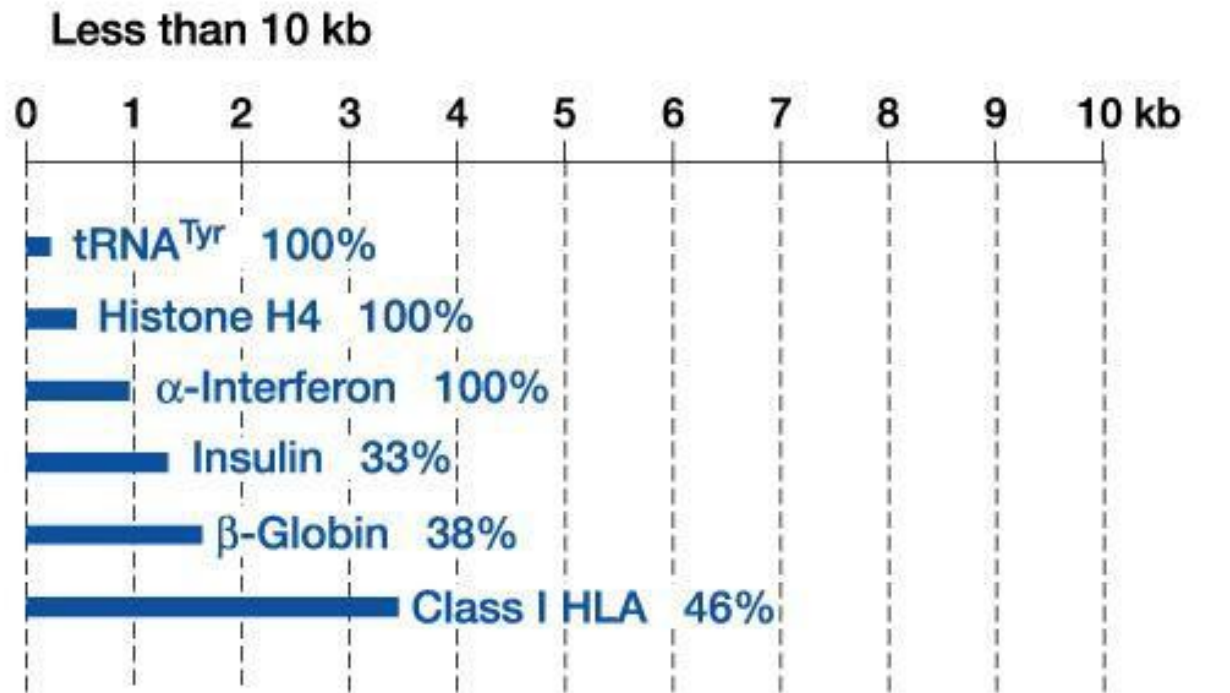


Organization of the human genome

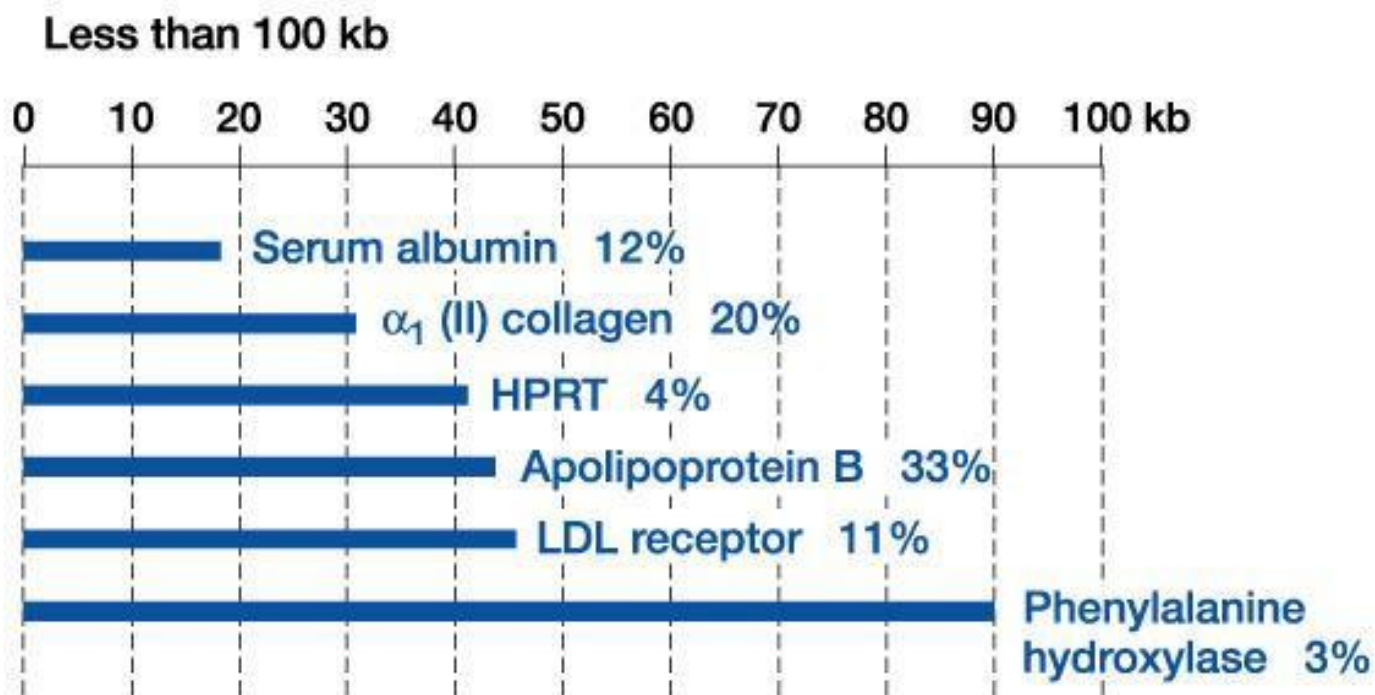
- Genes vary in size and exon content



(A)

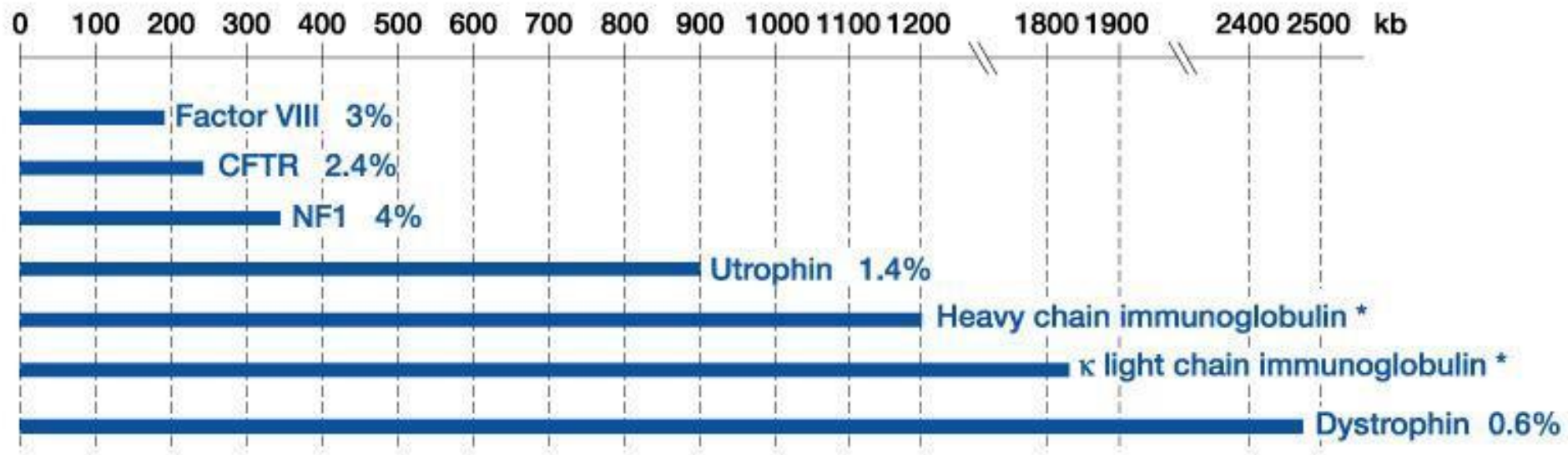


(B)



(C)

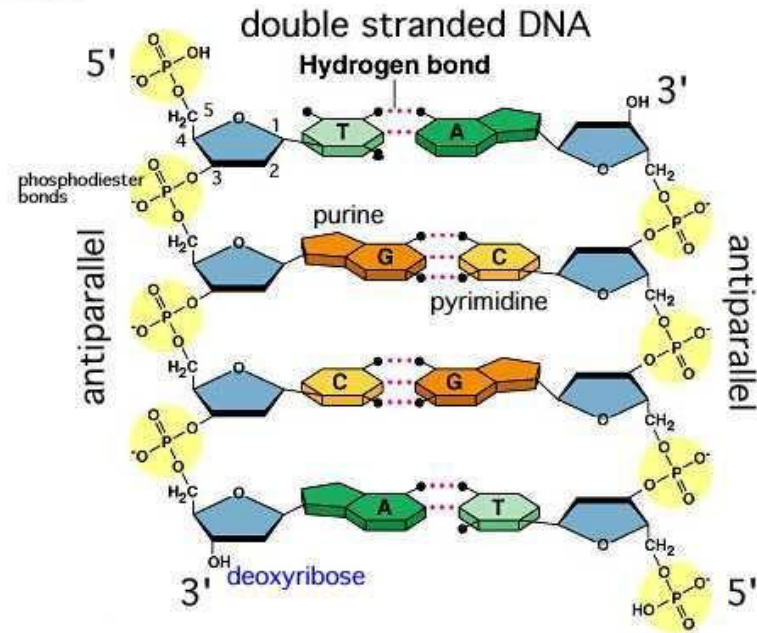
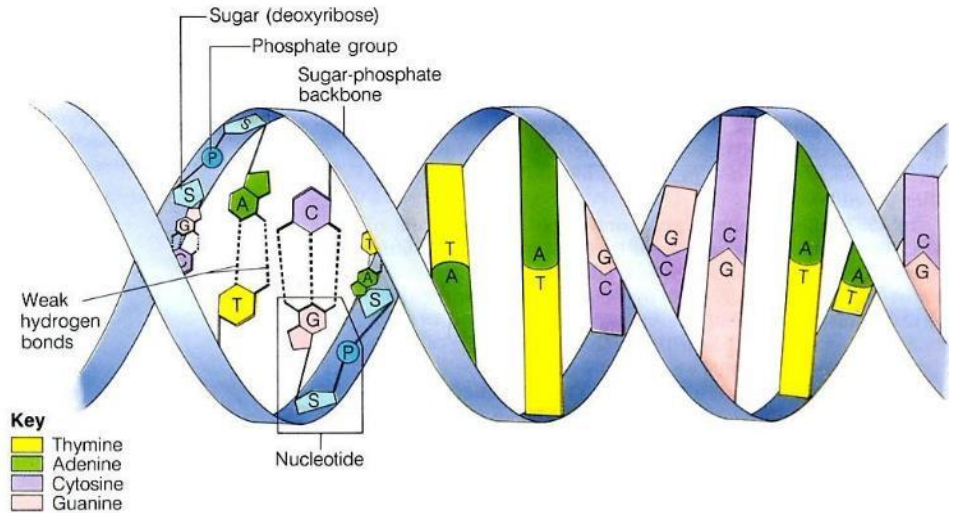
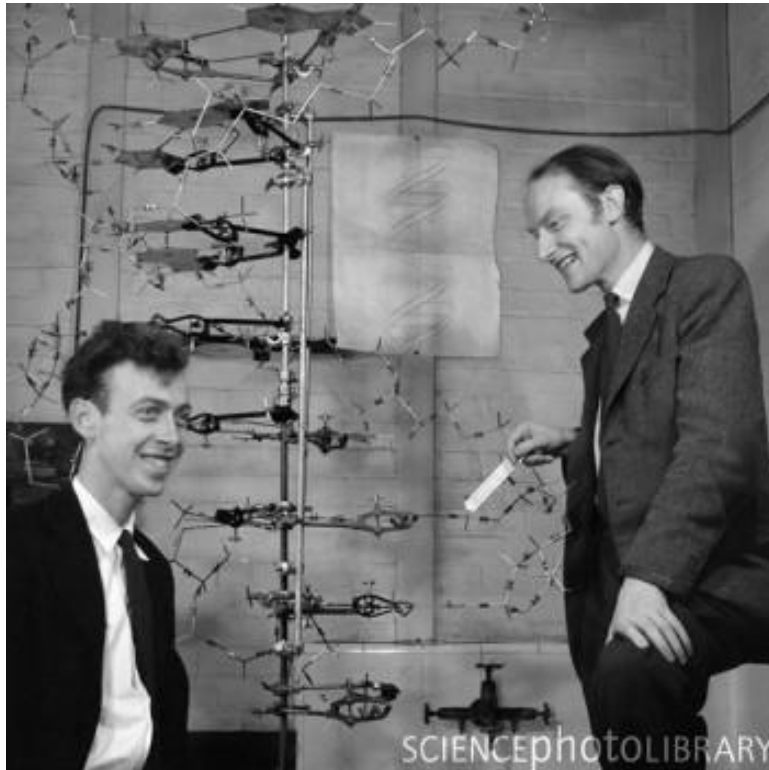
More than 100 kb



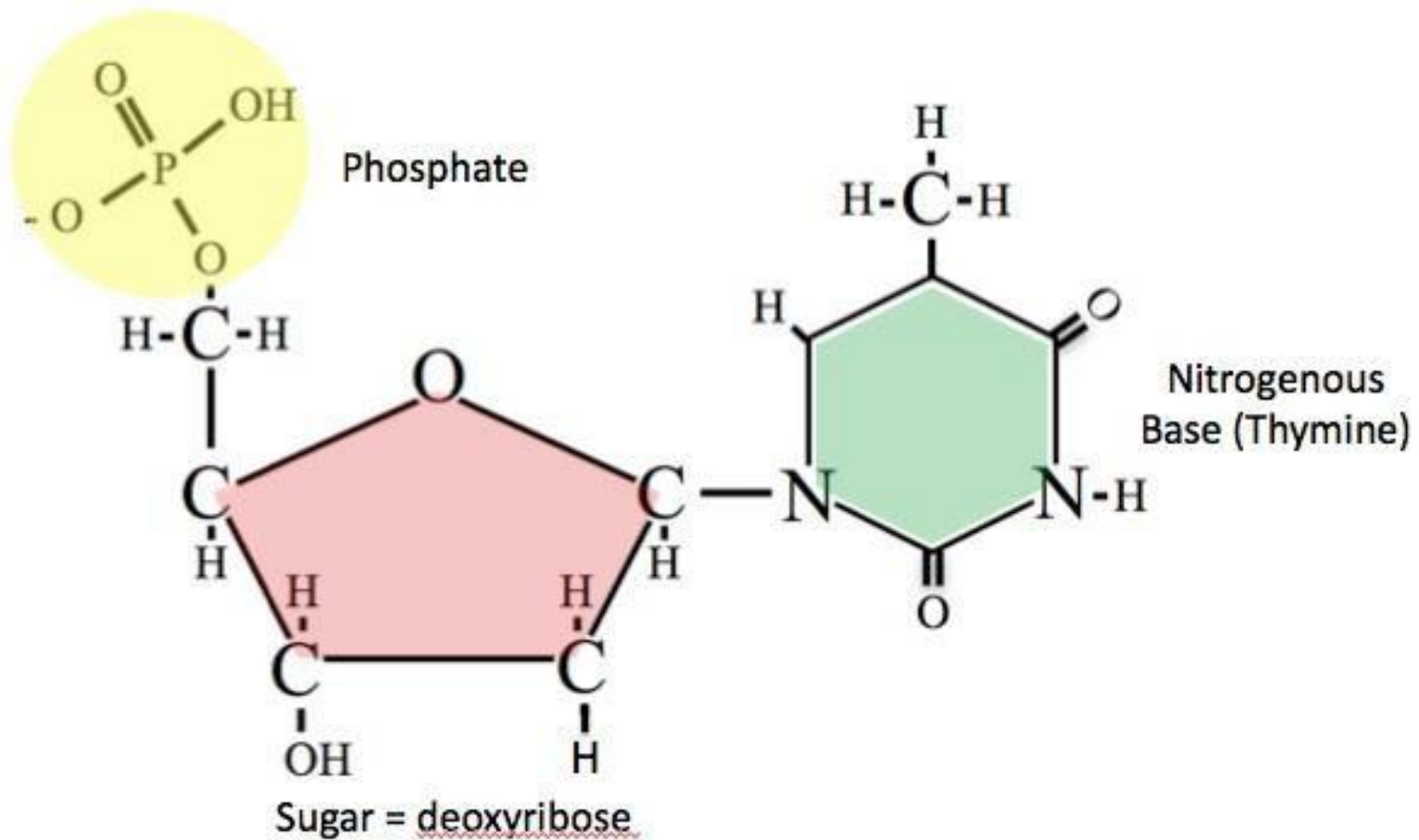
Human Genetic variation and disease

- What is a SNP?
- Types of SNP
- SNPs as genetic markers
- GWAS

Watson and Crick



Nucleotide



SNPs

- Single/simple nucleotide polymorphism – SNP
- A single nucleotide variant in the DNA
- SNP is a DNA sequence variation within a single nucleotide— A, T, C or G — in the genome
- E.g. Adenine to Guanine, Thymine to Cytosine
- Mostly biallelic (two alleles) polymorphism,
 - but large number of tri- and quadri-allelic SNPs is now described
- Could also be a 1bp indel, duplication, etc

SNP or mutation?

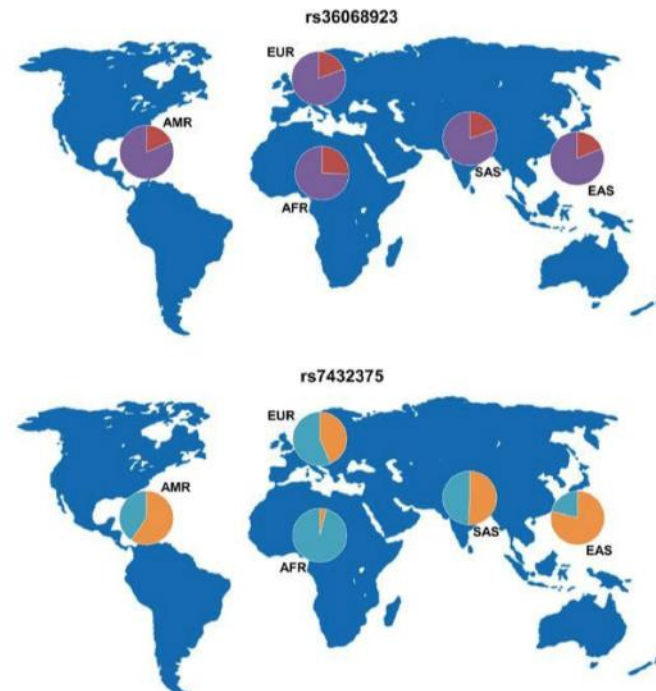
- Typically not considered to have a functional effect, hence polymorphism and not mutation
- However, SNP is often used as a term for all single-base changes, functional or not.
- The difference is that a mutation has a functional effect and a polymorphism does not necessarily

Minor Allele Frequency (MAF)

- This is how often the less frequent allele of a biallelic variant occurs in a group (often a percentage)
- As the total allele frequency is 1 (100%), a MAF must always be less than 0.5 (50%), otherwise it would be a major allele
- E.g. if we genotype a variant (A/G) in 1000 people
 - 550 are (A,A), 400 are (A,G) and 50 are (G,G)
 - There are 2000 alleles in total
 - The G allele is less common, accounting for 500 alleles
 - Therefore, the MAF is $500/2000 = 0.25$ or 25%

Minor allele frequency

- Rare variants: $MAF < 1\%$
- Low-frequency variants: $MAF 1-5\%$
- Common variants: $MAF \geq 5\%$



MAF is population specific *Translational Psychiatry volume 7, page e988 (2017)*

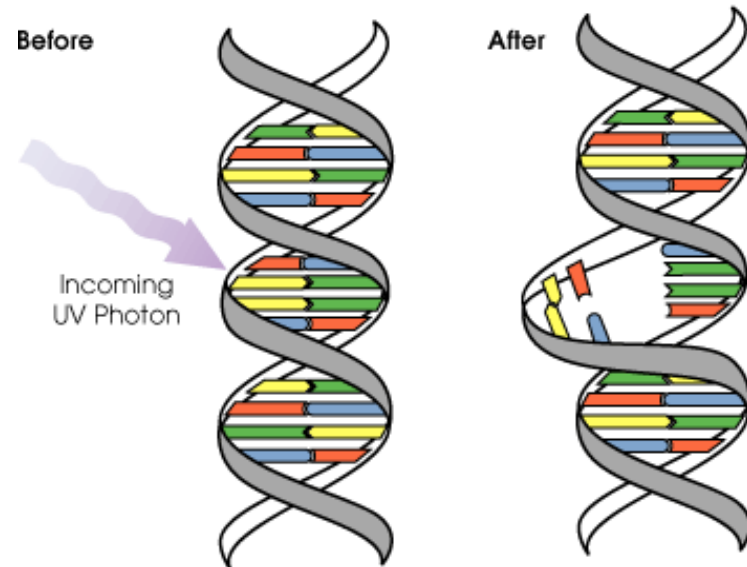
Causes of SNPs

External

- Ionising radiation
- Ultraviolet light
- Environmental chemicals

Internal

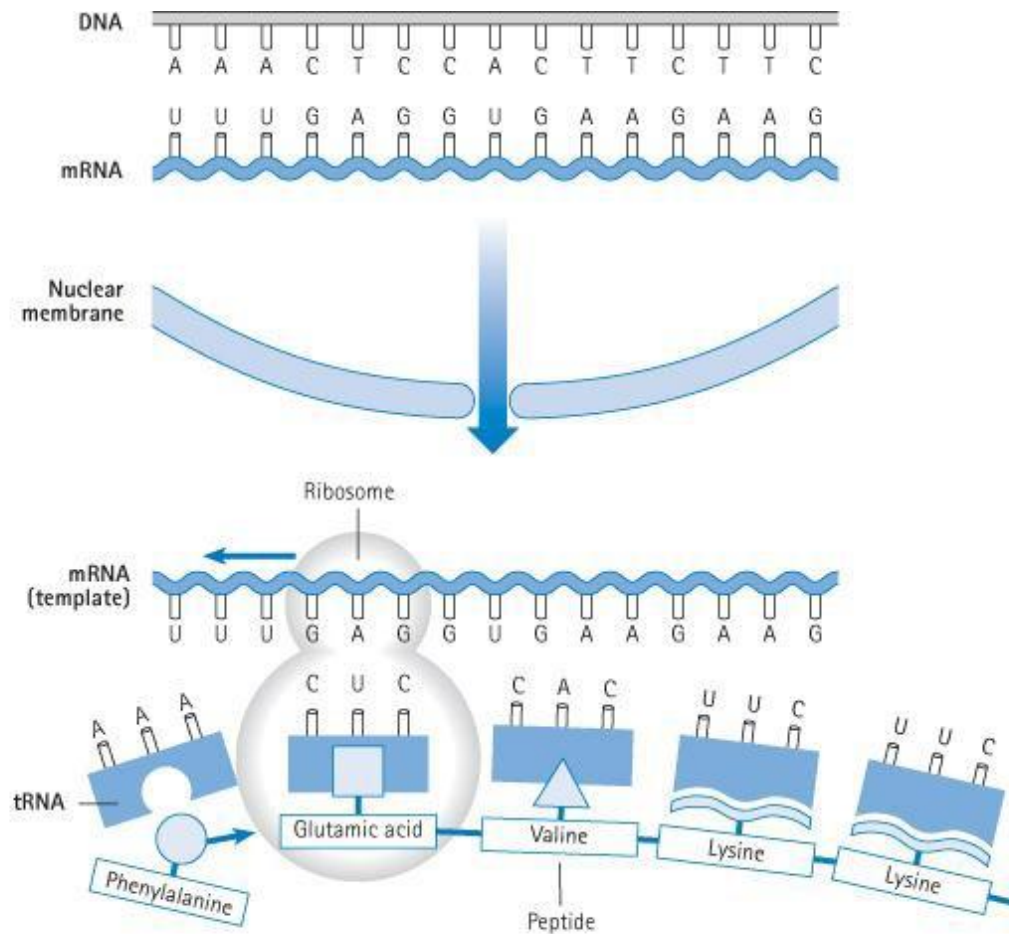
- Spontaneous depurination
- Spontaneous deamination
- DNA replication/repair errors
- DNA replication/recombination errors



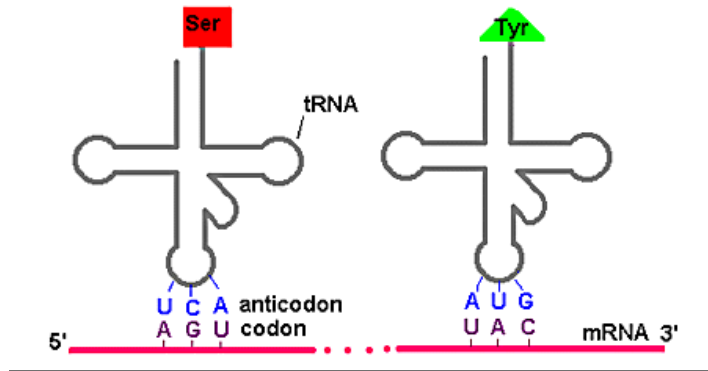
Coding for a protein

- Coding regions in a gene are transcribed to RNA,
- spliced to mRNA and then translated into protein
- Each amino acid is encoded by three bases in the DNA, known as a codon
- Some amino acids are encoded by >1 codon
- (CAT/CAC – His)
- SNPs alter the codon but not always the amino acid
- InDels may or may not change the reading frame

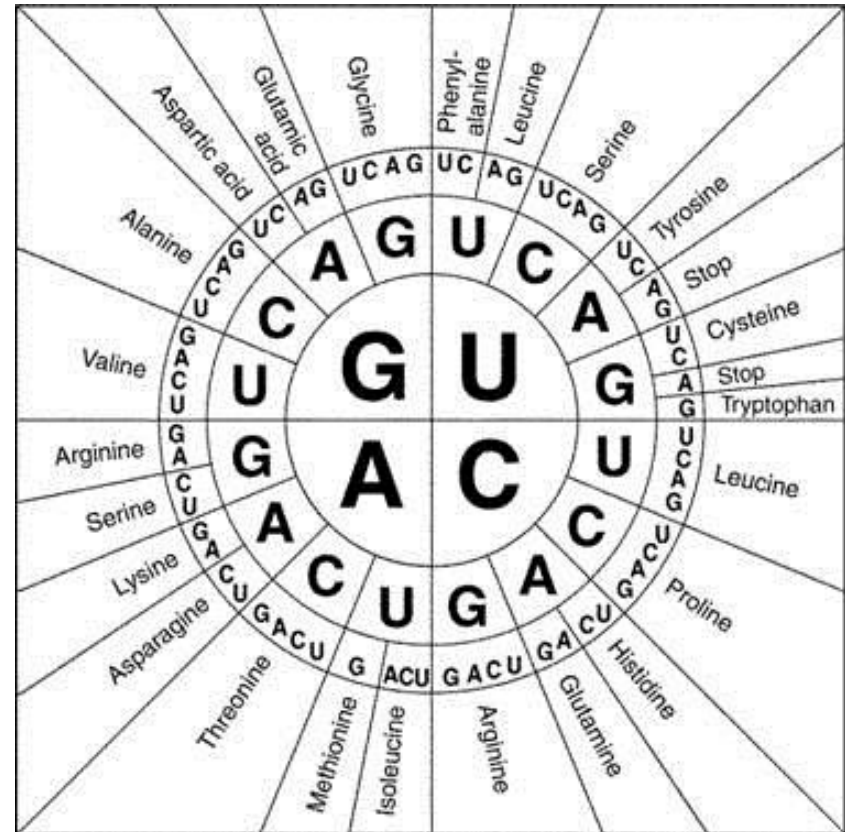
DNA-RNA-Protein



Codon Tables



2nd base in codon						
	U	C	A	G		
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	3rd base in codon
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	
		U C A G	U C A G	U C A G	U C A G	



The Genetic Code

Types of SNP

- Non-Synonymous
- Synonymous
- Promoter
- Terminator
- Splicing
- Neutral

Non-Synonymous SNPs

- Missense
 - – These change an amino acid codon to another amino acid.
- Nonsense
 - – These change an amino acid codon into a stop codon, causing premature truncation of the protein.
- e.g.
 - ...CAG...to ...TAG...
Gln to Stop
- e.g. Q39X in beta-globin causes β -thalassaemia

Synonymous SNPs

- These change the codon but due to codon degeneracy they have no effect on the amino acid in the protein
- e.g. ...TTG... changed to ...CTG...
- Both are Leucine codons and the protein is unaffected

Promoter SNPs

Changes in the gene promoter may alter the level of gene expression.

e.g.TATAAA... toTAGAAA....

Would remove the basal TATA box and could drastically reduce or abolish transcription of the gene

Terminator SNPs

- These could affect the correct termination and
 - polyadenylation of the messenger RNA
- e.g. ...AATAAA... to ...AATAGA...
- In alpha-globin this change disrupts the polyadenylation signal making the mRNA unstable

Splicing SNPs


These lead to the creation or deletion of splice donor, acceptor or branch sites, affecting the final mRNA and hence protein.

e.g. ...CAGGTAAGT... to ...CAGATAAGT...

Removes the underlined splice donor site leading to utilisation of a different or cryptic splice site.



SNPs as Genetic Markers

- Bi-allelic
 - Very common across the genome
 - Good evidence that they can directly cause disease
 - Easily genotyped using high-throughput technologies
 - Widely used for association and linkage studies
- 

PubMed Gene

<http://www.ncbi.nlm.nih.gov/pubmed/gene/>

➤ Database of genes in humans (n=43,828) and other organisms

- gene name
- alias (e.g. *TCF7L2*)
- function
- lineage in other organisms
- biological pathways

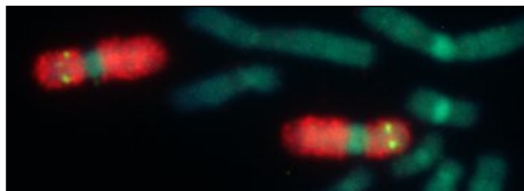
PubMed Gene

<http://www.ncbi.nlm.nih.gov/pubmed/gene/>

 **NCBI** [Resources](#) [How To](#) [Sign in to NCBI](#)

Gene

[Limits](#) [Advanced](#) [Help](#)



Welcome to Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

Using Gene

[Gene Quick Start](#)

[FAQ](#)

[Download/FTP](#)

[RefSeq Mailing List](#)

[Gene News](#) 

Gene Tools

[Submit GeneRIFs](#)

[Submit Correction](#)

[Statistics](#)

[BLAST](#)

[Genome Workbench](#)

[Splign](#)

Other Resources

[HomoloGene](#)

[OMIM](#)

[RefSeq](#)

[RefSeqGene](#)

[UniGene](#)

[Protein Clusters](#)

You are here: [NCBI](#) > [Genes & Expression](#) > [Gene](#)

[Write to the Help Desk](#)

GETTING STARTED

[NCBI Education](#)

[NCBI Help Manual](#)

[NCBI Handbook](#)

[Training & Tutorials](#)

RESOURCES

[Chemicals & Bioassays](#)

[Data & Software](#)

[DNA & RNA](#)

[Proteins & Structures](#)

POPULAR

[PubMed](#)

[Nucleotide](#)

[BLAST](#)

[PubMed Central](#)

FEATURED

[Genetic Testing Registry](#)

[PubMed Health](#)

[GenBank](#)

[Reference Sequences](#)

NCBI INFORMATION

[About NCBI](#)


[Research at NCBI](#)

[NCBI Newsletter](#)

[NCBI FTP Site](#)

PubMed Gene

<http://www.ncbi.nlm.nih.gov/pubmed/gene/>

 NCBI [Resources](#) [How To](#) [Sign in to NCBI](#)

Gene [Limits](#) [Advanced](#) [Help](#)

[Display Settings:](#) ☒ Full Report [Send to:](#)

TCF7L2 transcription factor 7-like 2 (T-cell specific, HMG-box) [*Homo sapiens*]
Gene ID: 6934, updated on 13-Nov-2012

Official Symbol	TCF7L2 <small>provided by HGNC</small>
Official Full Name	transcription factor 7-like 2 (T-cell specific, HMG-box) <small>provided by HGNC</small>
Primary source	HGNC:11641
Locus tag	RP11-357H24.1
See related	Ensembl:ENSG00000148737 ; HPRD:03751 ; MIM:602228 ; Vega:OTTHUMG00000019070
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	TCF4; TCF-4
Summary	This gene encodes a high mobility group (HMG) box-containing transcription factor that plays a key role in the Wnt signaling pathway. The protein has been implicated in blood glucose homeostasis. Genetic variants of this gene are associated with increased risk of type 2 diabetes. Several transcript variants encoding multiple different isoforms have been found for this gene.[provided by RefSeq, Oct 2010]
Annotation information	Note: TCF4 (GeneID: 6925) and TCF7L2 (GeneID: 6934) loci share the TCF4 symbol/alias in common. TCF4 is a widely used alternative name for T-cell-specific transcription factor 4 (TCF7L2) conflicting with the official symbol for transcription factor 4 (TCF4). [26 Jun 2008]

Table of contents

[Summary](#)

[Genomic context](#)

[Genomic regions, transcripts, and products](#)

[Bibliography](#)

[Phenotypes](#)

[HIV-1 protein interactions](#)

[Interactions](#)

[General gene info](#)

[General protein info](#)

[Reference sequences](#)

[Related sequences](#)

[Additional links](#)

Related information

[Order cDNA clone](#)

[3D structures](#)

[BioAssay](#)

[BioProjects](#)

[BioSystems](#)

[CCDS](#)

(db)SNP

<http://www.ncbi.nlm.nih.gov/snp>

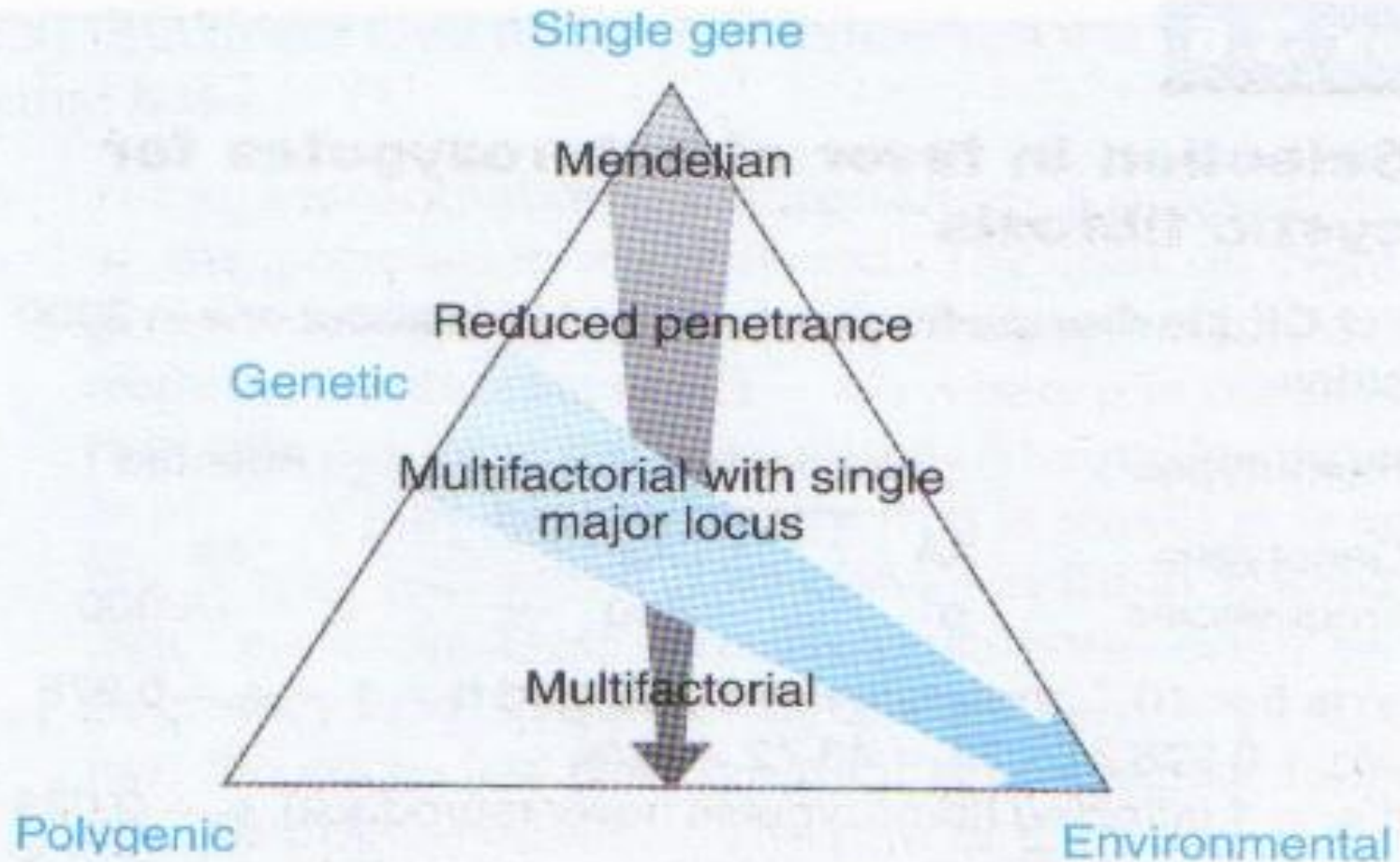
➤ More information on genetic variation

- genotype
- allele frequencies
- chromosome position
- sequence
- population diversity
- visual displays

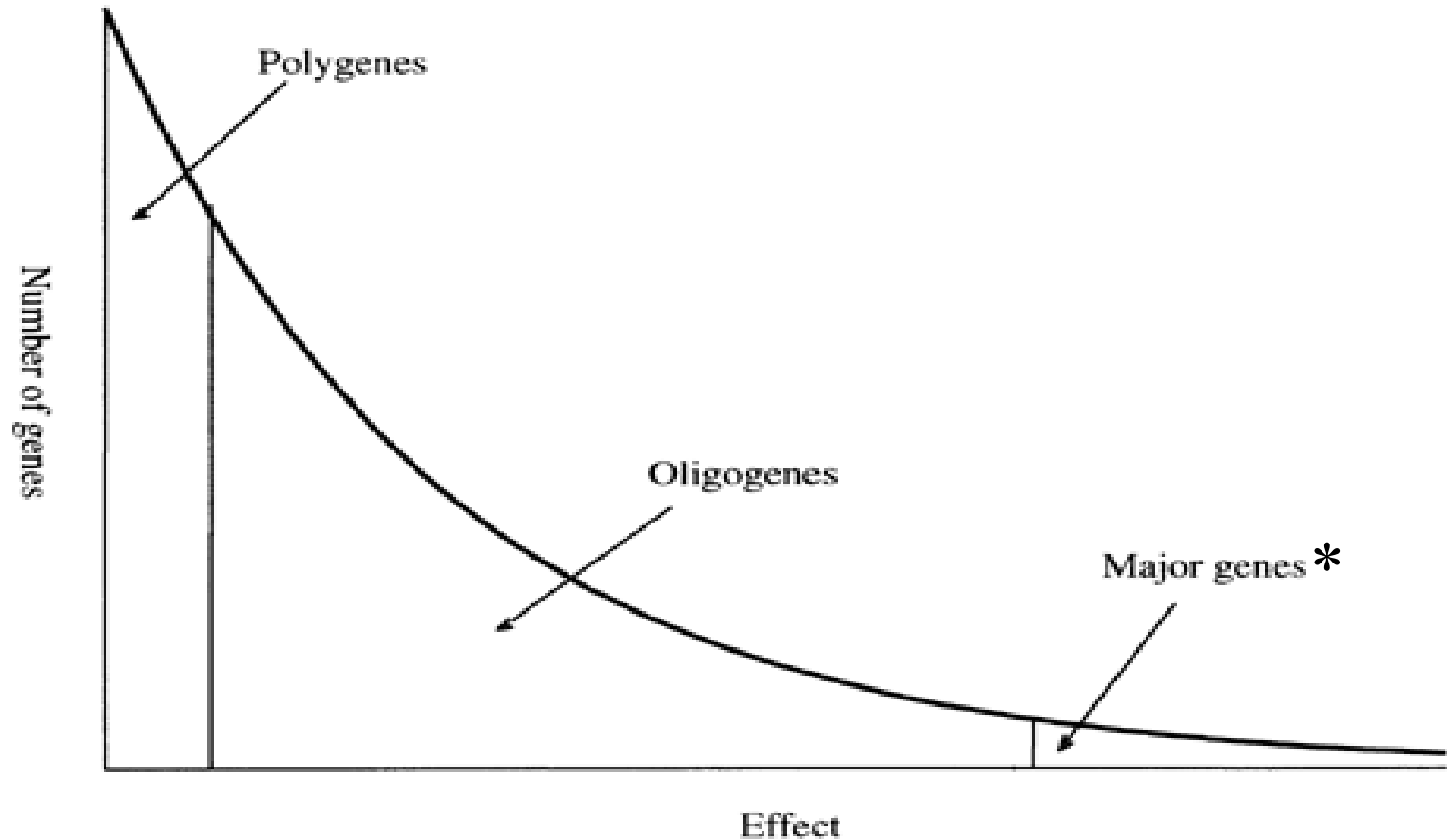
Genetic terms

- Each variant of a gene (at a given locus) is an **allele**
 - e.g. *A* or *a*
- For a genetic marker, the two parentally-inherited variants combined are called a **genotype**
 - e.g. *A,a*
- The set of variants along one chromosome (at different loci) is called a **haplotype**
 - e.g. *ABcDefGH*
- The two sets of variants on both chromosomes are sometimes called a **diplotype**
 - e.g. *ABcDefGH, aBCDefgh*

Step back: Monogenic vs. polygenic vs. environmental vs. multifactorial diseases



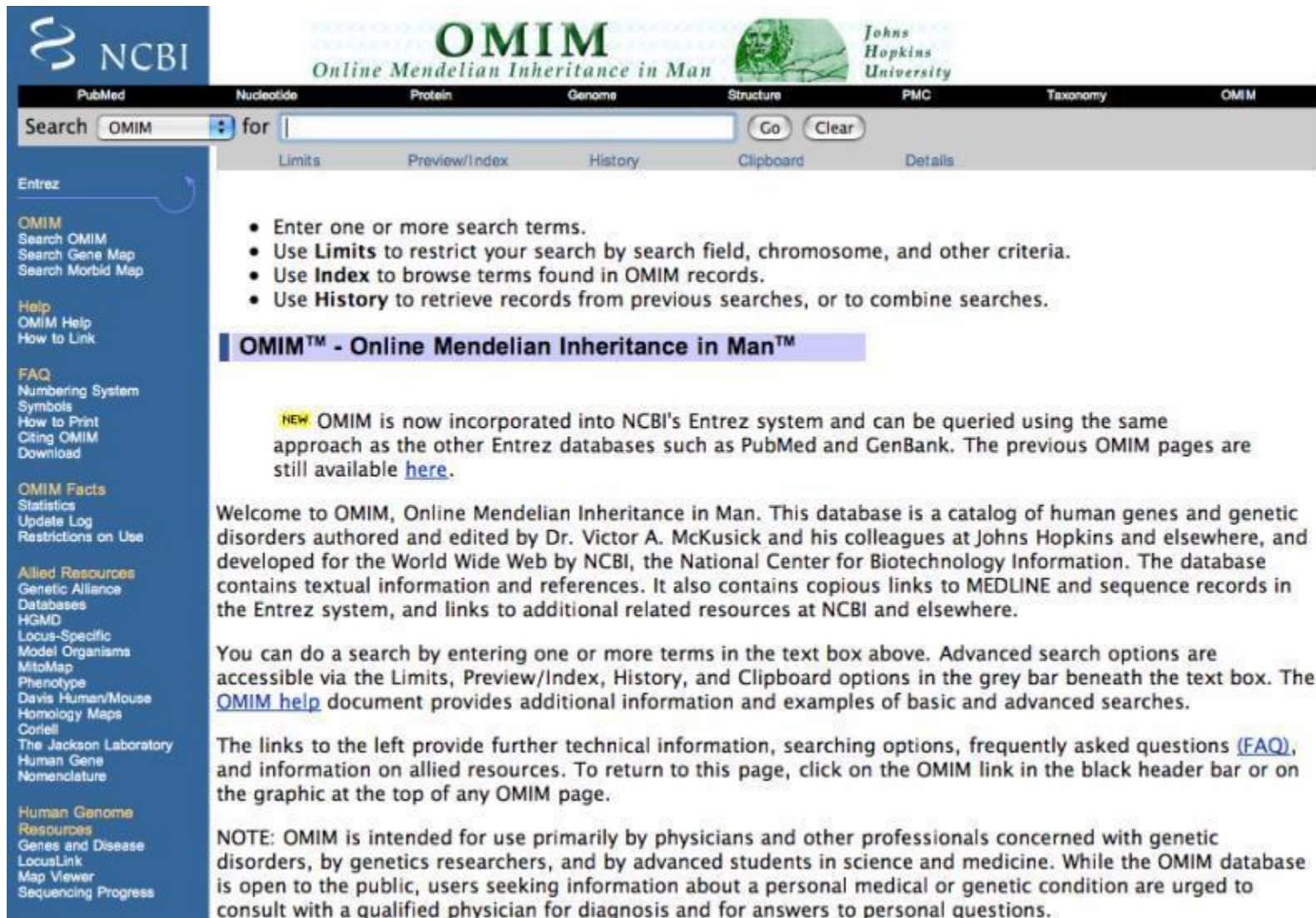
Number of genes vs. effect size



* about 2000 established monogenic disorders on OMIM (e.g. Huntington chorea, cystic fibrosis, retinitis pigmentosa)

OMIM (Online Mendelian Inheritance in Man)

<http://www.ncbi.nlm.nih.gov/omim/>



Search OMIM for

Limits Preview/Index History Clipboard Details

- Enter one or more search terms.
- Use **Limits** to restrict your search by search field, chromosome, and other criteria.
- Use **Index** to browse terms found in OMIM records.
- Use **History** to retrieve records from previous searches, or to combine searches.

OMIM™ - Online Mendelian Inheritance in Man™

NEW OMIM is now incorporated into NCBI's Entrez system and can be queried using the same approach as the other Entrez databases such as PubMed and GenBank. The previous OMIM pages are still available [here](#).

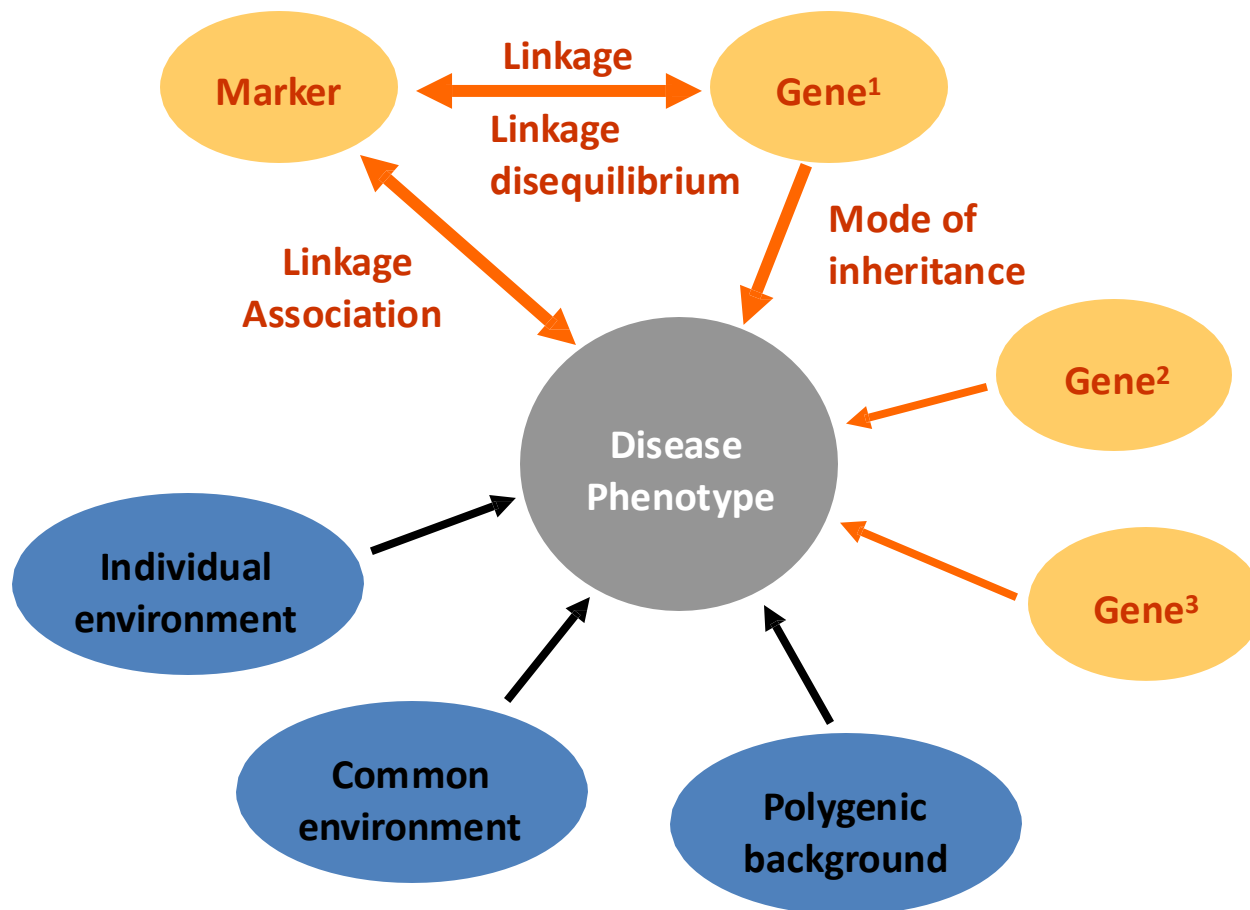
Welcome to OMIM, Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information and references. It also contains copious links to MEDLINE and sequence records in the Entrez system, and links to additional related resources at NCBI and elsewhere.

You can do a search by entering one or more terms in the text box above. Advanced search options are accessible via the Limits, Preview/Index, History, and Clipboard options in the grey bar beneath the text box. The [OMIM help](#) document provides additional information and examples of basic and advanced searches.

The links to the left provide further technical information, searching options, frequently asked questions ([FAQ](#)), and information on allied resources. To return to this page, click on the OMIM link in the black header bar or on the graphic at the top of any OMIM page.

NOTE: OMIM is intended for use primarily by physicians and other professionals concerned with genetic disorders, by genetics researchers, and by advanced students in science and medicine. While the OMIM database is open to the public, users seeking information about a personal medical or genetic condition are urged to consult with a qualified physician for diagnosis and for answers to personal questions.

Complex Trait Model



Step back: Key concepts of population genetics

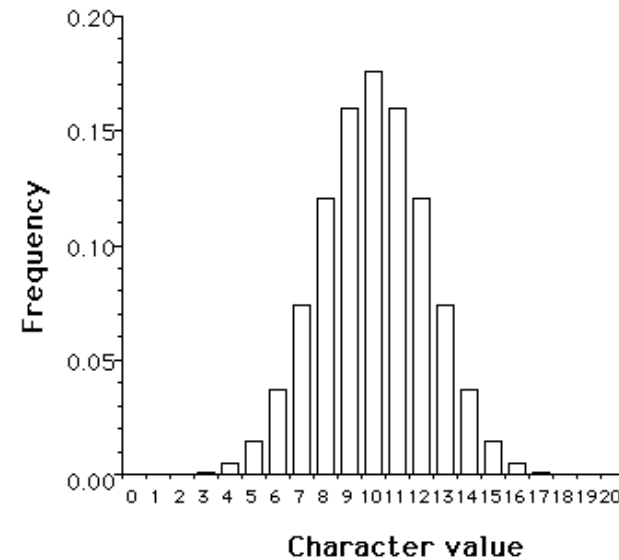
- Heritability
- Hardy-Weinberg equilibrium
- Linkage disequilibrium

Heritability (of a trait) definitions

- **Fraction of phenotypic variability that is attributable to genetic variation**
- IS NOT: how much genetics influences trait in one person
- Is relative to specific population in a particular environment (since contribution of genetic factors is relative to contribution of other factors such as environment)

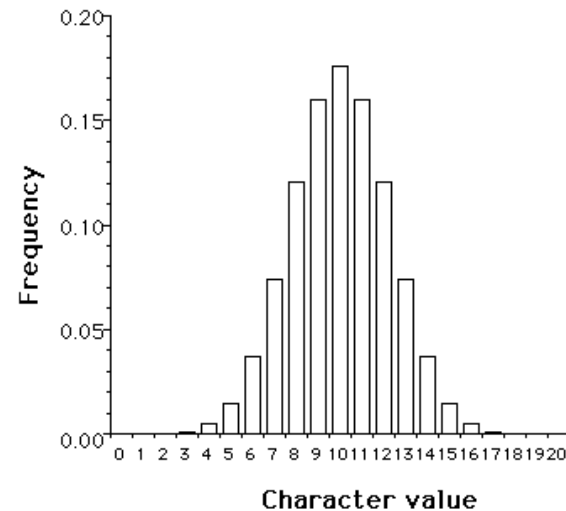
The total phenotypic variance:

- VP = total phenotypic variance
- VG = total genetic variance
- VE = environmental variance
- $VP = VG + VE$



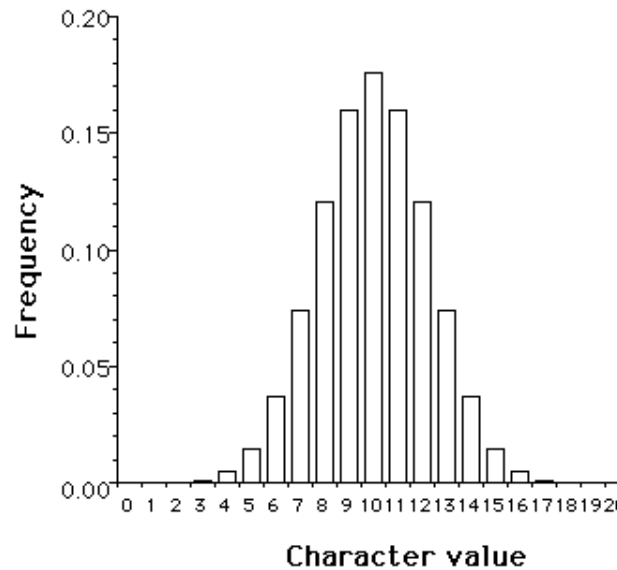
The total genetic variance (VG):

- V_A = additive genetic variance
- V_D = dominance genetic variance
- V_I = epistatic (interactive) genetic variance
- $V_G = V_A + V_D + V_I$



Estimating heritability

- Broad sense heritability = V_G / V_P
- Narrow sense heritability = V_A / V_P
- Narrow sense heritability is more commonly used



Estimating heritability

One common approach is to compare phenotypic scores of parents and their offspring:

Mid Parent	Offspring
168	171
182	178
161	163

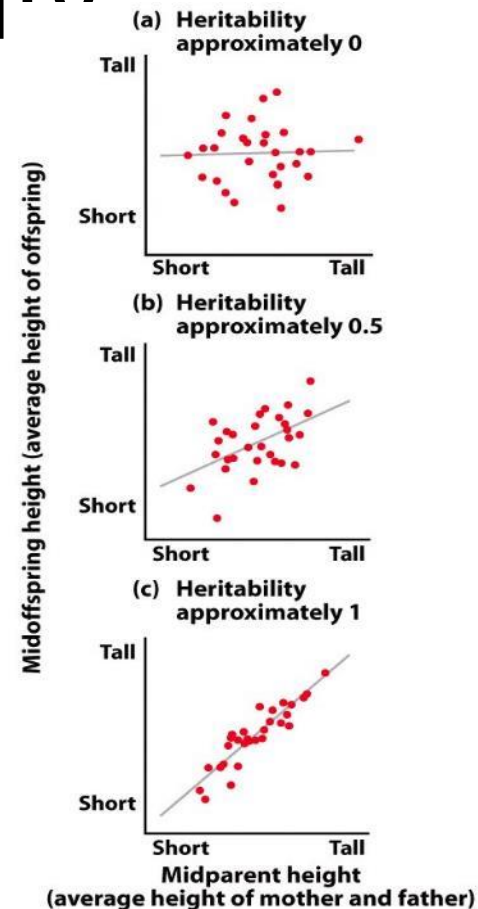
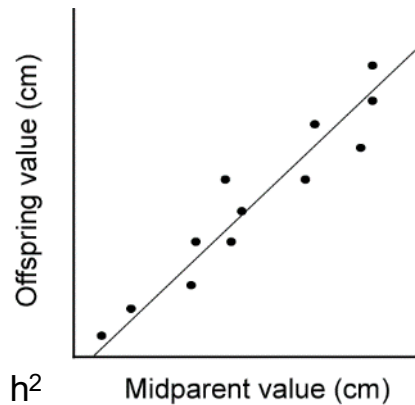
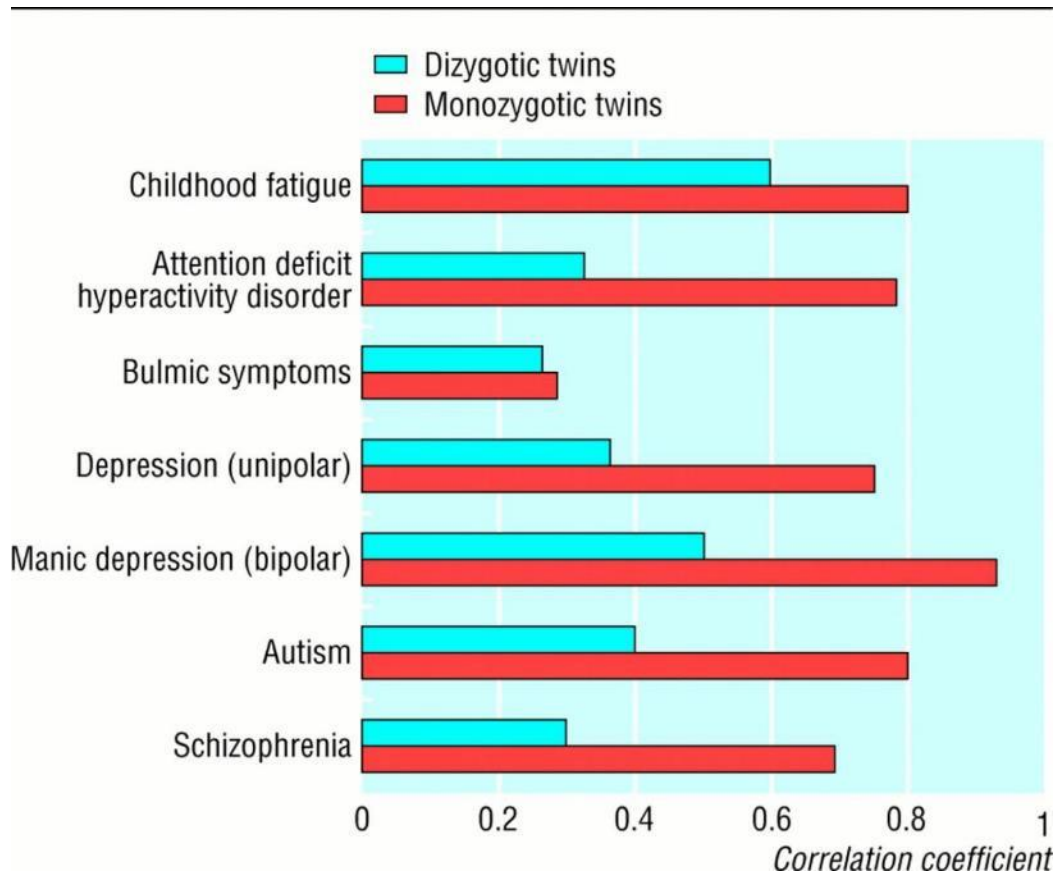


Figure 9-13 Evolutionary Analysis, 4/e
© 2007 Pearson Prentice Hall, Inc.

Heritability of Traits



Schizophrenia

$$r(MZ) = 0.7$$

$$r(DZ) = 0.3$$

$$H^2 = 2(0.7 - 0.3)$$

$$\text{Heritability} = 0.8$$

Some points regarding heritability

- Heritability could be estimated using family studies and adoption studies but is very common in twins study
- Heritability is not the proportion of a phenotype that is genetic, but rather the proportion of phenotypic variance that is due to genetic factors.
- Heritability is a population parameter and, therefore, it depends on population-specific factors.
- Heritability is not constant. It can change over time or in life course.

HARDY-WEINBERG THEOREM



Hardy-Weinberg Equilibrium (HWE)

- GH Hardy
- Wilhelm Weinberg
- Mathematical model of expected genotype frequencies in a population
- Allele and genotype frequencies will remain constant from generation to generation in the absence of other evolutionary influences



Godfrey Hardy
Mathematician



Wilhelm Weinberg
Physician



William Castle
Geneticist

Hardy-Weinberg theorem

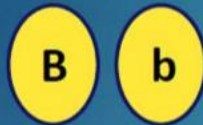
- frequency of homozygous major allele: $p \times p = p^2$
- frequency of homozygous minor allele: $q \times q = q^2$
- frequency of heterozygotes: $(p \times q) + (q \times p) = 2pq$
- frequencies of all individuals must add to 1:

- $p^2 + 2pq + q^2 = 1$

Example of HWWE

► Alleles:

$$p + q = 1$$



$$p = .60$$

$$q = .40$$

► Individuals:

$$p^2 + 2pq + q^2 = 1$$



$$p^2 = .36$$

$$2pq = .48$$

$$q^2 = .16$$

$$p^2 = .20$$

$$2pq = .64$$

$$q^2 = .16$$

In a population the allele frequency of b is 0.4

How many of each genotype?

$$q^2 (bb): 16/100 = .16$$

$$q (b): \sqrt{.16} = \mathbf{0.4}$$

$$p (B): 1 - 0.4 = \mathbf{0.6}$$

Our population (n=400)

Hardy-Weinberg

OEGE - Online Encyclopedia for Genetic Epidemiology studies

- Overview
- GWAS
- NGS
- Population Studies
- Case & Family Studies
- Genetic Databases
- Genotyping
- Phenotyping
- Software
- Disease Genes
- Genetic Diversity
- Journals

Hardy-Weinberg equilibrium calculator including analysis for ascertainment bias

Chi-sq Hardy-Weinberg equilibrium test calculator for biallelic markers (SNPs, indels etc), including analysis for ascertainment bias for dominant/recessive models (due to biological or technical causes)
Enter observed counts for each genotype, then click "Calculate". (Copyright TRG, SR, INMD, 2008)

If you use this web-tool please cite:

Santiago Rodriguez, Tom R. Gaunt and Ian N. M. Day.

Hardy-Weinberg Equilibrium Testing of Biological Ascertainment for Mendelian Randomization Studies.

American Journal of Epidemiology Advance Access published on January 6, 2009, DOI 10.1093/aje/kwn359.

Common homozygotes 20 Heterozygotes 64 Rare Homozygotes 16

Calculate Reset

Result

$\chi^2 = 7.96$

(100 samples counted)

for likelihoods of calculated χ^2 value see below.

Genotype	Expected	Observed
Common homozygotes	27.04	20
Heterozygotes	49.92	64
Rare homozygotes	23.04	16

p allele freq = 0.52; q allele freq = 0.48

Solutions for perfect HWE, under a model of ascertainment (+/-) of one group

Group affected	Common Hz	Heterozygotes	Rare Hz	p allele freq	q allele freq
Common Hz	64	64	16	0.67	0.33
Heterozygotes	20	35.78	16	0.53	0.47
Rare Hz	20	64	51.2	0.38	0.62

$$p^2 = .20$$

$$2pq = .64$$

$$q^2 = .16$$

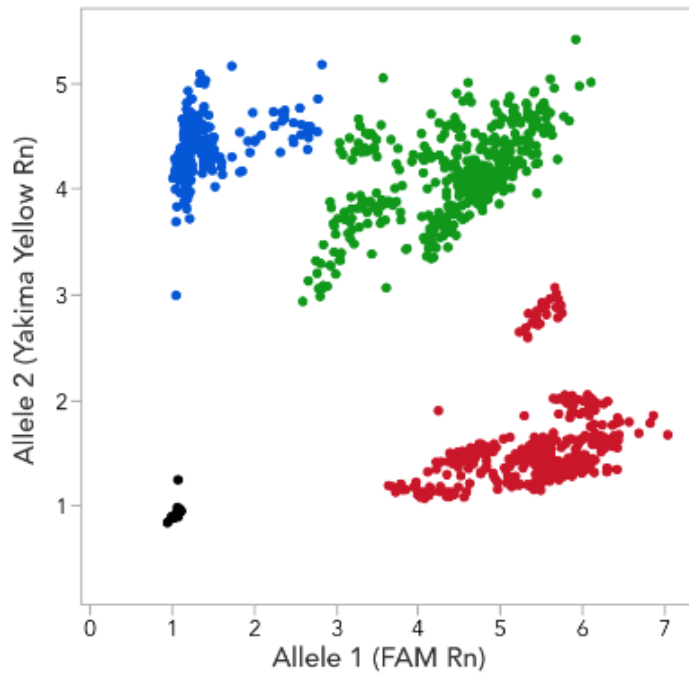
Our population (n=400)

Hardy-Weinberg Equilibrium (HWE)

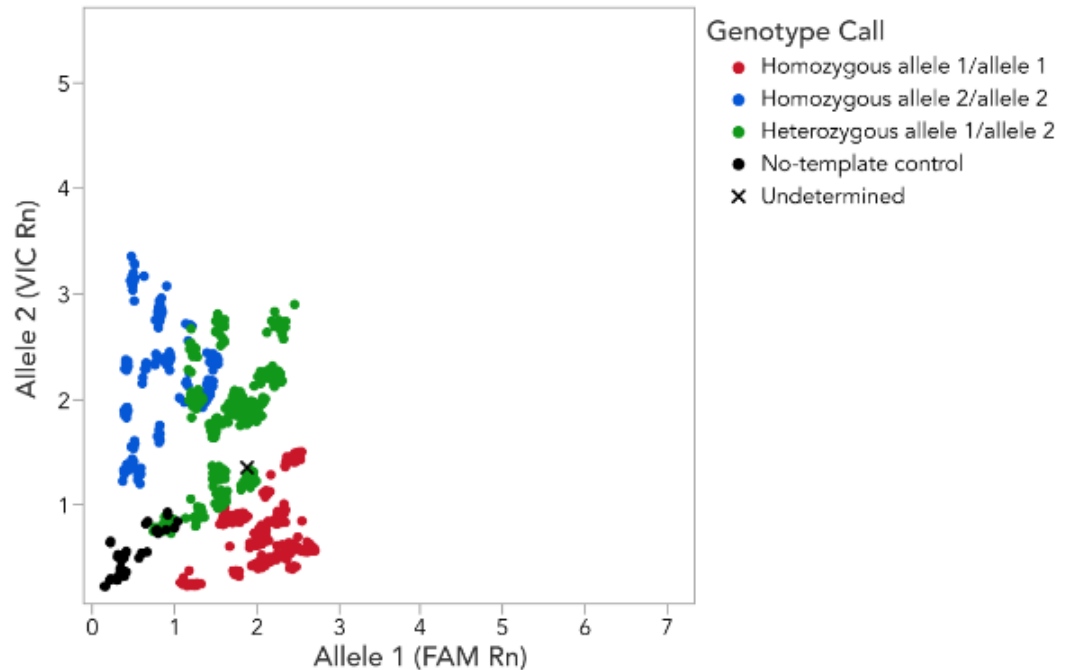
- Violations of HWE could be due to:
 - Non-random mating (i.e., inbreeding)
 - Natural selection
 - Mutation
 - Migration
 - Chance (in small populations)

Genotype call and HWE error

Filtering on HWE is mainly done to filter genotyping error



A. rhAmp SNP Assays



B. SNP assays from Supplier T

LINKAGE DISEQUILIBRIUM (LD)



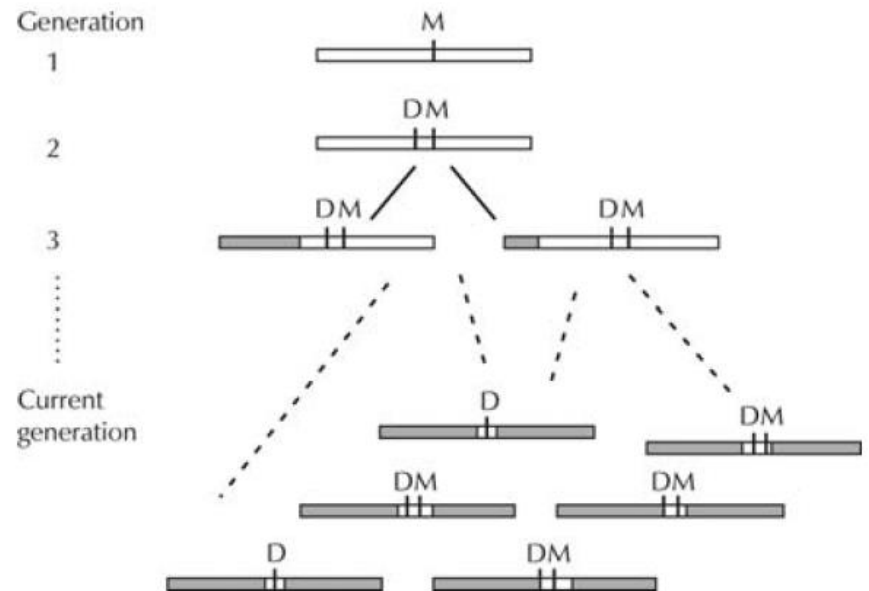
Linkage disequilibrium (LD)

- Non-random association of alleles at different loci
- Presence of statistical associations between alleles at different loci that are different from what would be expected if alleles were independently transmitted from generation to generation

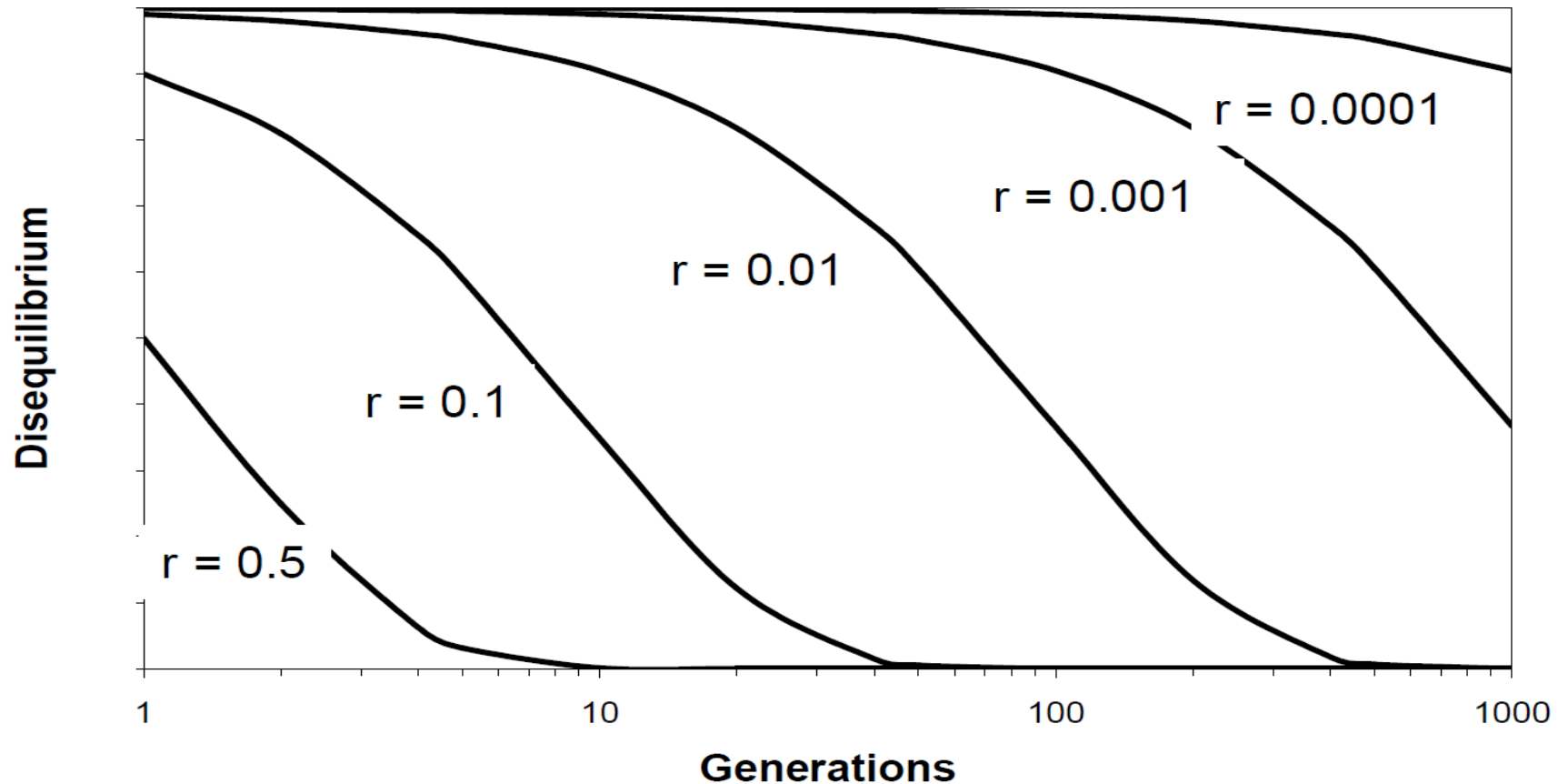
Linkage disequilibrium (LD)

➤ Measures of LD

- r^2 ($r^2=1$ implies the SNP alleles are perfectly correlated)

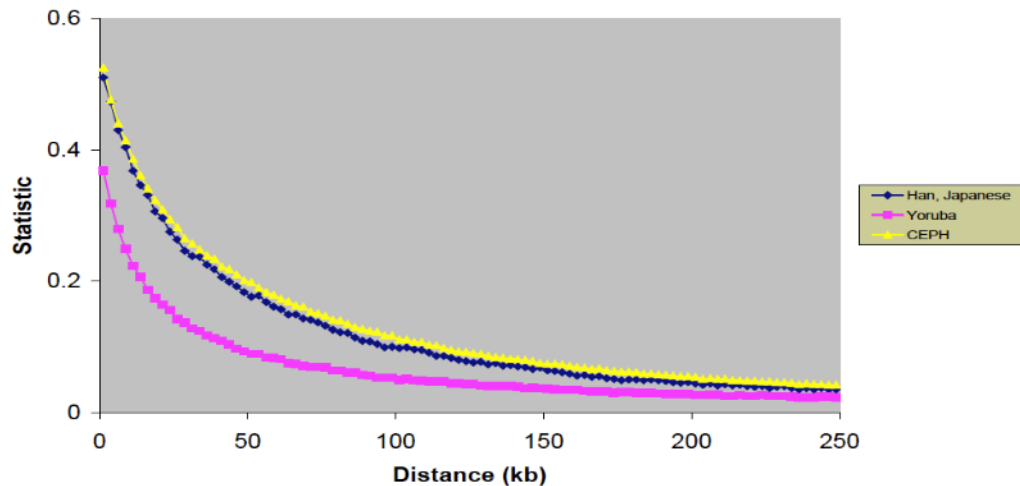


LD is diminished with time and increased recombination rate



LD and genomic distance

- The closer the markers, the stronger the LD since recombination will have occurred at a low rate



LD extends further in CEPH and the Han/Japanese than in the Yoruba

International HapMap Consortium, *Nature*, 2005

rs714180 rs39747 rs38845 rs9641562 rs38846 rs7798983 rs38849 rs10246585 rs10243024 rs38855 rs38857 rs10215153 rs2283053 rs2402118 rs2023748 rs1621

Block 1 (3 kb) Block 2 (11 kb) Block 3 (7 kb) Block 4 (38 kb)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

36 99 70 35 53 73 5 70 66 32 4 8 69 29 26 7

28 99 70 35 53 73 5 70 66 32 4 8 69 29 26 7

99 70 35 53 73 5 70 66 32 4 8 69 29 26 7

70 35 53 73 5 70 66 32 4 8 69 29 26 7

35 53 73 5 70 66 32 4 8 69 29 26 7

53 73 5 70 66 32 4 8 69 29 26 7

73 5 70 66 32 4 8 69 29 26 7

5 70 66 32 4 8 69 29 26 7

70 66 32 4 8 69 29 26 7

66 32 4 8 69 29 26 7

32 4 8 69 29 26 7

4 8 69 29 26 7

8 69 29 26 7

69 29 26 7

29 26 7

26 7

7

Role of LD in genetic epidemiology studies

- If all genetic polymorphisms were independent of one another at a population level, then we would need to genotype all of them to find the genetic cause of a disease (\$\$\$)
- If a polymorphism is found to be associated with a disease, then that means that either this polymorphism is the cause of the disease or that it is in LD with the causal genetic variant

Indirect tests of association using “tag SNP” genetic markers

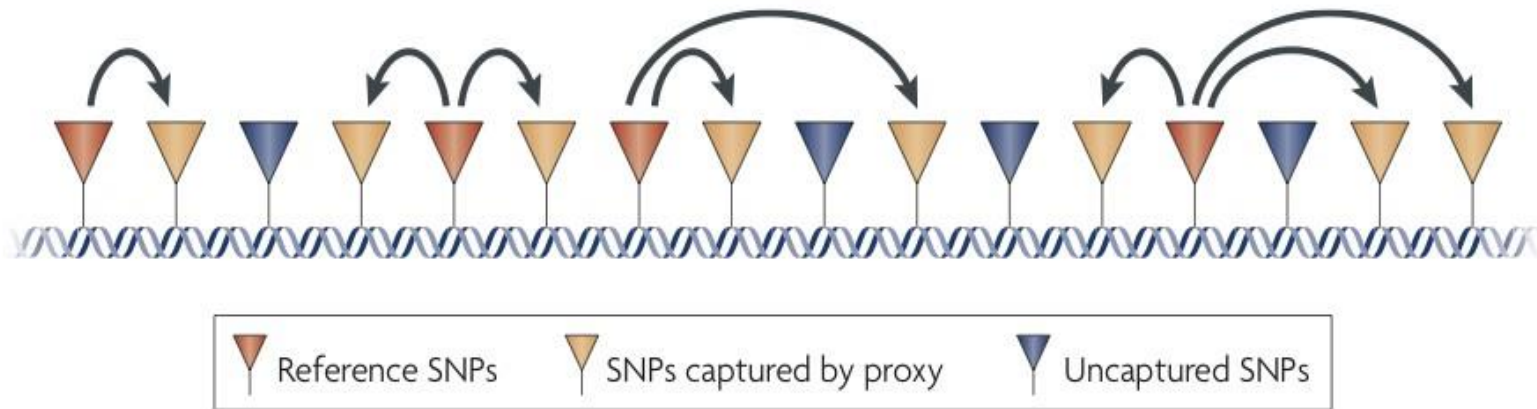
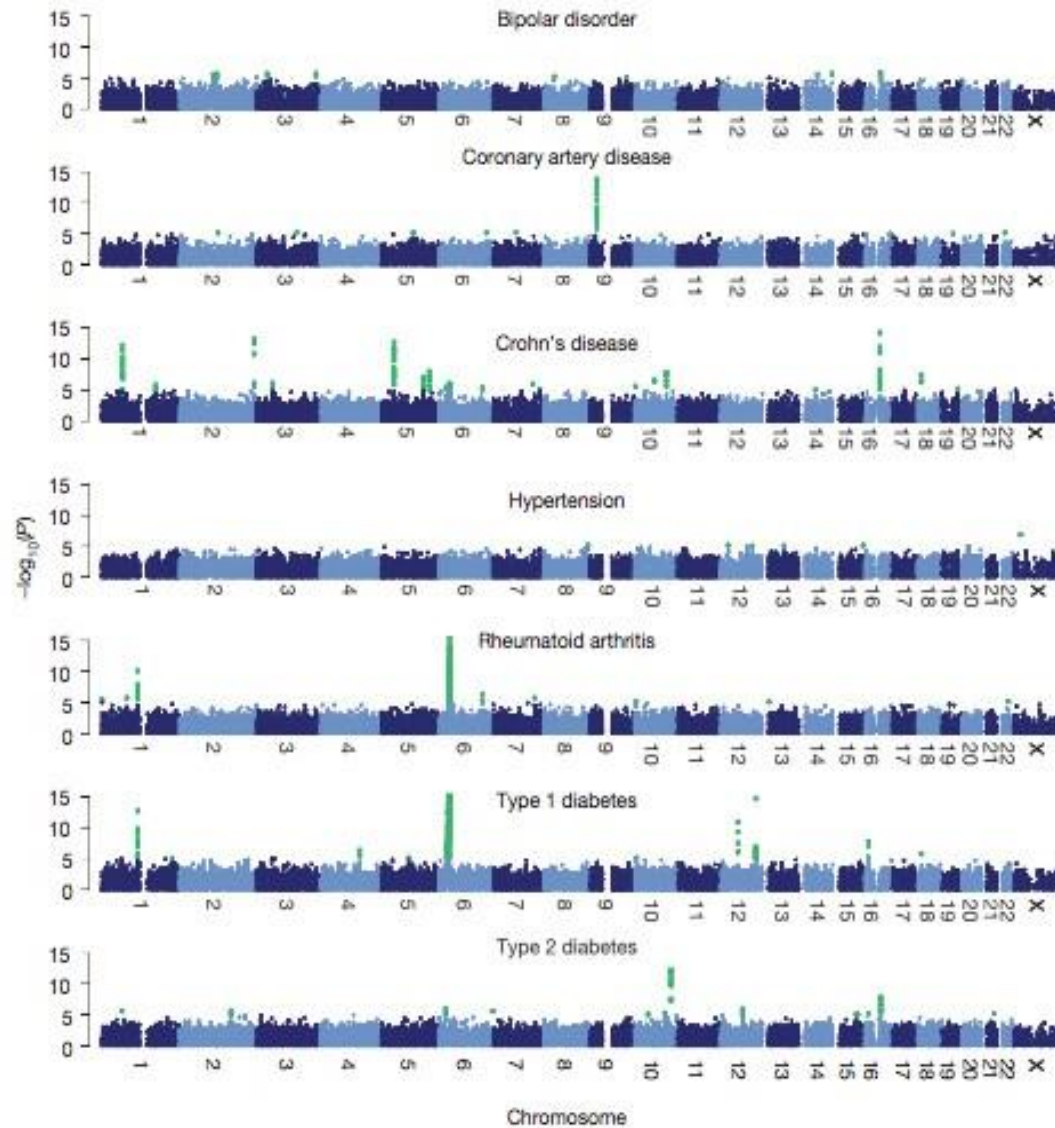


Figure 3 | **Schematic of a genomic region to be tested for association with a phenotype.** The four reference SNPs in the mapping panel are indicated by red triangles; these are genotyped directly. The eight SNPs indicated by yellow triangles are captured through linkage disequilibrium (by proxy) with the reference SNPs denoted by arrows. The four SNPs indicated by blue triangles are neither genotyped nor in linkage disequilibrium with the reference SNPs; phenotypic association that is due to one of these would be missed.

Genome-wide association studies (GWAS)



WTCC1 paper, Nature Vol 447, June 2007

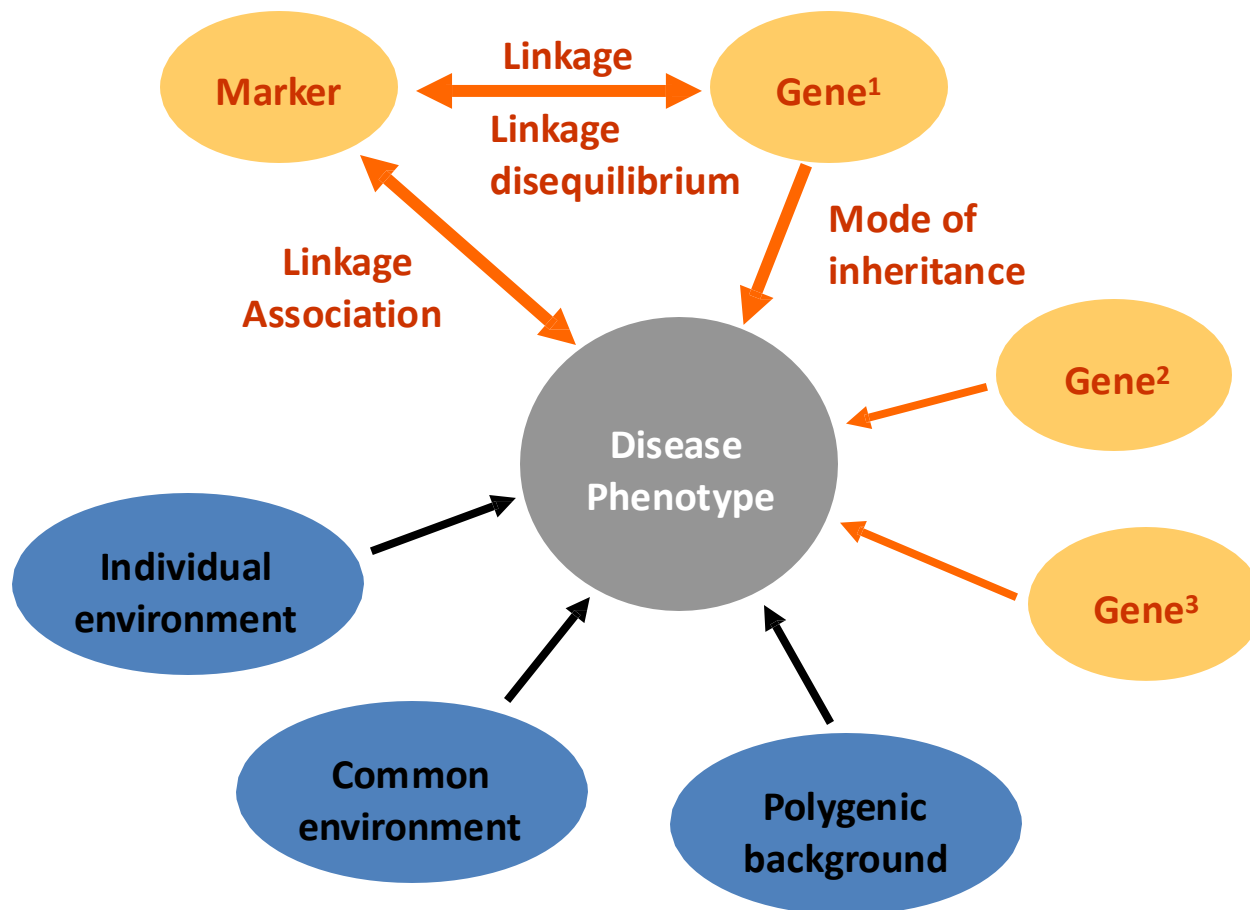
Genetic Models and Association

Penetrance reflects the risk of disease in an individual with respect to the genotype :

multiplicative, additive, recessive and dominant

Genotype	Genetic model			
	Genotype (general)	Recessive	Dominant	Additive
AA (reference)	f_0	0	0	0
AB	f_1	0	1	1
BB	f_2	1	1	2

Complex Trait Model



<http://www.hapmap.org>

THE HAPMAP PROJECT

The HapMap project

Aim: to identify the variation in DNA sequence across the whole human genome within and between ethnic groups

- A resource for disease researchers
 - Identify “common” variations ($MAF > 5\%$)
 - Identify patterns of linkage disequilibrium
 - Provide efficient strategy for disease mapping

The Original HapMap subjects

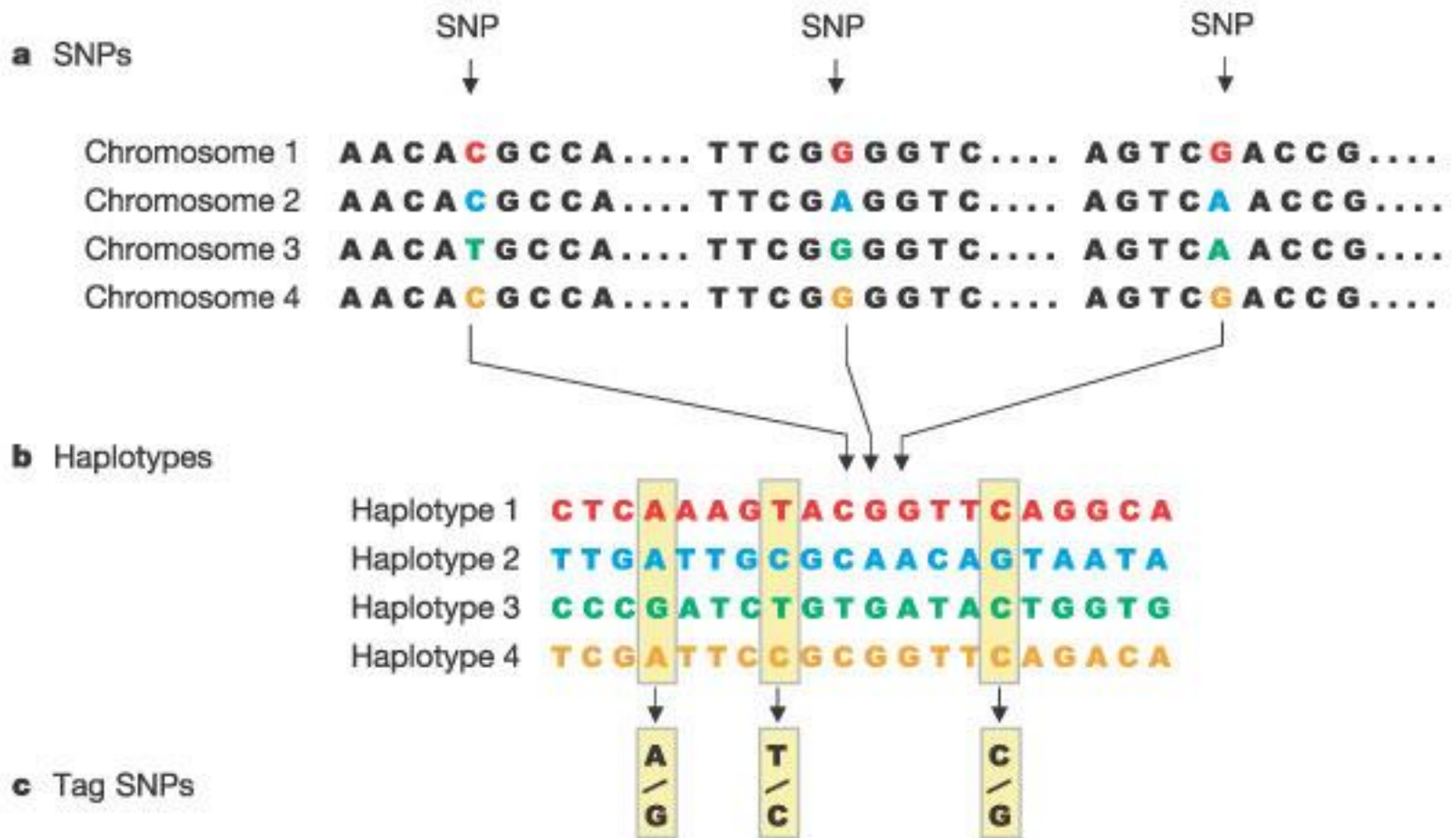
269 people:

- 30 African trios from Yoruba, Nigeria (YRI)
- 44 Japanese from Tokyo (JPT)
- 45 Chinese from Beijing (CHB)
- 30 US trios with European ancestry (CEU)

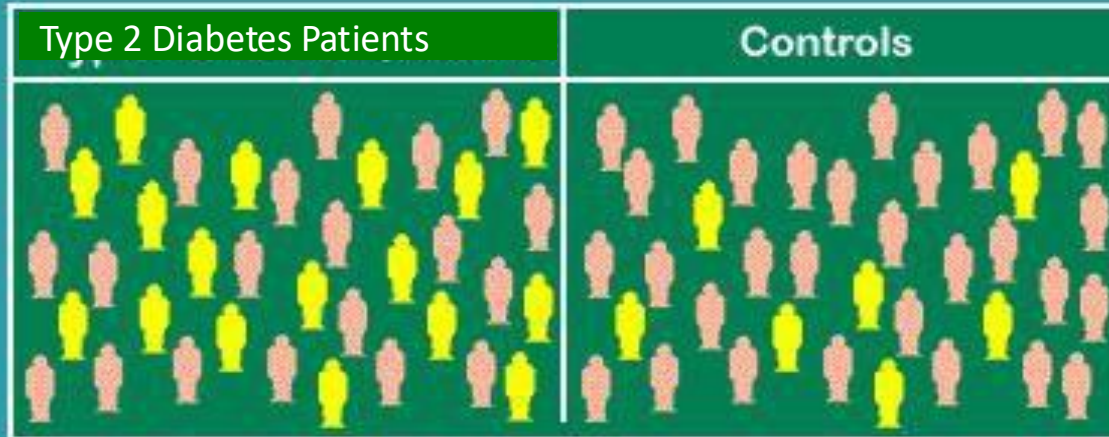
Utility of the HapMap

- Linkage between two genetic markers can be measured
- If they are close together on a chromosome they will tend to be inherited together and are said to be in linkage disequilibrium
- This allows efficient typing of markers across the genome by choosing markers that “tag” large numbers of other markers

Using SNPs to track predisposition to disease



Association Studies



Genotype	T2D	Controls	Total
TCF7L2-A	17	7	24
non-TCF7L2-A	20	30	50
	37	37	

$$\chi^2_{.05} = 5.377$$

$p < 0.025$

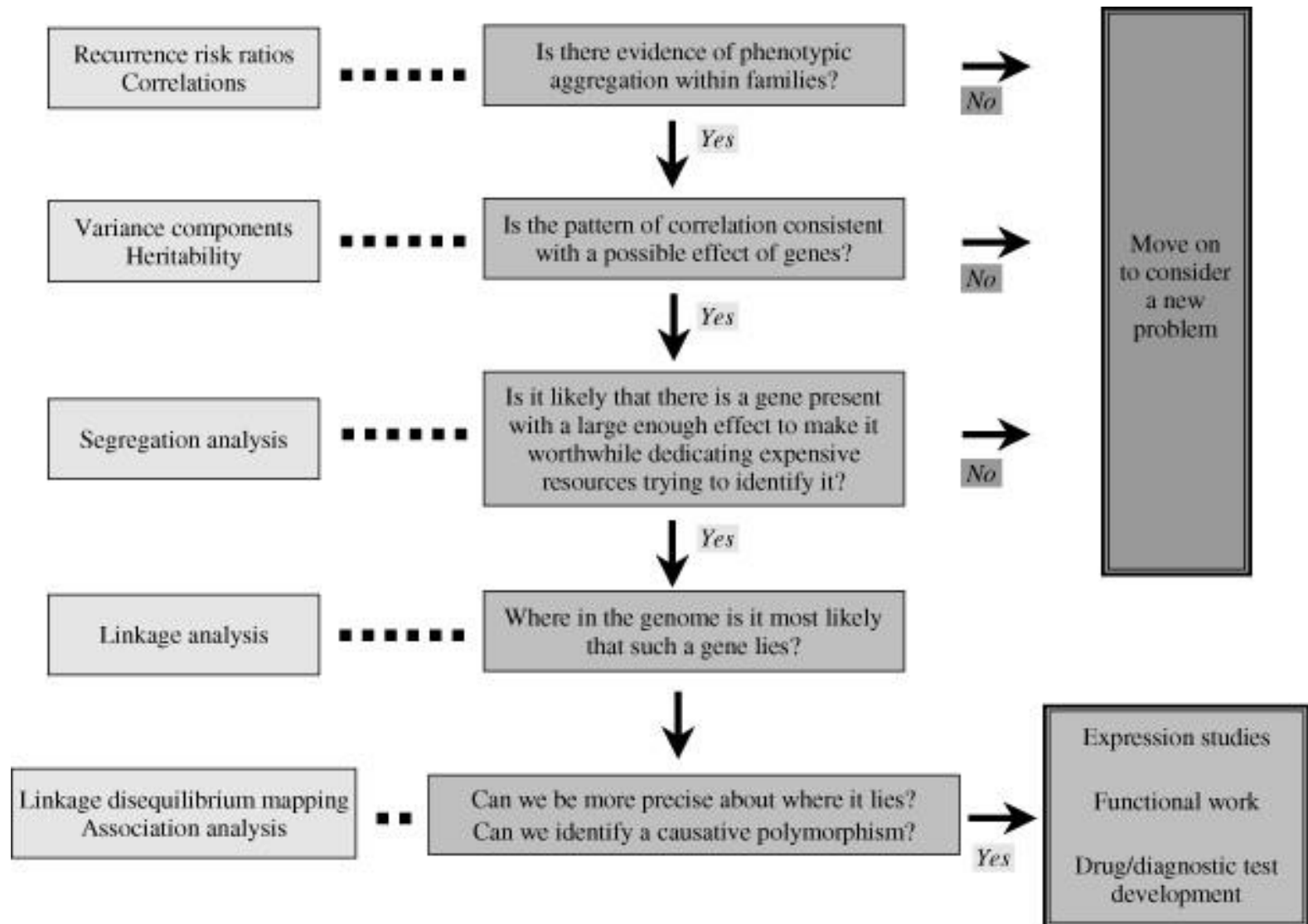
 = TCF7L2-A

 = non-TCF7L2-A

Odds Ratio: 3.6
95% CI = 1.3 to 10.4

Study designs in Genetic Epidemiology

GENETIC EPIDEMIOLOGY RESEARCH METHODS



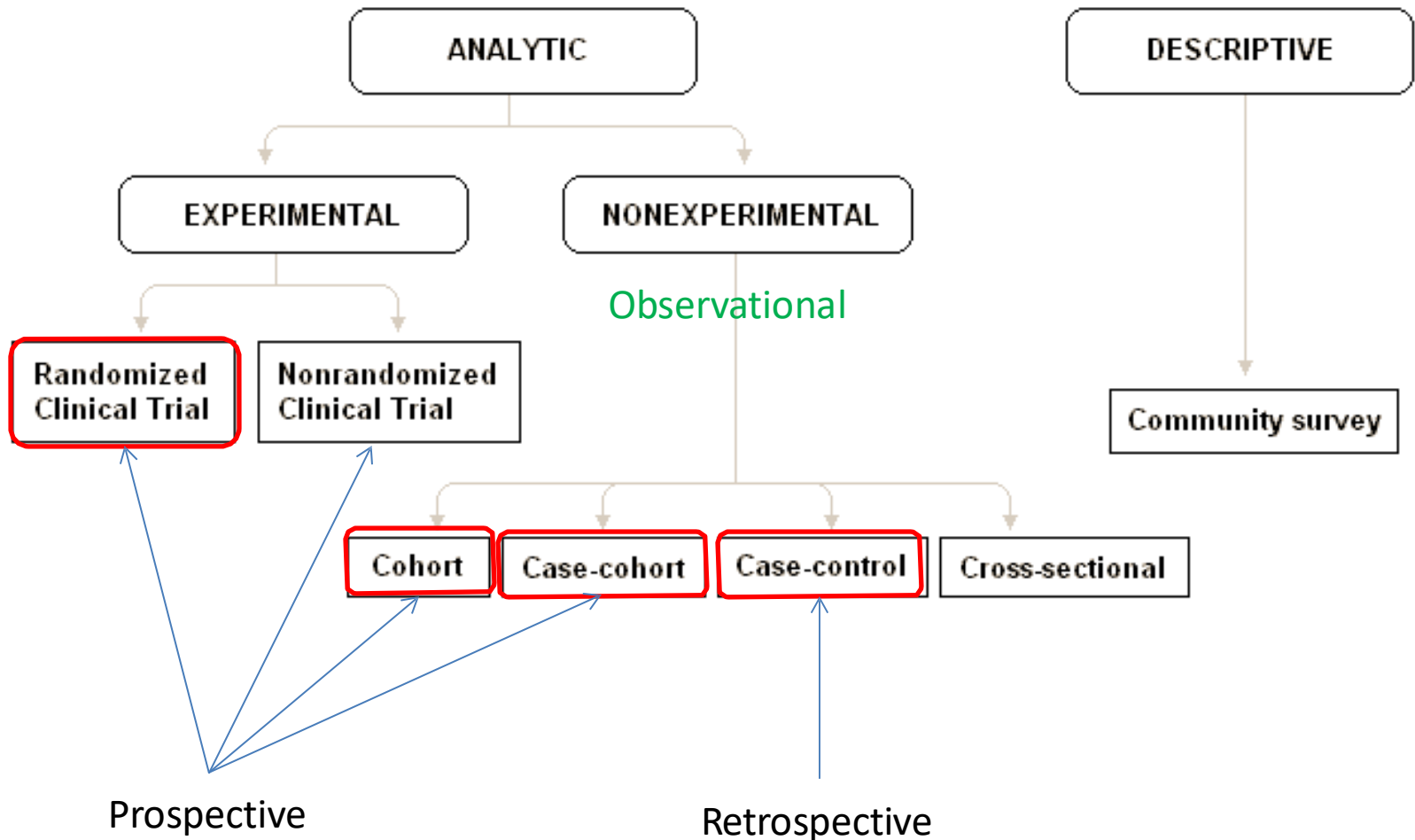
Outline

- **Study designs**
 - Overview
 - Case-control studies
 - Cohort studies
 - Randomized/experimental studies
- **The road to Genome-wide association (GWA) studies**
 - Overview
 - Family studies
 - Candidate-gene association studies
 - GWA studies

Which study design?

- Purpose of the study
 - Hypothesis-testing versus hypothesis generating
 - Finding signal versus quantifying the signal
- Available resources
- Need for data collection
- Choice of outcome
- Ability to draw valid causal inference

Epidemiologic study designs for genetic studies

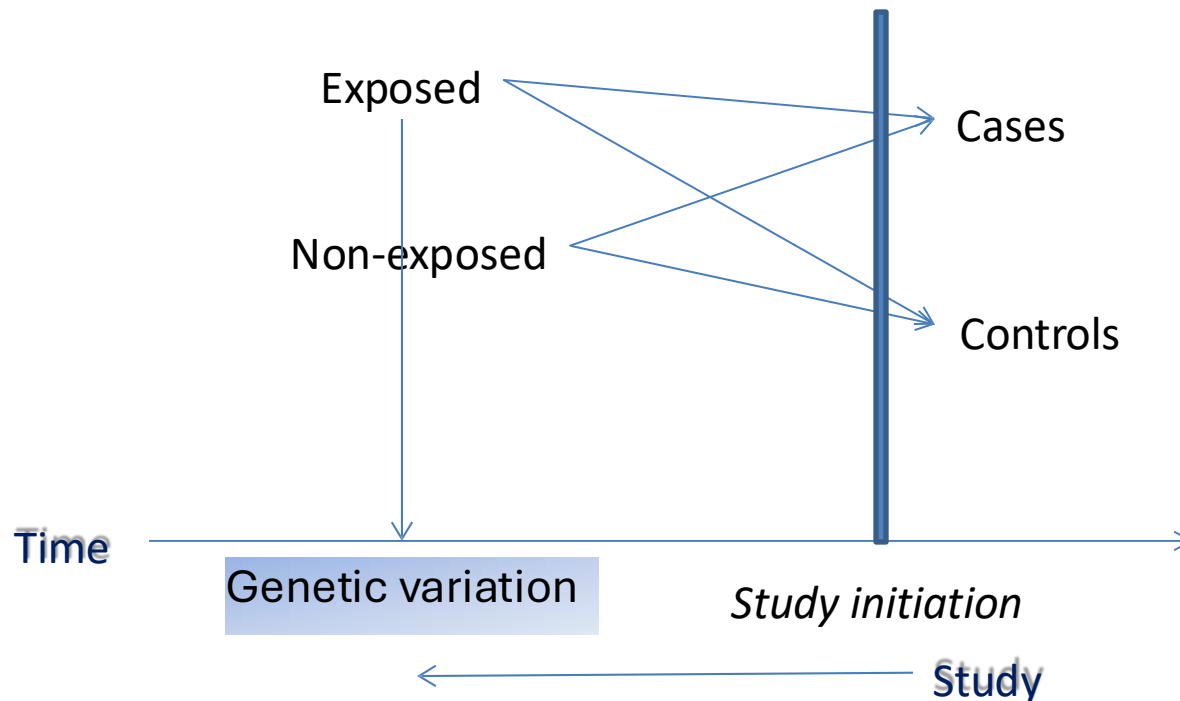


Outline

- **Learning objectives**
- **Study designs**
 - Overview
 - Case-control studies
 - Cohort studies
 - Randomized/experimental studies
- **The road to Genome-wide association (GWA) studies**
 - Overview
 - Family studies
 - Candidate-gene association studies
 - GWA studies

Case-control study design in genetics

- Design: identify participants based on their disease/outcome status, compare presence of genetic variant



Assumptions

- Cases representative of all cases of disease
- Controls drawn from the same population as cases (and at risk for the outcome)
- Exposure data (**genetic information**) collected similarly in cases and controls
 - **Genetics:** T2D cases DNA is extracted from whole blood, controls DNA is from cell lines

Measures of Risk or Association

	Cases	Controls	Totals
Exposed Allele A	a	b	M ₁
Not exposed Allele B	c	d	M ₂
Totals	N ₁	N ₂	N

Odds Ratio (OR)

$$= (a/b) / (c/d)$$

Odds of being a case if you are exposed = a/b

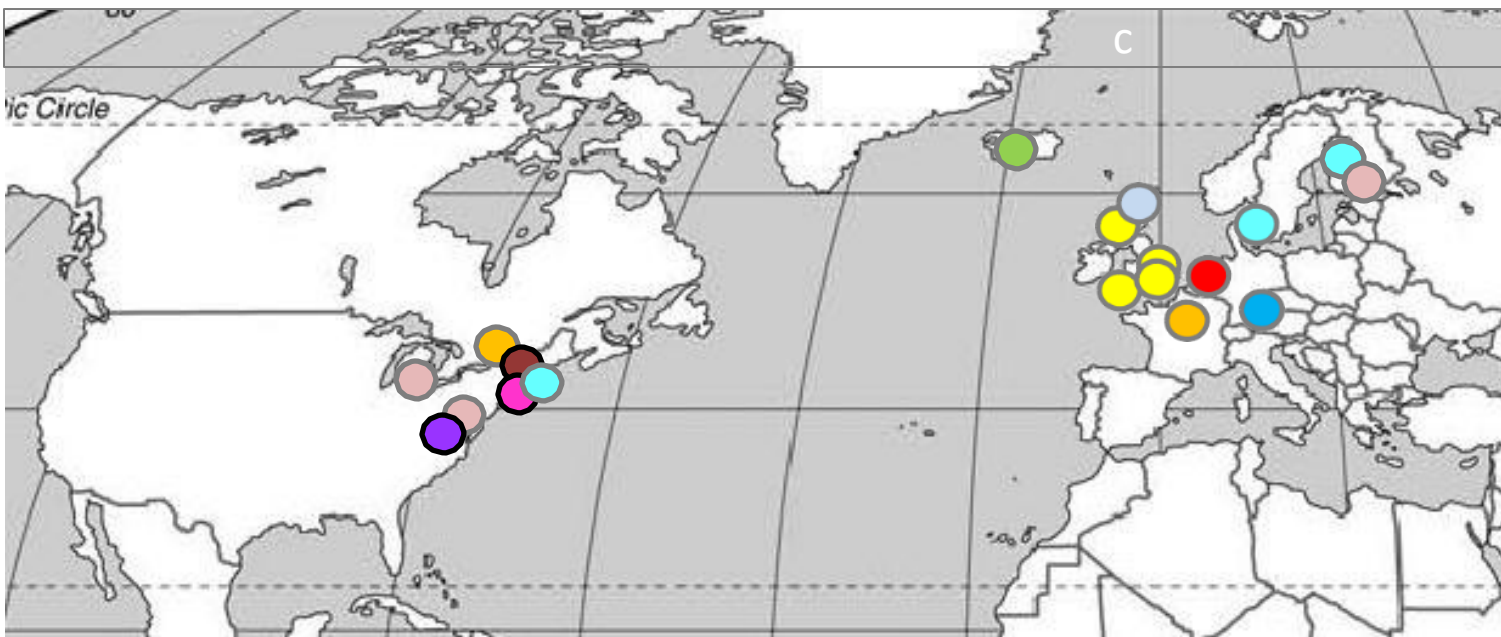
Odds of being a case if you are not exposed = c/d

Assumes log-additive genetic effects, implemented through logistic regression

Examples of case-control studies

- Oral contraceptives and reduced risk of ovarian/endometrial cancer
- *LOXL1* and exfoliation glaucoma
 - *LOXL1* catalyses the formation of elastin fibres found to be a major component of the lesions in XFG (damage to nerve fibres in the eye)
- *TCF7L2* and type 2 diabetes

The **DIabetes **G**enome-wide **R**eplication **A**nd **M**eta-analysis [**DI**AGRAM] Consortium**



WTCCC

FUSION (US/Finland)

DGI (US/Sweden/Finland)

DeCODE

ARIC

HPFS

KORA

Rotterdam

DGDG (France/Canada)

GODARTS

FHS

NHS

DIAGRAM+ 1000G imputation

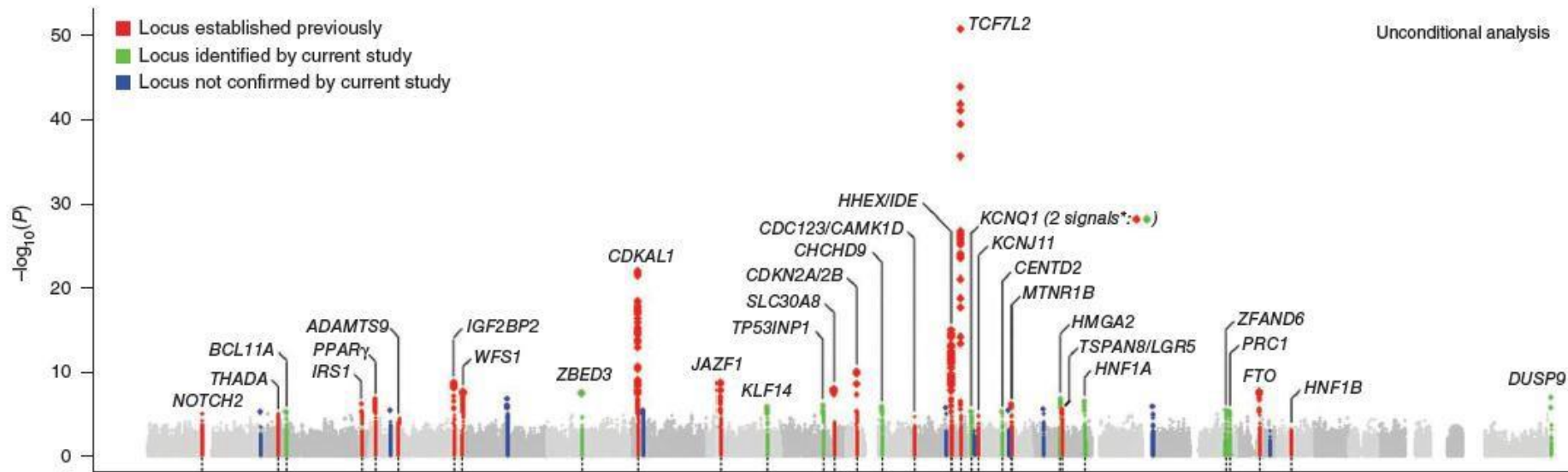
T2D case-control sets

Study	Cases	Controls	Effective size
ARIC	755	7,009	1,363
deCODE	2,012	30,114	3,772
DGDG	679	697	688
DGI	1,022	1,075	1,048
ERGO	614	4,984	1,093
FHS	673	7,660	1,237
FUSION	1,160	1,173	1,166
GODARTS	3,010	2,668	2,829
HPFS	1,124	1,298	1,205
KORAgen	433	1,438	666
NHS	1,467	1,754	1,598
WTCCC	1,924	2,938	2,325
Total	14,873	62,808	18,990

Blue – **nested** case-control studies

Compare to case-control studies

DIAGRAM+ meta-analysis



Advantages of a case-control study

- Suitable for rare outcomes
- Suitable for outcomes with long induction period
- Cheaper
- Need fewer people in some cases
- Readily evaluate multiple exposures
- Convenient
- If assumptions are met, valid estimates of relative risk

Disadvantages of a case-control study in genetics

- Retrospective (not so much of a problem in genetic epi)
- Difficult to study rare exposures
- Genetic confounding (population stratification)
- Problematic when investigating G*E interactions
- Special considerations (more later)
 - Exposure-related
 - Recall bias: Disease status may influence reporting (not so much of a problem in genetic epidemiology as genetic variation is determined at the time of gamete formation)

Subtypes of case-control studies

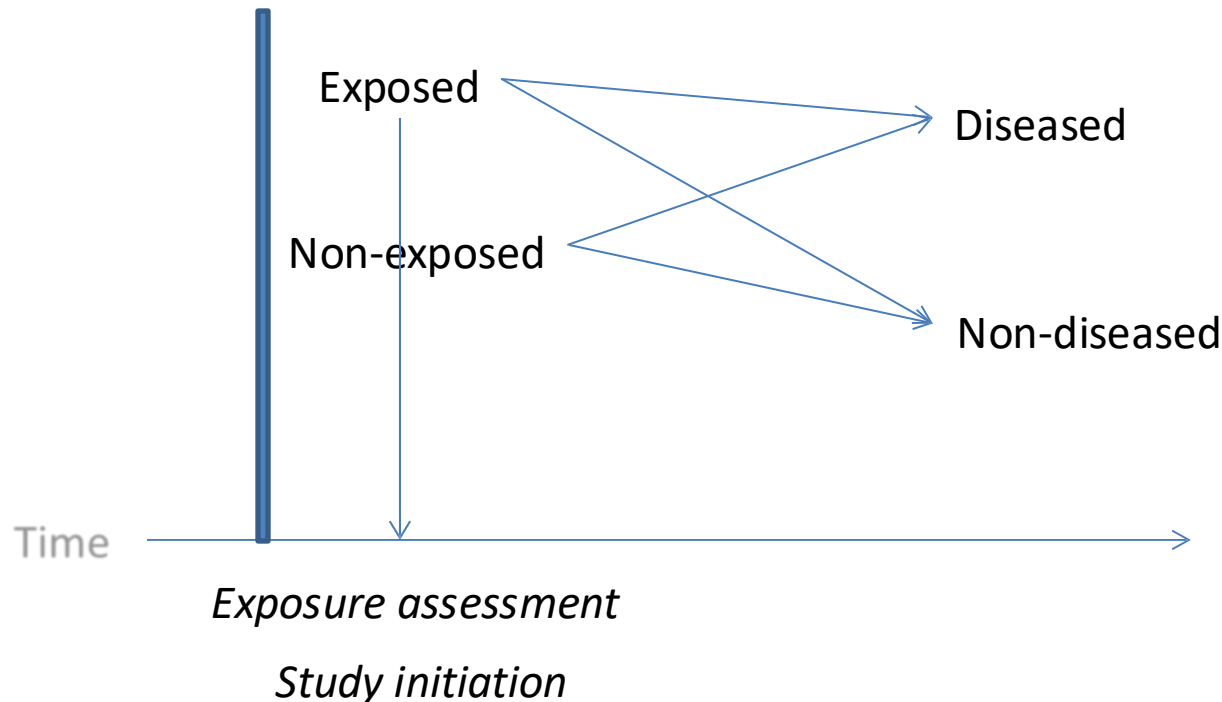
- Nested case-control
 - Within a cohort study, compares all cases to a subset of persons who did not develop disease
- Case-cohort
 - Within a cohort study, compares all cases to a random subsample of the cohort
 - Sub-cohort can be used for multiple case groups
- Super-cases and super-controls
 - Extremes of the phenotypes
 - Maximizes opportunity to detect signal

Outline

- **Learning objectives**
- **Study designs**
 - Overview
 - Case-control studies
 - **Cohort studies**
 - Randomized/experimental studies
- **The road to Genome-wide association (GWA) studies**
 - Overview
 - Family studies
 - Candidate-gene association studies
 - GWA studies

Cohort studies

- Identify individuals based on their exposure status, follow-up to ascertain disease/outcome status



Assumptions

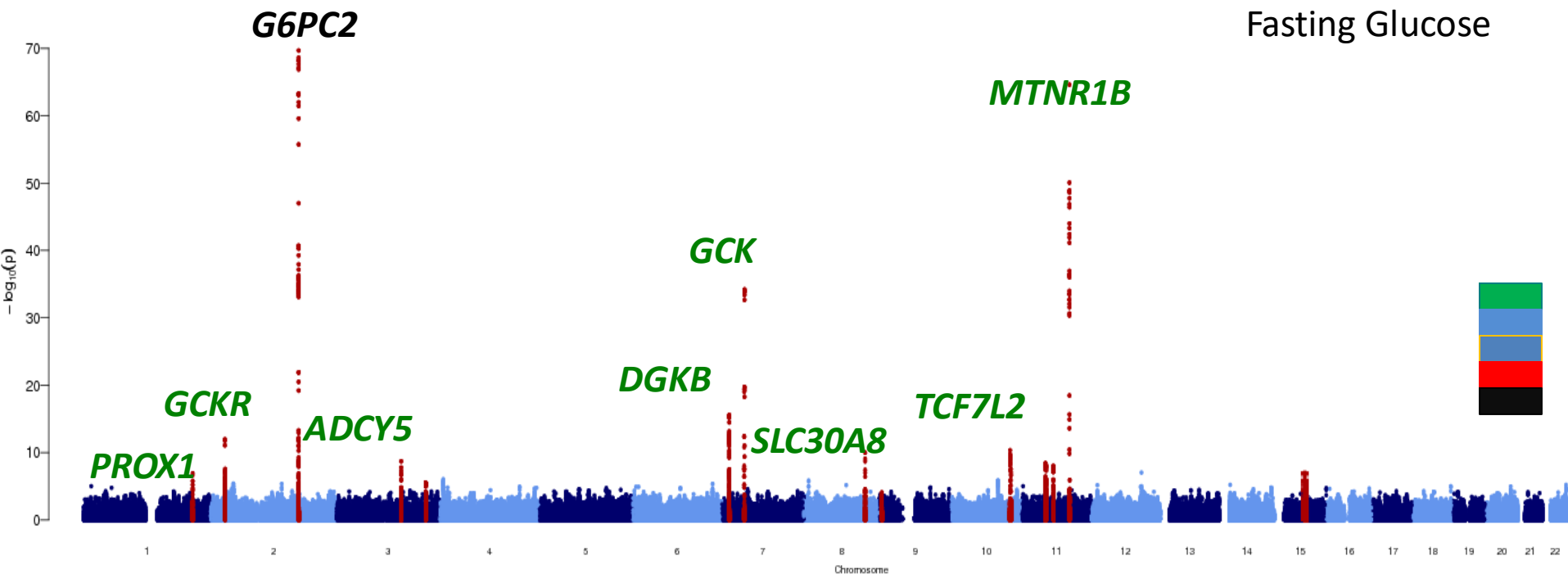
- Exposed and non-exposed groups are representative of a well-defined general population
- Outcome assessment comparable between exposed and non-exposed

Measure of genetic effects

- Cohort studies are often used for quantitative outcomes in genetic studies
 - BMI, eye colour
- Genetic model assumes additive genetic effects to test for association
 - r -fold increase in phenotype values for each risk allele
 - Uses linear regression with number of risk alleles as predictor and trait value as outcome
 - Trend test, 1df

MAGIC meta-analysis: 16 loci associated with FG & HOMA-B

- FG in 21 cohorts, **n=46,186**
- normoglycaemic individuals (FG<7mmol/l)



Advantages of a cohort study in genetics

- Able to directly estimate disease incidence
- Optimal for short induction periods
 - Induction period = time from exposure to manifest disease
- Can look at multiple outcomes
- Potential to investigate natural history of disease
- Amenable to both quantitative and binary outcomes
- Risk factors ascertained prior to disease
- Ideal for gene*environment interaction analyses

Disadvantages of a cohort study

- Not suitable for rare exposures or rare outcomes
- Requires large populations
- May be more expensive, time consuming

Outline

- **Learning objectives**
- **Study designs**
 - Overview
 - Case-control studies
 - Cohort studies
 - **Randomized/experimental studies**
- **The road to Genome-wide association (GWA) studies**
 - Overview
 - Family studies
 - Candidate-gene association studies
 - GWA studies

Randomized designs

- Definition: a comparative study in which study subjects are assigned by a formal chance mechanism between two or more intervention strategies
- Gold standard for inferring causality
- Also called “randomized controlled trials, randomized clinical trials, experimental studies”

Randomized design

- Methods of randomization
 - Several choices, from “flipping a coin” to stratified randomization
 - Alleles are distributed randomly during meiosis and mating in the population is random = alleles can be used as method for randomization

A special case:
„Mendelian randomization“

Parameter of interest:

γ – the causal effect of X on Y

Complication:

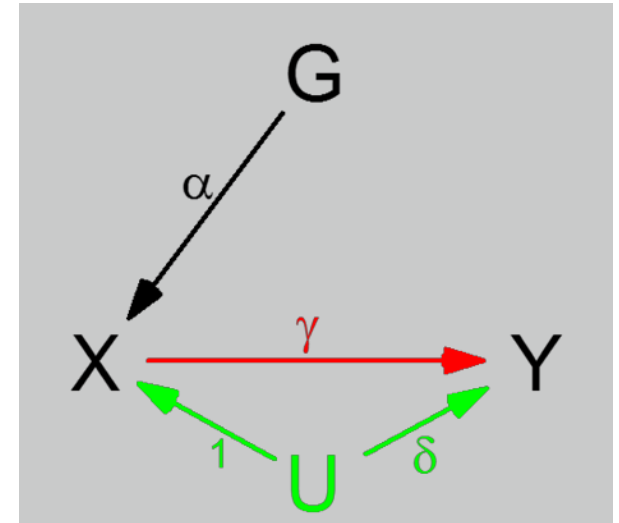
Association between X and Y is
confounded by U

Solution:

genotype G serves as an **instrument** –
correlation between G and Y provides
evidence on γ

Untestable assumptions:

no direct effect of G on Y; no association between G and U



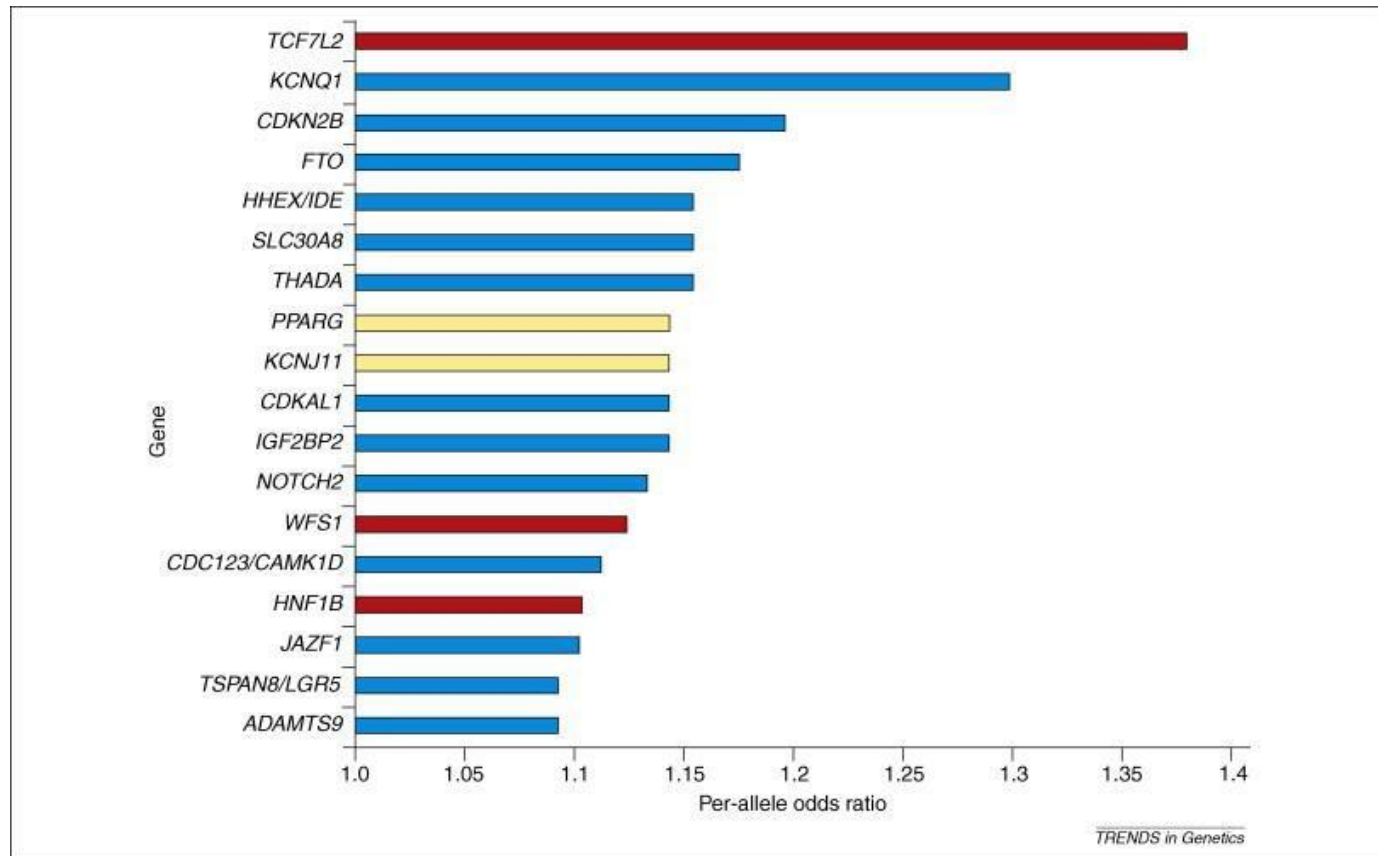
Outline

- Learning objectives
- Study designs
 - Overview
 - Case-control studies
 - Cohort studies
 - Randomized/experimental studies
- **The road to Genome-wide association (GWA) studies**
 - Overview
 - Family studies
 - Candidate-gene association studies
 - GWA studies

Progression of genetic epidemiology

- Twin studies, family studies → candidate SNPs → candidate gene → genome-wide association
- Intersection of developments in biology, technology and statistical methods
- Emphasis shifting from hypothesis-driven to agnostic study designs
- Expanding focus from single gene disorders to common, multi-genic diseases

Identification of T2D associated loci



Loci shown in blue are those identified by GWA approaches, whereas those found by candidate-gene approaches and by large-scale association analyses are shown in yellow and red, respectively.

Prokopenko, Lindgren, McCarthy, *Trends in Genetics*, 2008

Outline

- Learning objectives
- Study designs
 - Overview
 - Case-control studies
 - Cohort studies
 - Randomized/experimental studies
- **The road to Genome-wide association (GWA) studies**
 - Overview
 - Family studies
 - Candidate-gene association studies
 - GWA studies

Why family studies?

- Good route for gene discovery in Mendelian disorders
 - Strong familial clustering suggests genetic basis
 - Sentinel families good for studying specific phenotypes
 - Less susceptible to population stratification
- Estimation of special parameters
 - Familial relative risk
 - Risk penetrance

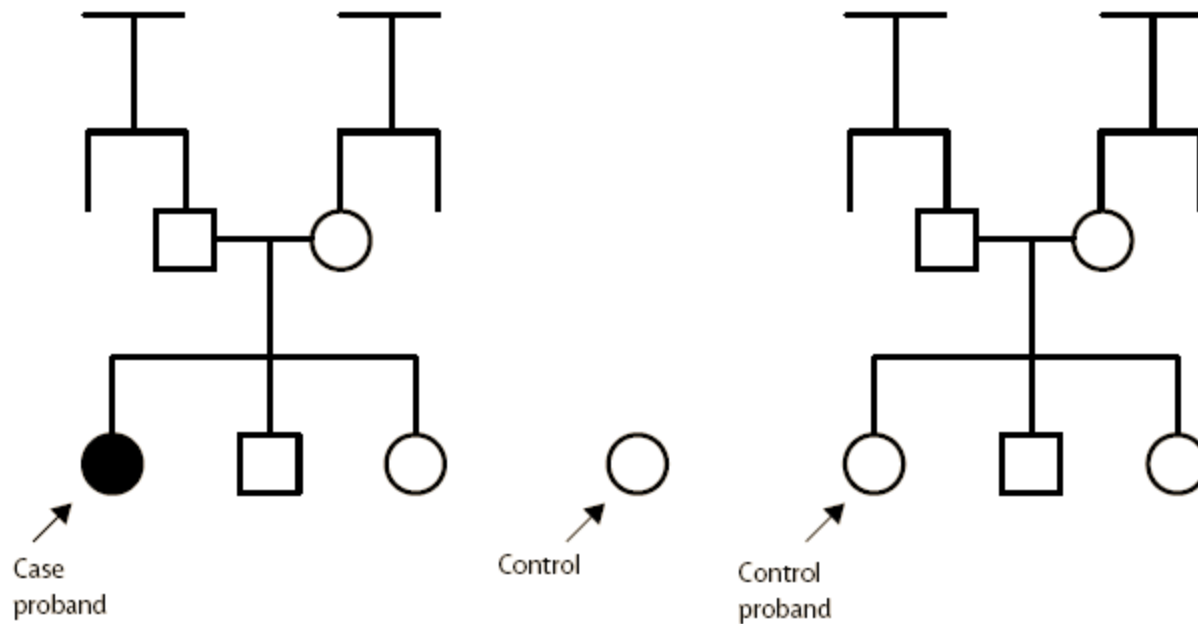
Early family study designs

- Heritability analysis
 - Objective: quantify the fraction of total phenotypic variance attributed to genetic differences
- Linkage analysis
 - Objective: identify genomic regions where genes associated with the phenotype might lie
- At best, identify large chromosomal regions, not specific genes
- Further fine mapping of causal locus required

Family-based association studies

- A twist on a familiar theme: cases + their relatives
 - Family history, e.g., first-degree relative
 - Parent-child trios: compare observed to expected transmission of alleles
 - Extension to siblings, nuclear families, extended pedigrees

Example of family-based study



Advantages of family studies

- Less prone to population stratification
- Rich context for evaluating shared genetic and environmental influences

Disadvantages of family studies

- Difficult to separate shared environmental from genetic influences
- Reduced power due to exclusion of uninformative families
- Challenging for outcomes of older age
- Risk estimates may not apply to general population

Outline

- Learning objectives
- Study designs
 - Overview
 - Case-control studies
 - Cohort studies
 - Randomized/experimental studies
- **The road to Genome-wide association (GWA) studies**
 - Overview
 - Family studies
 - Candidate-gene association studies
 - GWA studies

Candidate gene study -biology

- Driven by current state of knowledge
- Assumptions about genes, SNPs
- Common disease, common variant hypothesis
- One or a few common ($\geq 5\%$) SNPs in one or a few genes, associated with outcome

Candidate gene study -methods

- Started by interrogating known functional regions –promoters, exons
- Increasing knowledge about linkage disequilibrium → tagSNPs
- HapMap
- Concern for false positives moderate
- Problems with replication

Candidate gene studies -examples

- *APOE* and Alzheimer's Disease
- *BRCA* and breast cancer
- *PPARG* and type 2 diabetes

Outline

- Learning objectives
- Study designs
 - Overview
 - Case-control studies
 - Cohort studies
 - Randomized/experimental studies
- **The road to Genome-wide association (GWA) studies**
 - Overview
 - Family studies
 - Candidate-gene association studies
 - **GWA studies**

GWA study -biology

- Robust associations not always with functional variants
- Success of candidate gene approach depended on correct specification of genes
- Early GWA studies identified promising regions that were previously unknown
- “Agnostic” approach

GWA study -methods

- Genotyping platforms developed to look at hundreds of thousands of genes
- Same analysis (and relative risks or odds ratios) as before, but repeated hundreds of thousands of times
- False positive results a major concern
- Statistical adjustment of p-values, replication

A new era for type 2 diabetes (T2D) genetics: genome-wide association studies

Genome-wide association study

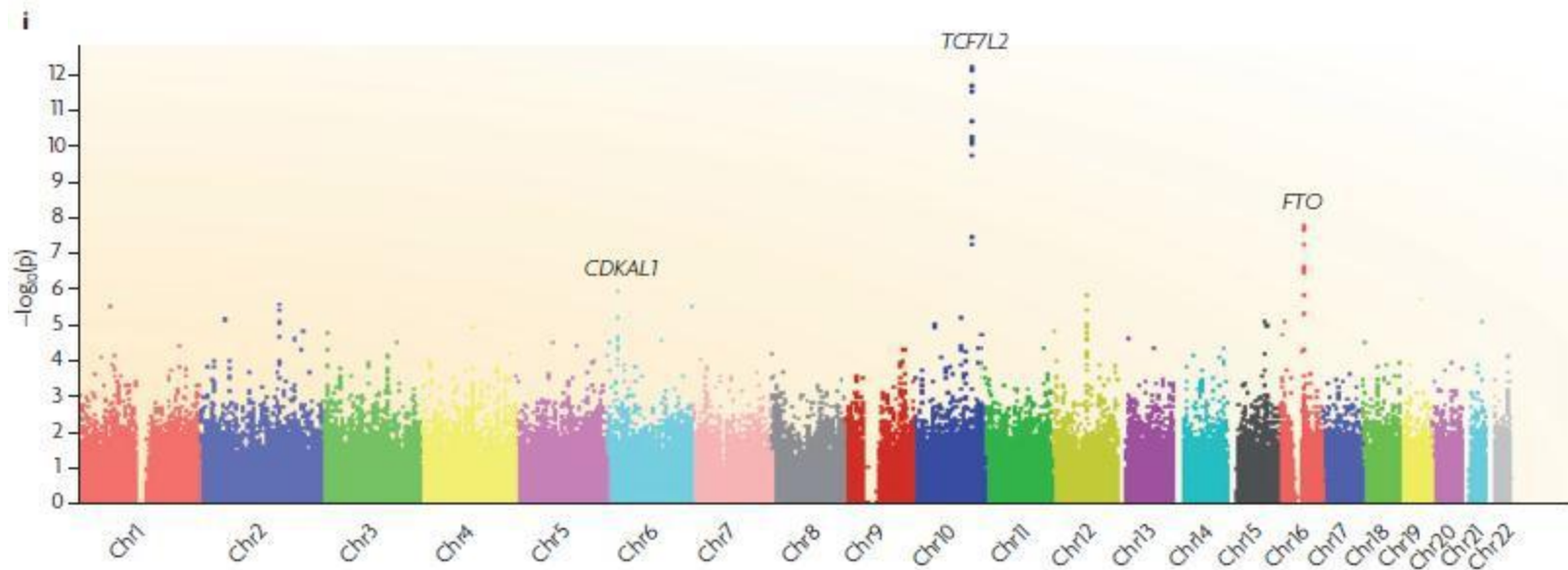
T2D Cases



500K SNP chip



Controls



HuGE Navigator

An integrated, searchable knowledge base of genetic associations and human genome epidemiology.

- Human Genome Epidemiology Navigator
- Literature database of genetic epidemiology associations
 - GWAS
 - Candidate-gene studies
 - Linkage studies
 - Meta-analyses
- Search engine by
 - Phenotype (i.e., T2D)
 - Gene (i.e., *TCF7L2*)



Public Health Genomics and Precision Health Knowledge Base (v7.0)

PHGKB

About

MyPHGKB

Specialized PHGKB



Genomics (A-Z)

Office of Genomics and
Precision Public Health

My Family Health
Portrait

State Public Health
Genomics Programs
Map

Genomics Precision
Health Weekly Scan
(Current Edition)

Advanced Molecular
Detection Weekly Clips
(Current Edition)

Non-Genomics
Precision Health



About HuGE Navigator

HuGE Navigator provides access to a continuously updated knowledge base in human genome epidemiology, including information on population prevalence of genetic variants, gene-disease associations, gene-gene and gene-environment interactions, and evaluation of genetic tests. .

Site citation: W Yu, M Gwinn, M Clyne, A Yesupriya & M J Khoury. [A Navigator for Human Genome Epidemiology](#). *Nat Genet* 2008 Feb;40(2): 124-5.



Phenopedia

Look up genetic associations and human genome epidemiology summaries by disease



Genopedia

Look up genetic associations and human genome epidemiology summaries by gene.

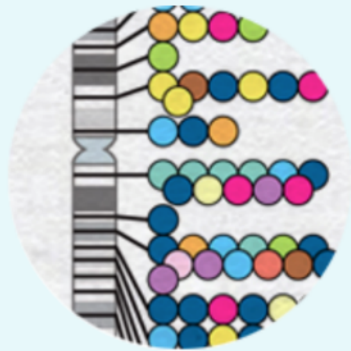


HuGE Literature Finder



Variant Name Mapper

Map common names and rs numbers of genetic



GWAS Catalog

The NHGRI-EBI Catalog of human genome-wide association studies



Examples: [breast carcinoma](#), [rs7329174](#), [Yao](#), [2q37.1](#), [HBS1L](#), [6:16000000-25000000](#)

- Contains information on genetic associations from published GWA studies
 - >100,000 tag-SNPs
 - $P < 10^{-5}$
- Information on: author, title, journal, PubMed link, phenotype, SNP, risk allele and frequency, sample size, country, P-value, magnitude of association, genetic locus, etc.

GWAS Catalog

<http://www.genome.gov/gwastudies/>

Date Added to Catalog (since 11/25/08)	First Author/Date/ Journal/Study	Disease/Trait	Initial Sample Size	Replication Sample Size	Region	Reported Gene(s)	Mapped Gene(s)	Strongest SNP-Risk Allele	Context	Risk Allele Frequency in Controls	P-value	OR or beta-coefficient and [95% CI]	Platform [SNPs passing QC]	CNV
08/07/12	Perry JR May 31, 2012 <i>PLoS Genet</i> Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases.	Type 2 diabetes	2,112 lean type 2 diabetes cases, 4,123 obese type 2 diabetes cases, 54,412 controls	2,881 lean type 2 diabetes cases, 8,702 obese type 2 diabetes cases, 18,957 controls	10q25.2 10q25.2	<i>TCF7L2</i> <i>TCF7L2</i>	<i>TCF7L2</i> <i>TCF7L2</i>	rs7903146-T rs7903146-T	intron intron	0.29 0.29	2×10^{-40} (Lean) 4×10^{-21} (Obese)	1.58 [1.47-1.68] 1.26 [1.20 - 1.32]	NR	N
12/17/11	Kho AN November 19, 2011 <i>J Am Med Assoc</i> Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study.	Type 2 diabetes	2,413 European ancestry cases, 810 African American cases, 2,392 European ancestry controls, 873 African American controls	NR	10q25.2	<i>TCF7L2</i>	<i>TCF7L2</i>	rs7903146-T	intron	0.26 (EA+AA)	2×10^{-15}	1.46 [NR]	Illumina [NR]	N
07/12/10	Voight BF June 27, 2010 <i>Nat Genet</i> Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.	Type 2 diabetes	8,130 European ancestry cases, 38,987 European ancestry controls	Up to 34,412 European ancestry cases, 59,925 European ancestry controls	10q25.2	<i>TCF7L2</i>	<i>TCF7L2</i>	rs7903146-T	intron		2×10^{-51}	1.4 [1.34-1.46]	Affymetrix & Illumina [2,426,886] (imputed)	N
05/07/09	Takeuchi F April 29, 2009 <i>Diabetes</i>	Type 2 diabetes	519 Japanese ancestry	5,629 Japanese ancestry cases, 7,370 Japanese ancestry	10q25.2	<i>TCF7L2</i>	<i>TCF7L2</i>	rs7903146-T	intron	0.04	8×10^{-12}	1.54 [1.36-1.74]	Illumina [482,625]	N

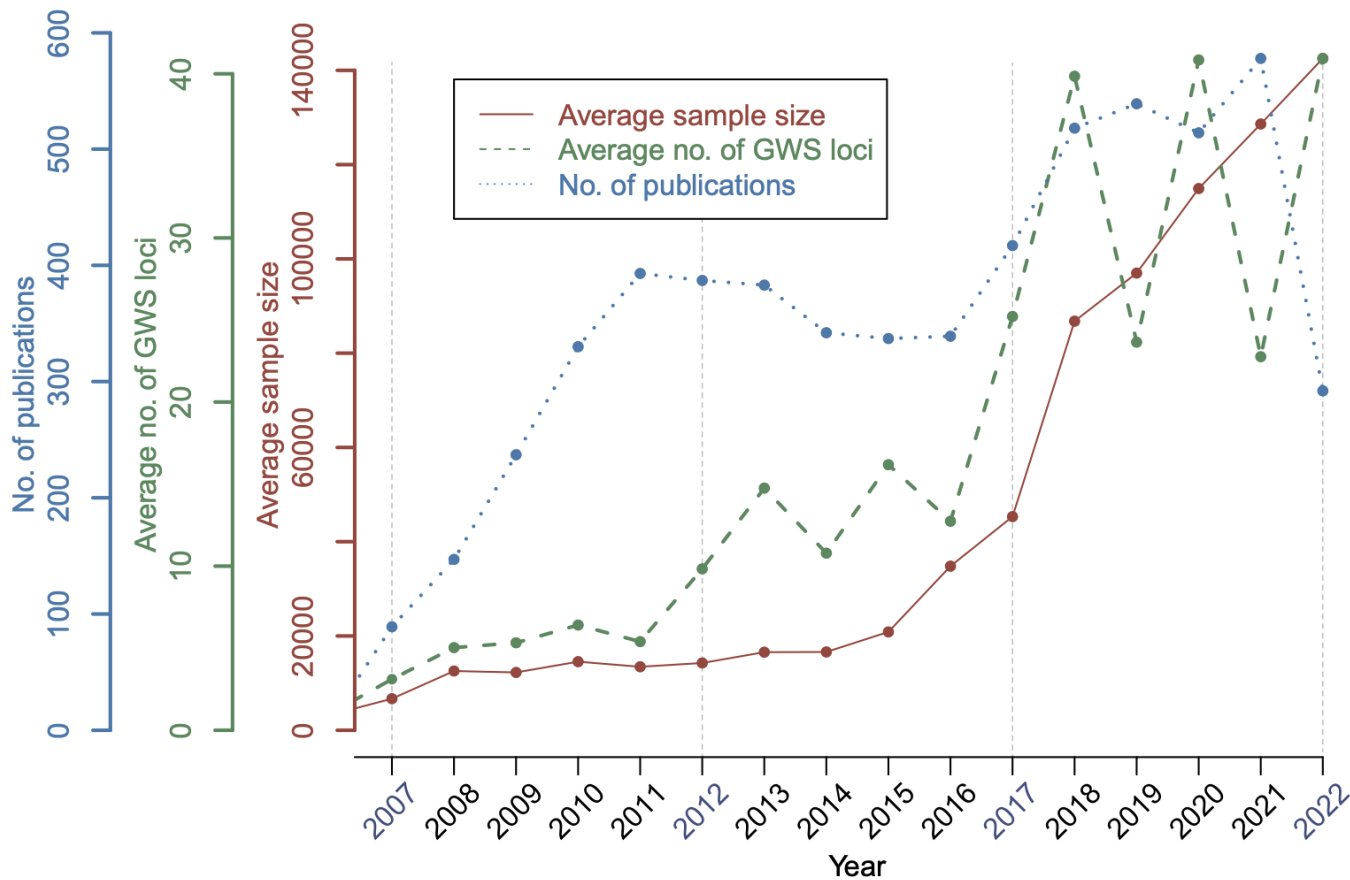


Figure 1. Average sample size and average number of genome-wide significant (GWS) loci per publication for each year during the 15 years history of GWAS discoveries

The data were extracted from 5,771 GWAS publications that used a genome-wide genotyping array and shared their summary statistics on GWAS Catalog before November 8, 2022.

