Introduction to Genome-wide Association Studies (GWAS)

Teresa Ferreira and Nilufer Rahmioglu

Wellcome Centre for Human Genetics/Big Data Institute

GWAS Course Overview

- Session 1 (9:30-10:30) Historical Perspective and Concepts
- Session 2 (10:30-11:30) Steps involved in Genome-wide Association Studies

Break: 11:30-12:00

- Session 3 (12:00-13:00): Post-GWAS Analyses: Characterising GWAS Loci

Lunch-break: 13:00-14:00

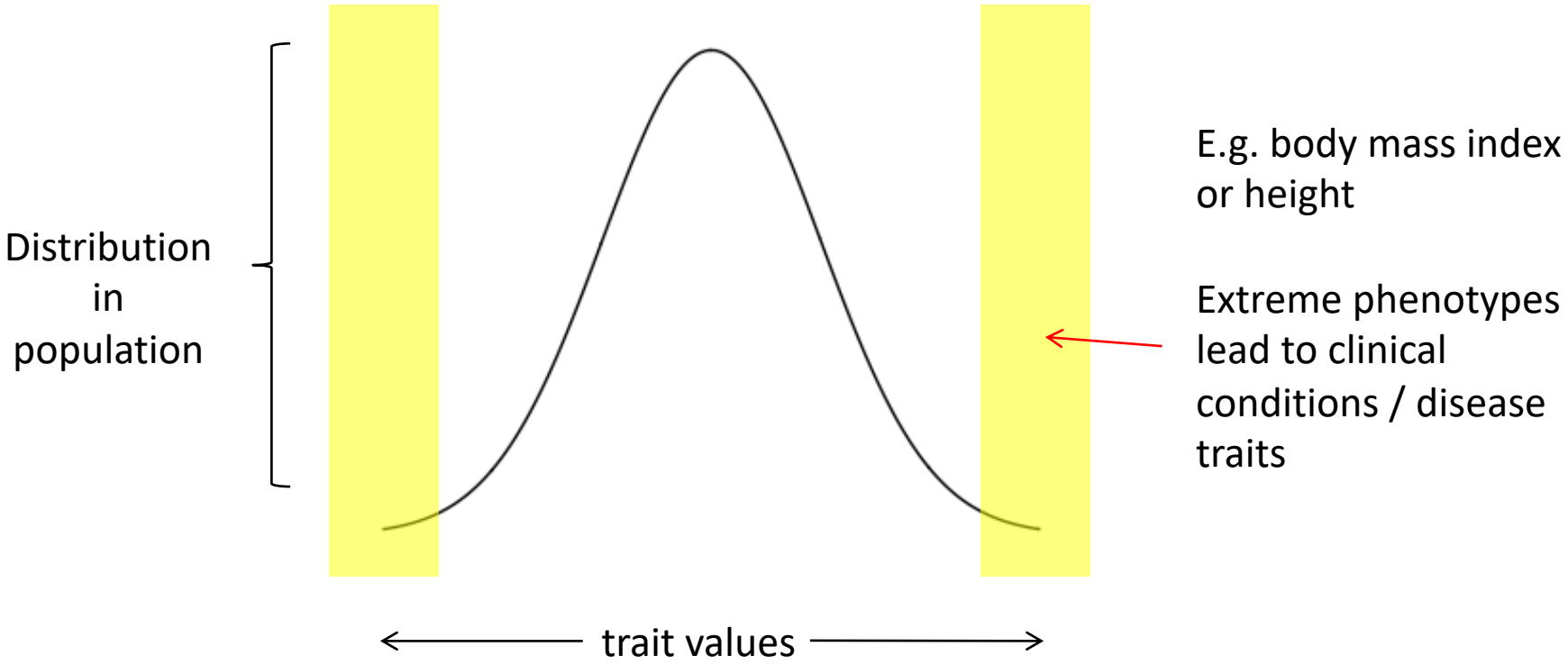
- Session 4 (14:00-16:30): Practical

Session 1: Historical Perspective and Concepts

Learning Objectives

- Understand the definition of a complex trait
- The concepts behind identification of disease variants
- Historical perspective on studying the human genome
- How to study the underlying genetic variants for Monogenic Disease? Linkage studies
- How to study the underlying genetic variants for Complex Disease? Association studies
- GWAS: For identification of common variants for complex traits

A complex trait

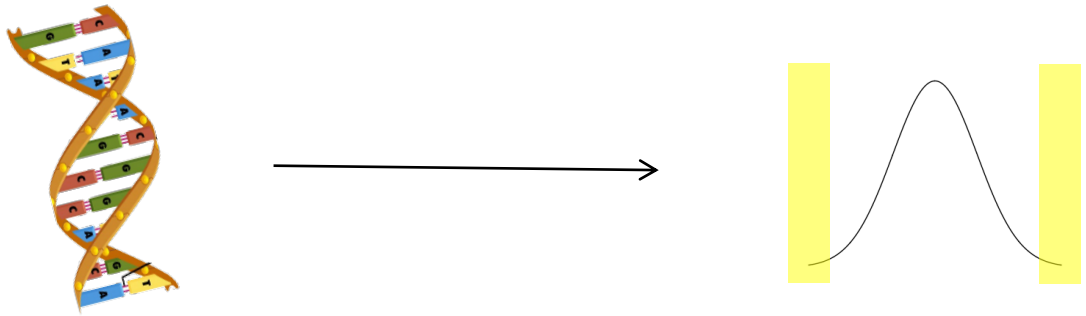


Variation due to age, sex, environmental factors (e.g. diet), and **genetic variation**. May be an effect of **multiple common variants that slightly alter normal physiological processes**.

Why find “disease genes”?

Genetic factors are particularly interesting because (unlike environmental factors) they are:

- inherited at birth
- essentially unchanging
- (often) easily measurable



This makes inferences about *causation* particularly simple.

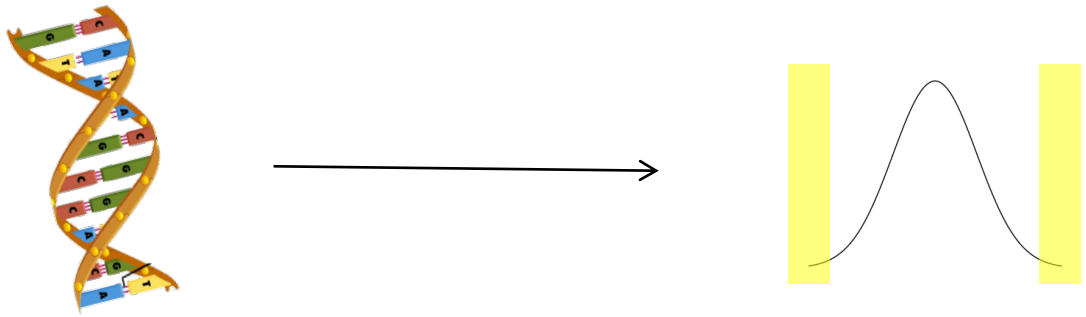
e.g. compare:
“*cyclists tend to be taller*” => causation could plausibly work either way
“*people with genotype AA tend to be taller*” => implies causation (all else being equal)

Inheritance = nature’s randomised control trial

Why find “disease genes”?

Genetic factors are particularly interesting because (unlike environmental factors) they are:

- inherited at birth
- essentially unchanging
- (often) easily measurable

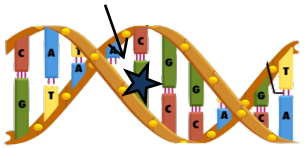


Reasons to look for disease genes:

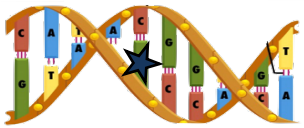
- Identify drug targets
- Predict risk of disease
- Personalised medicine (e.g. stratified by likely treatment response)
- Gene therapy?
- etc...
- ...understand the biology of disease

The circle of genetic causation

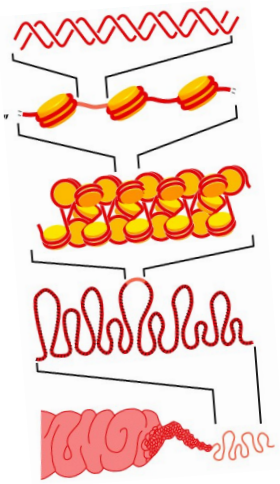
(a causal mutation)



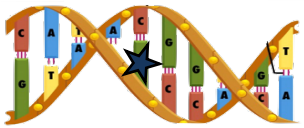
The circle of genetic causation



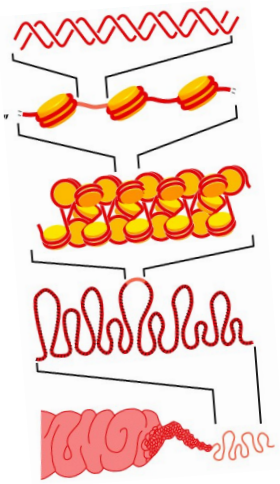
DNA gets physically packaged up into chromosomes...



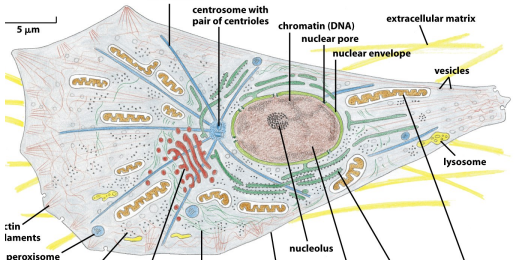
The circle of genetic causation



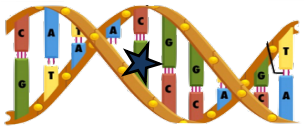
DNA gets physically packaged up into chromosomes...



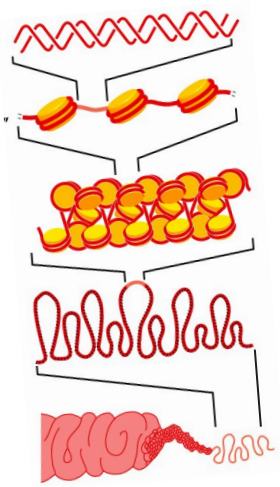
...inside cells, where it is **transcribed** to form proteins and other molecules...



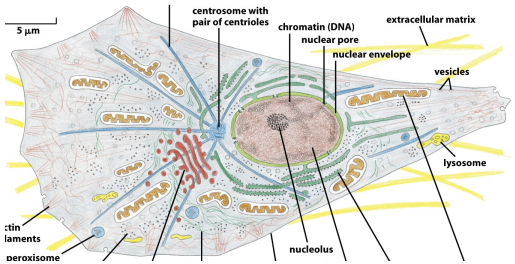
The circle of genetic causation



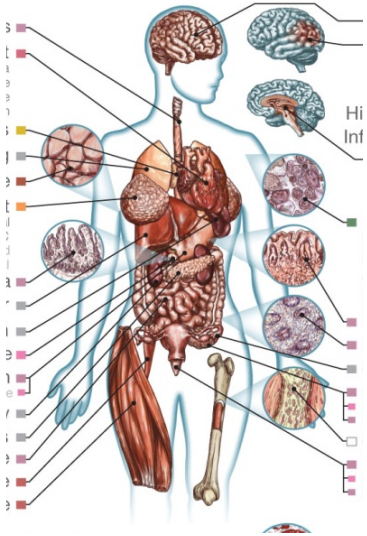
DNA gets physically packaged up into chromosomes...



...inside cells, where it is **transcribed** to form proteins and other molecules...

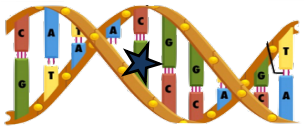


...that affect how the cells behave, forming different organs...

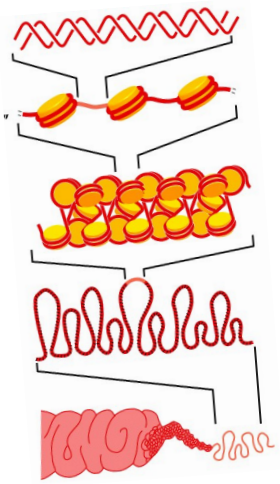


...that combine to make individuals...

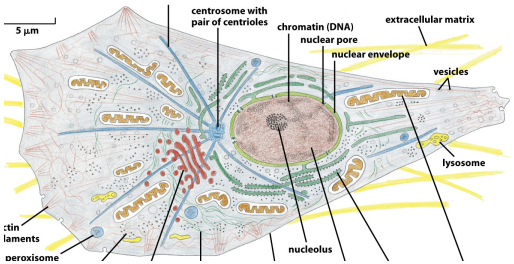
The circle of genetic causation



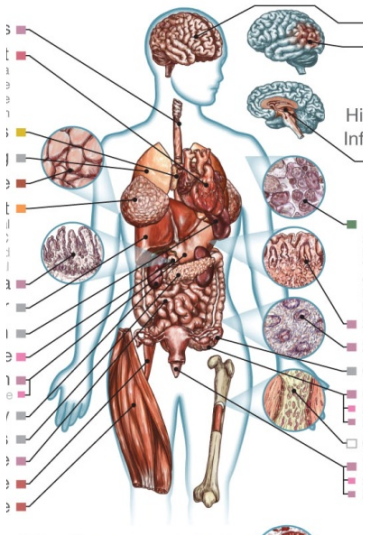
DNA gets physically packaged up into chromosomes...



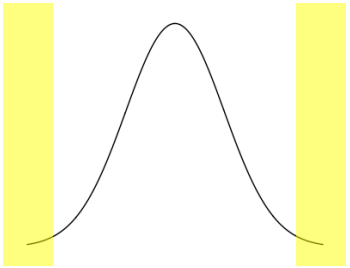
...inside cells, where it is **transcribed** to form proteins and other molecules...



...that affect how the cells behave, forming different organs...

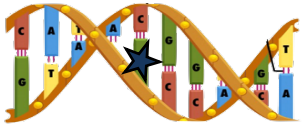


...that combine to make individuals...

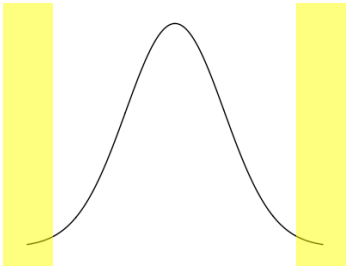


...whose success is affected by the traits they have...

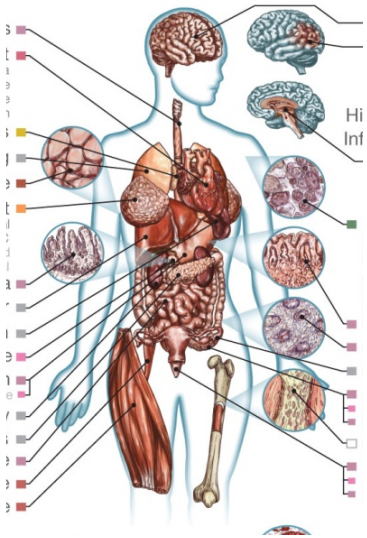
The circle of genetic causation



...passing on DNA, with **mutations** and **recombination**, to new generations...

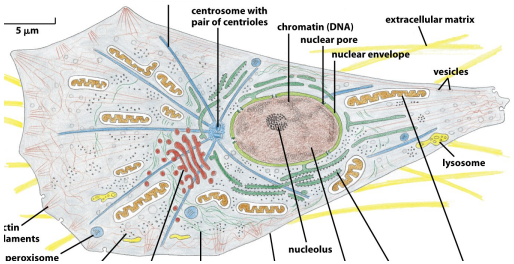


...whose success is affected by the traits they have...

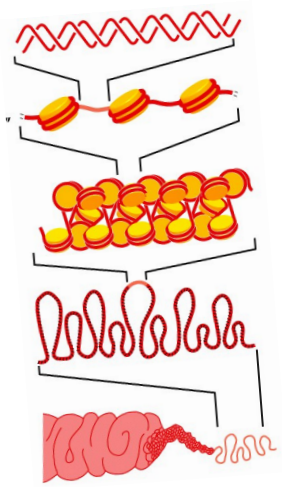


...that combine to make individuals...

...that affect how the cells behave, forming different organs...



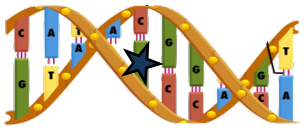
...inside cells, where it is **transcribed** to form proteins and other molecules...

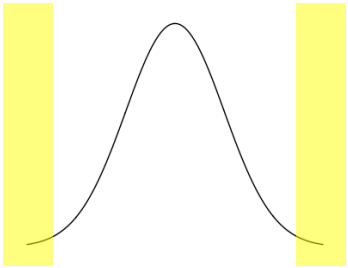


...that gets physically packaged up into chromosomes...

The circle of genetic causation

...passing on DNA, with **mutations** and **recombination**, to new generations...



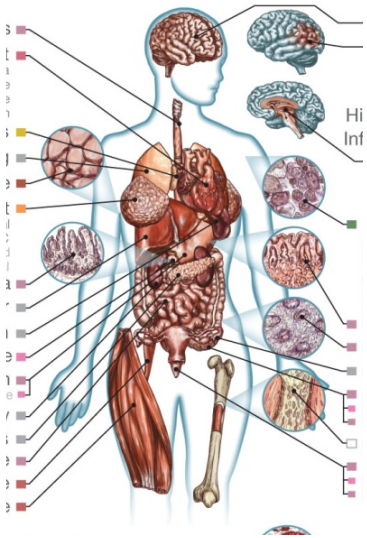


...whose success is affected by the traits they have...

RFLP, microarrays, genome sequencing

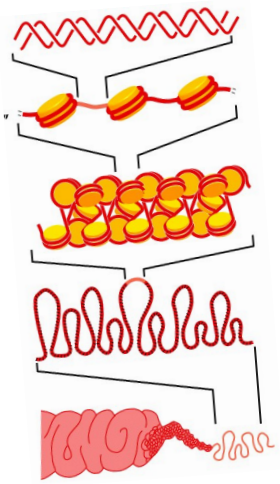
Clinical phenotype measurements

...that gets physically packaged up into chromosomes...



All of this can now be measured

Biomarker measurements

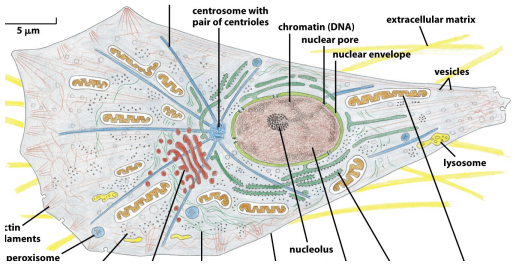


Chromatin state marker assays, ChIP-seq, ...

RNA-seq, spectroscopy, antibody binding

...that combine to make individuals...

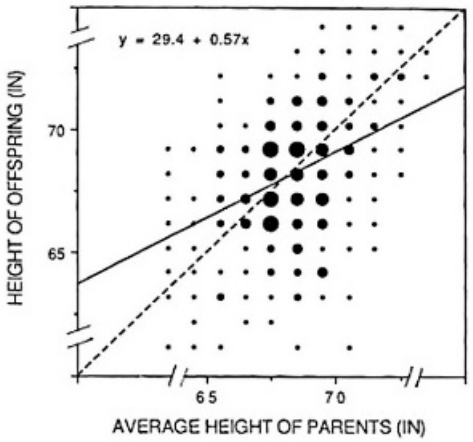
...inside cells, where it is **transcribed** to form proteins and other molecules...



...that affect how the cells behave, forming different organs...

Finding disease genes in practice

- I'm going to assume we've got a trait that we've established is heritable



← Demonstration by Francis Galton that human height is heritable (height of parents predicts height of offspring).

We want to find genetic variants influencing it. How?

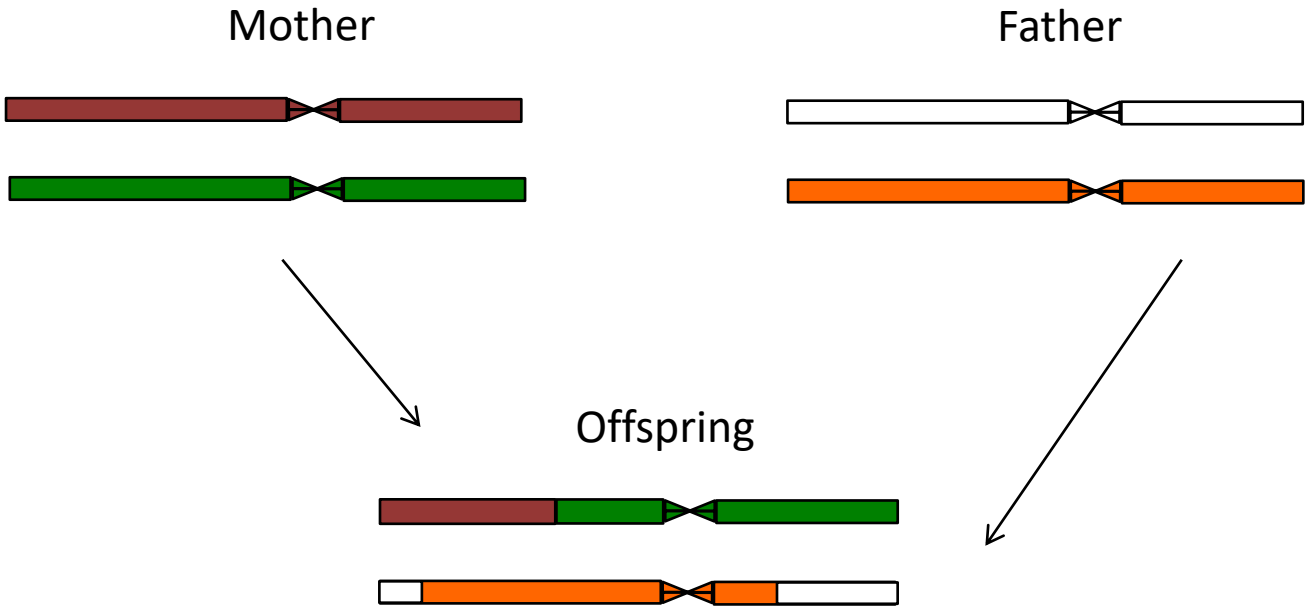
Finding needles in the haystack

The human genome is 3.2 billion base pairs long

We want to find a small number of ‘causal’ genetic mutations in there. How?

Luckily, nature has given us a way to narrow down on specific regions of the genome.

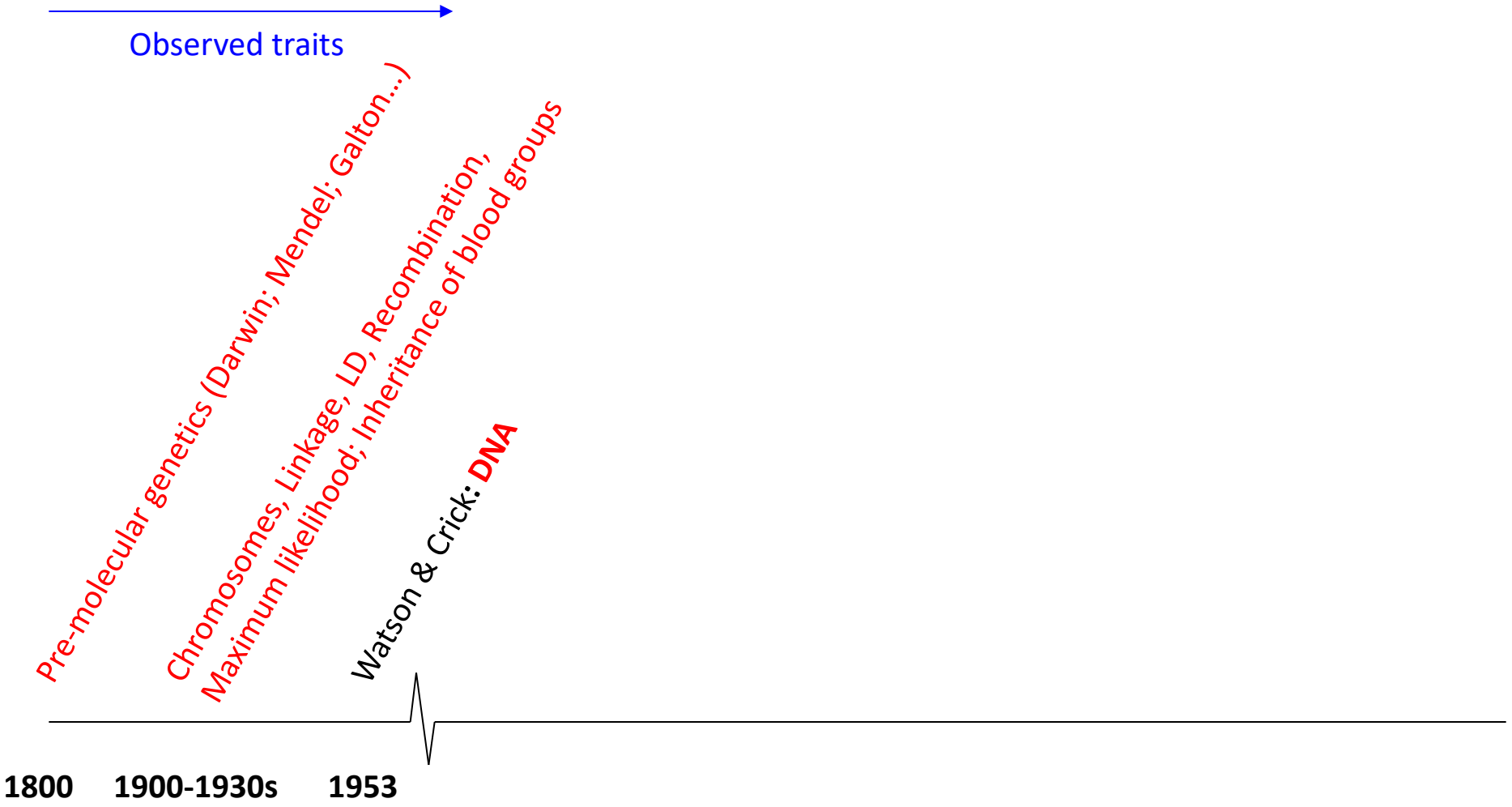
Recombination



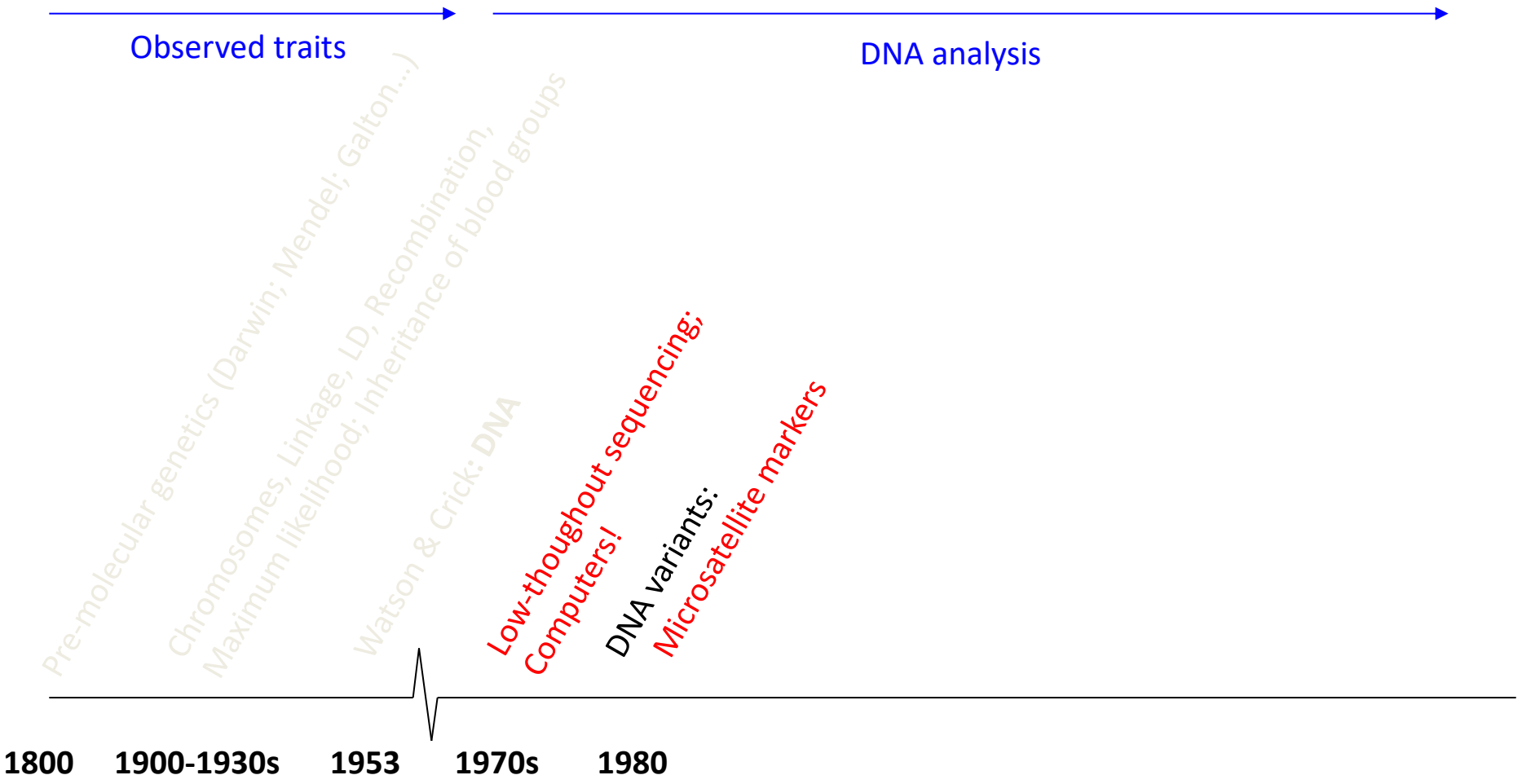
Genetic recombination breaks up the DNA into segments.
You inherit a mosaic of segments of your parents' DNA.

Recombination = nature's magnifying glass

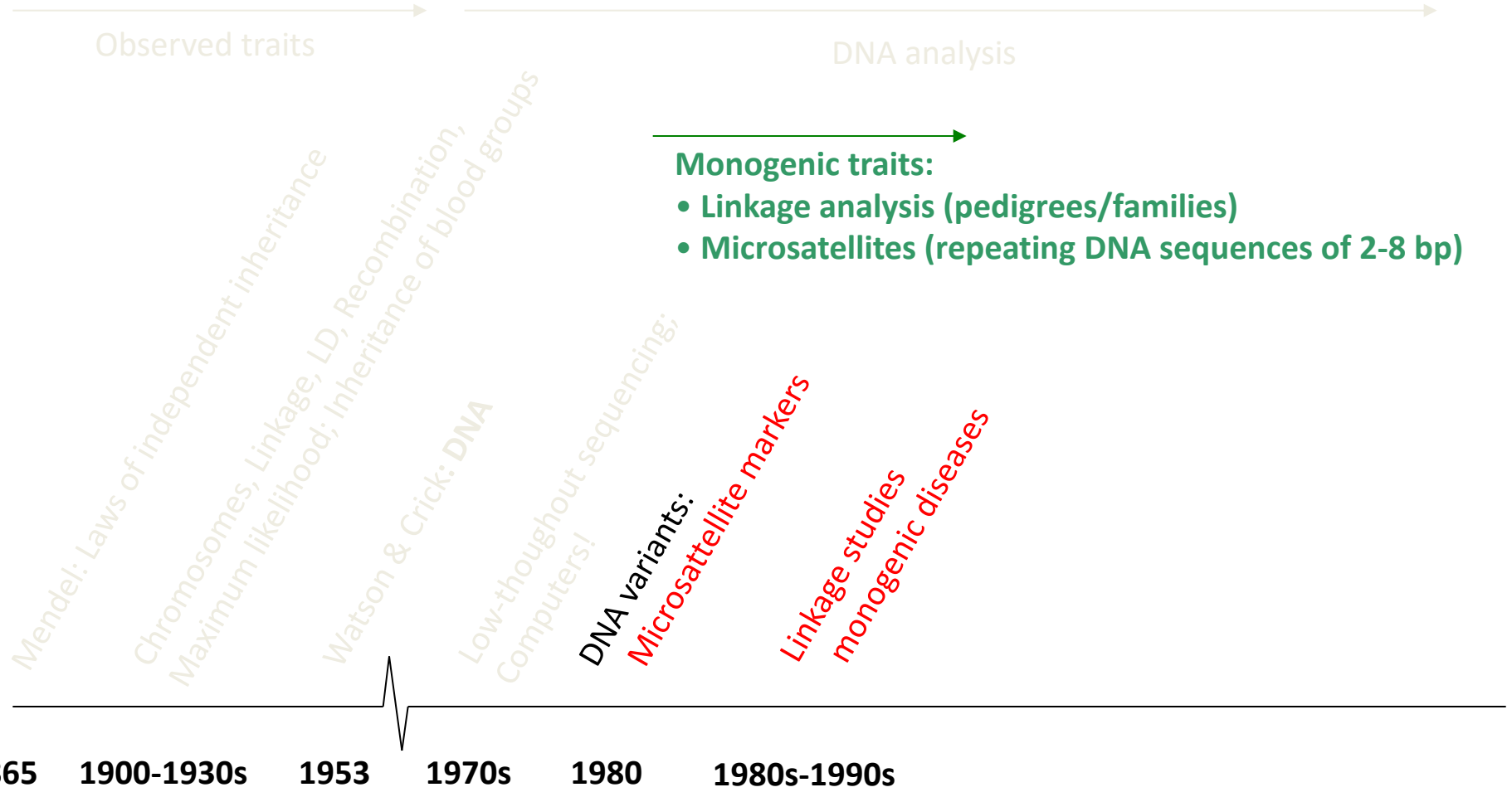
Studying the human genome



Studying the human genome



Studying the human genome



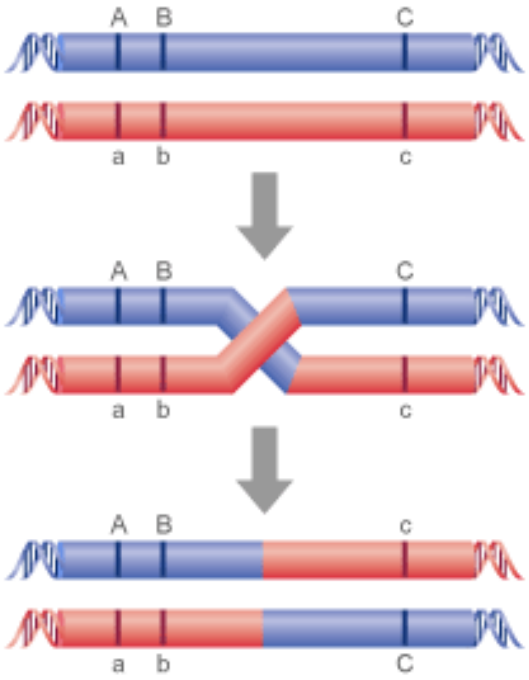
‘Mendelian’ vs. Complex diseases (traits)

Mendelian (mono-genic)

- Caused by single, **rare** mutations that confer a ***very high risk***
- Disease primarily identified in families
- Examples: Huntington’s chorea; Cystic fibrosis
- Typical study design: Family-based **Linkage Studies**

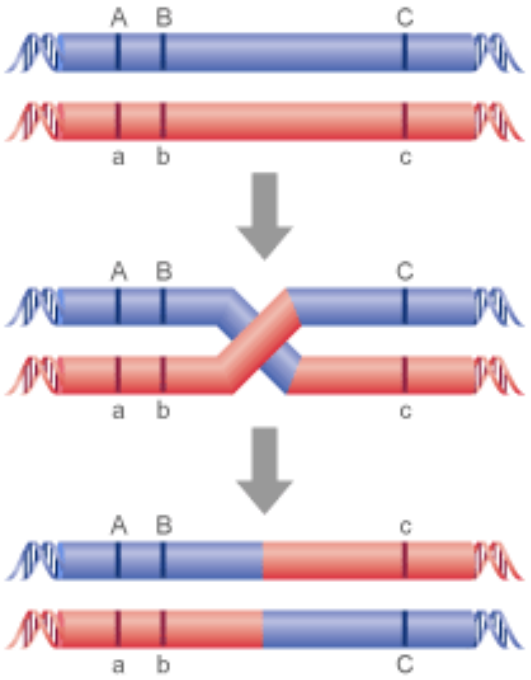
Linkage Study

- Aims to find a (large) section of chromosome that harbours a disease-predisposing gene variant, across ***families***
- Utilises chromosomal recombination events during meioses across generations



Linkage Study

- Aims to find a (large) section of chromosome that harbours a disease-predisposing gene variant, across ***families***
- Utilises chromosomal recombination events during meioses across generations



Process:

- Genotype ~ 400 microsatellite markers across the genome in families with multiple affecteds;
- Compare whether the allelic status of these variants are shared between affecteds in a family;
- Sum evidence of sharing across families - calculate if sharing is significantly more than expected under Mendelian laws of inheritance → evidence for *linkage*

Linkage Study

- Very successful for monogenic disorders, when carrying a variant means you develop disease (e.g. Cystic fibrosis, Huntingdon's chorea)

  Low success rate for complex diseases:

  caused by >1 gene; variant=susceptibility; environment

- Most powerful to detect variants responsible for disease in families, but otherwise rare in the general population

‘Mendelian’ vs. Complex diseases (traits)

Mendelian (mono-genic)

- Caused by single, **rare** mutations that confer a ***very high risk***
- Disease primarily identified in families
- Examples: Huntingdon’s chorea; Cystic fibrosis
- Typical study design: Family-based **Linkage Studies**

Complex

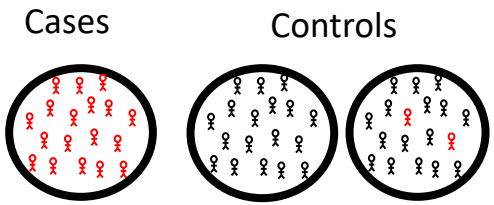
- Caused by multiple **common** genetic variants + environmental factors, each conferring a ***modest increase in susceptibility***
- Disease observed in families as well as sporadic cases
- Examples: Cancers, CVD, Type I/II Diabetes, Body Mass Index.....
- Typical study design: **Association Studies**

NB Sub-types of complex diseases can behave as Mendelian traits,
e.g. Familial breast and ovarian cancer caused by *BRCA1/2* mutations.....

Association Study

*

Case-control



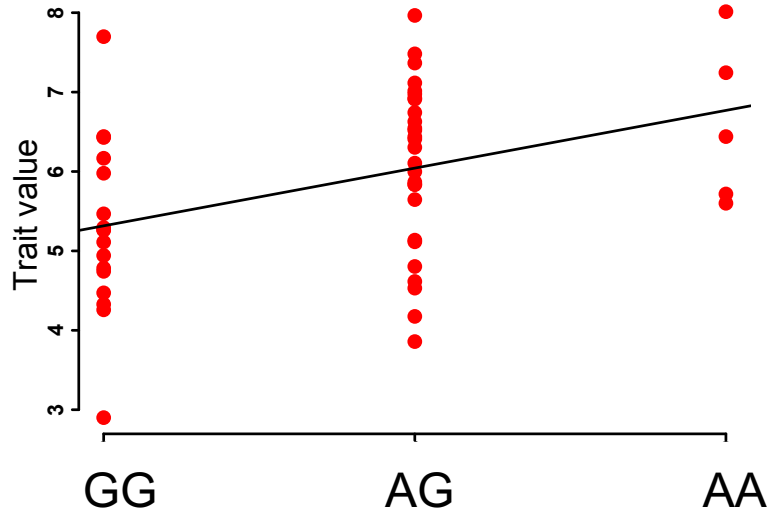
SNP allele			
	A	0.17	0.15
	G	0.83	0.85
SNP genotype			
	AA	0.03	0.03
	AG	0.28	0.24
	GG	0.69	0.73

- Chisquare tests
- Logistic regression
- Significance value (P)
- ORs + CIs

- Uses single DNA base pair changes (SNPs)
- Aims to find a common *disease-predisposing* genetic variant (or get close by examining a SNP nearby)
- Powerful method for complex disease gene mapping, provided genetic variant causing disease is not rare

Association Study

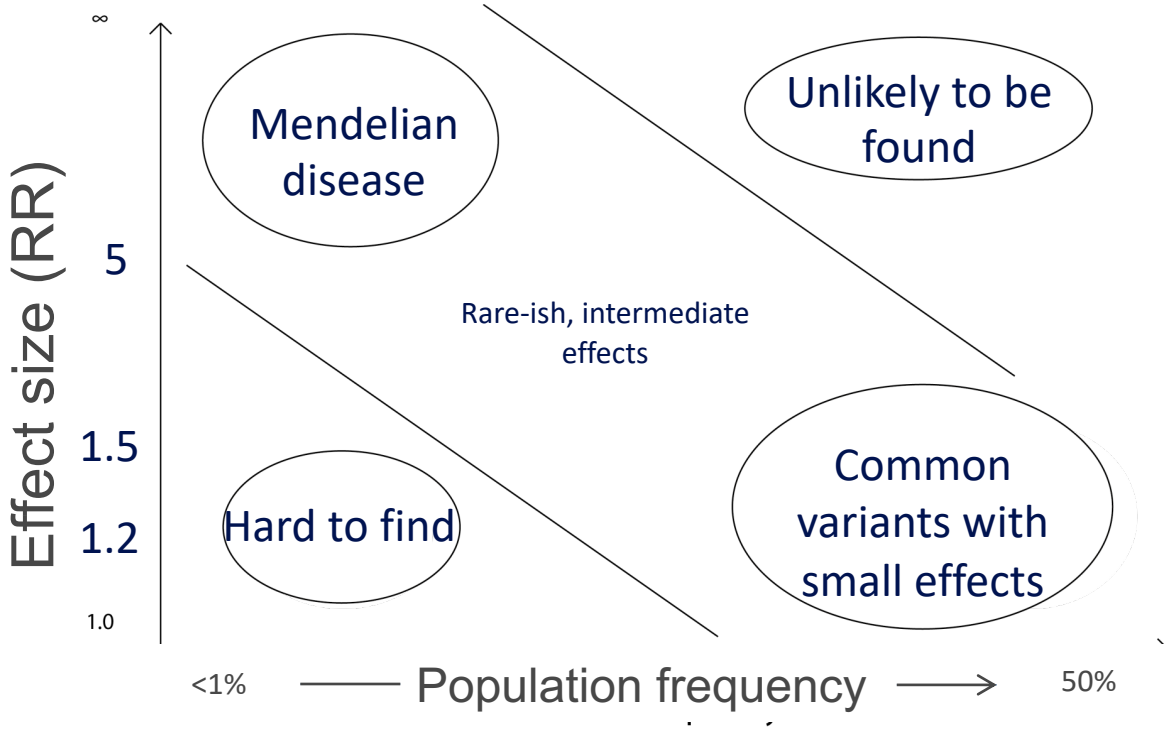
Continuous trait



- T-test/ANOVA
- Linear regression
- Significance value (P)
- Mean values +SEs/CIs ; Betas + SEs/CIs

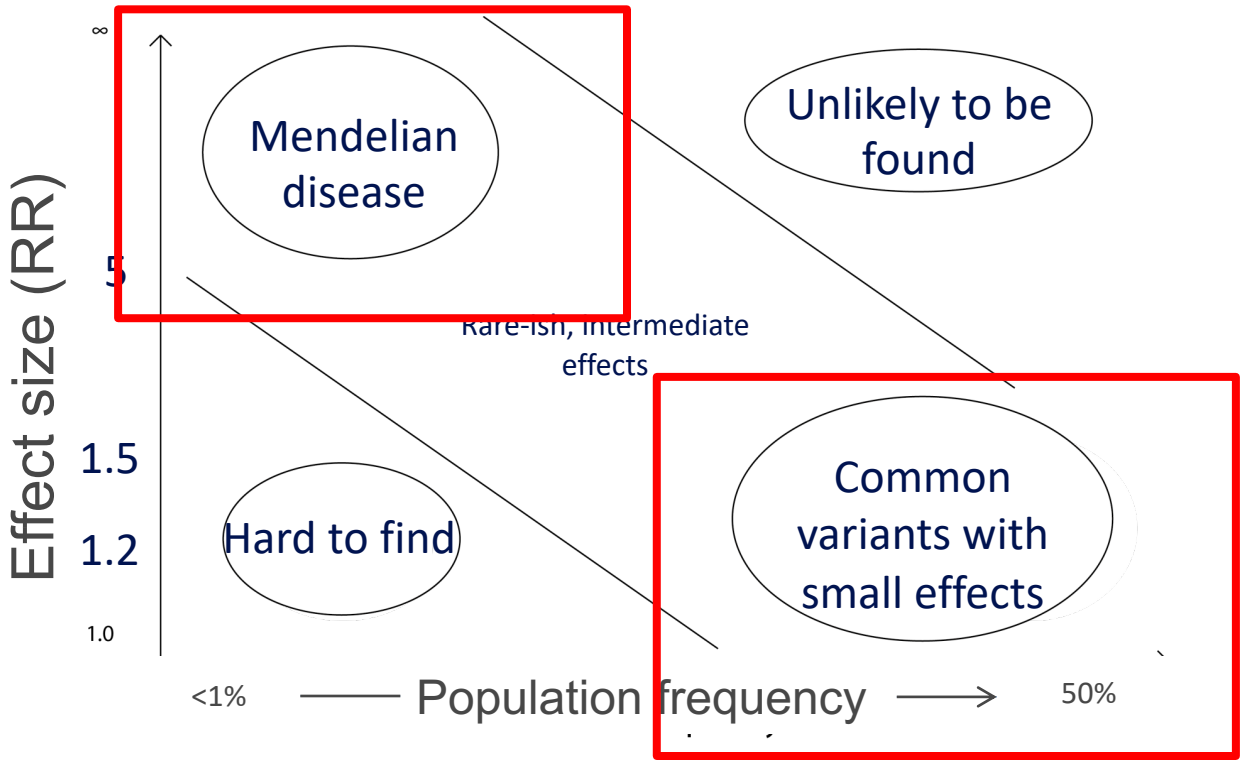
- Uses single DNA base pair changes (SNPs)
- Aims to find a common *trait-affecting* genetic variant (or get close by examining a SNP nearby**)**
- Powerful method for complex trait mapping, provided genetic variant causing disease is not rare

Complex diseases



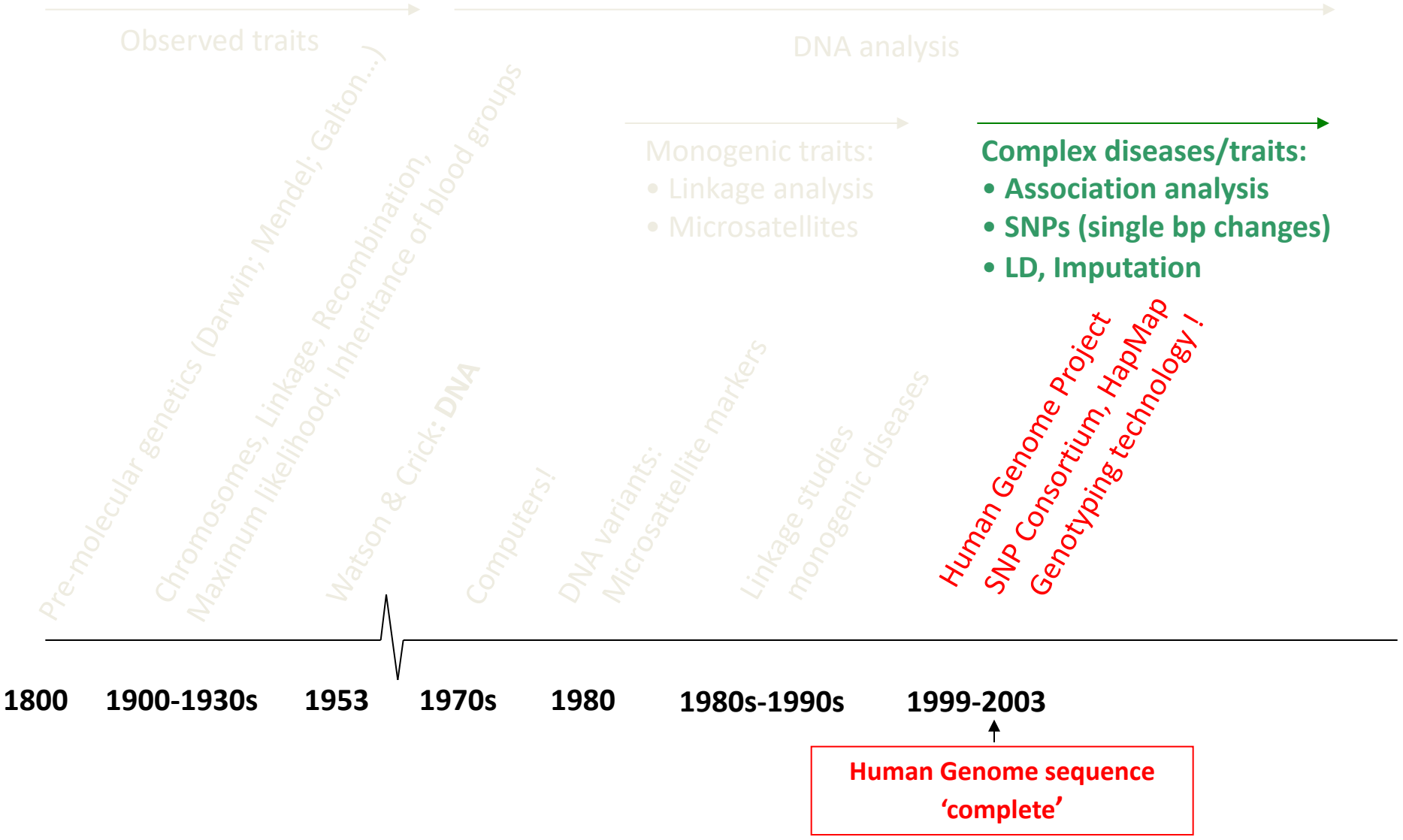
Complex diseases

Where linkage studies are likely to work



Where most complex disease effects are

Studying the human genome



The International HapMap Project: LD

(International HapMap Consortium, Nature 2007; 2009; 2010)

- Common SNPs across the genome (frequency > 5%) are *not independent* when tested in a population, due to common ancestry between individuals, due to **Linkage disequilibrium (LD)**

- SNPs located near each other: the variant status of one predicts ('tags') the other
- The HapMap Project provided a map that shows SNP dependencies across the genome in different ancestral populations (Phase I/II: African-YRI, Asian-CHB/JPT, European– CEU; Phase III: + 9 other)
- We now 'only' need to genotype ~ a few 100K SNPs, to reliably predict the genotype status of all common SNPs in the genome, in a given population
- Means that **when you find an association with a SNP, you have not found the causal variant**, but are most likely 'tagging' the causal variant

Studying the human genome

Observed traits

DNA analysis

Pre-molecular genetics (Darwin; Mendel; Galton…)

Chromosomes, Linkage
Recombination, Maximum likelihood

Watson & Crick: **DNA**

Computers!

DNA variants:
Microsattellite markers

Linkage studies
monogenic diseases

1000 Genomes Project

Monogenic traits:
- Linkage analysis
- Microsatellites

Complex disease/traits:
- **Association analysis**
- **SNPs (single bp changes)**
- **LD, Imputation**

1800 **1900-1930s** **1953** **1970s** **1980** **1980s-1990s** **1999-2003** **2005-2014**

1000G: Sequencing data on 2,500 people, 26 populations

From ~ 2000: Major developments in genetics knowledge and technology

*Utility: GWAS to detect **common variants** underlying complex diseases*

Human Genome Project

- Human genome: ~ 3 billion base-pairs
- SNPs: 1 per 100-300 bp : ~ 10M common SNPs (freq >0.05)

Do we need to genotype 10M SNPs for each individual?

- Unnecessary – reference panel for imputation
- Too costly

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

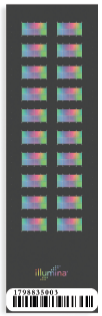
NATURE | VOL 431 | 21 OCTOBER 2004 | www.nature.com/nature

A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms

The International SNP Map Working Group*

NATURE | VOL 409 | 15 FEBRUARY 2001 | www.nature.com

NATURE | VOL 426 | 18/25 DECEMBER 2003



High through-put genotyping/sequencing

NATURE | VOL 526 | 1 OCTOBER 2015

Technology improvements

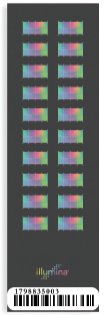
Cost and coverage

- Affymetrix 100K
- Affymetrix 500K
- Affymetrix 6.0 (~1M SNPs)

…

- Illumina 650Y
- Affymetrix UK Biobank array
- Illumina 1M
- Illumina 2.5M
- Illumina 5M

…



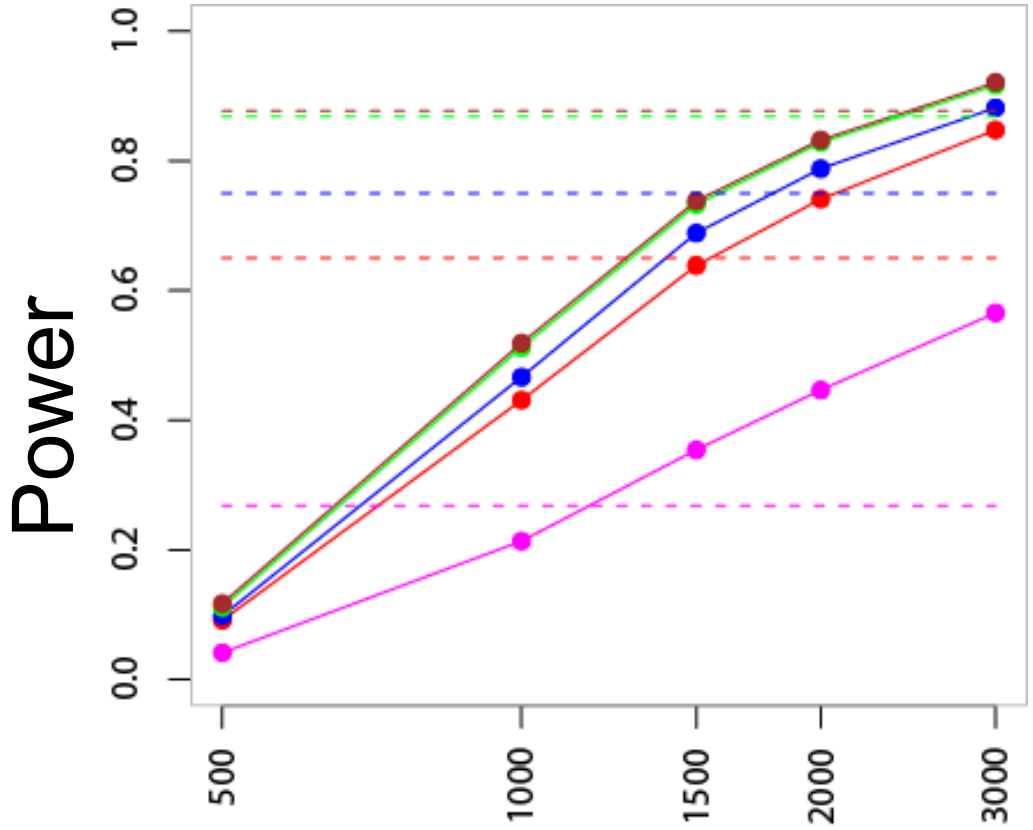
Costs are decreasing with time.

Example

- The UK Biobank has typed ~500,000 individuals on the Affymetrix UK Biobank array (containing ~800k SNPs).

- This array might now cost ~£20 per sample. So this project would cost in the order of £10,000,000 for genotyping.

Power to find weak effects



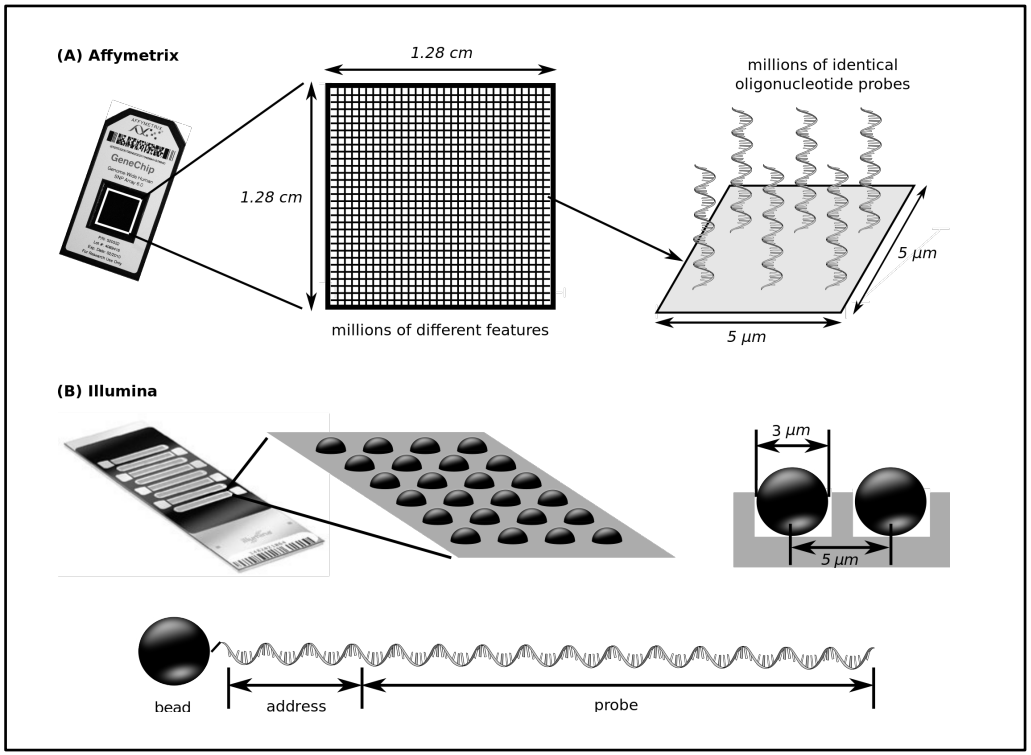
Higher statistical power with higher coverage (and more samples)

- Illumina 650k
- Illumina 550k
- Illumina 300k
- Affymetrix 500k
- Affymetrix 100k

Sample size (number of cases and controls)

Relative risk of 1.2

How a microarray works

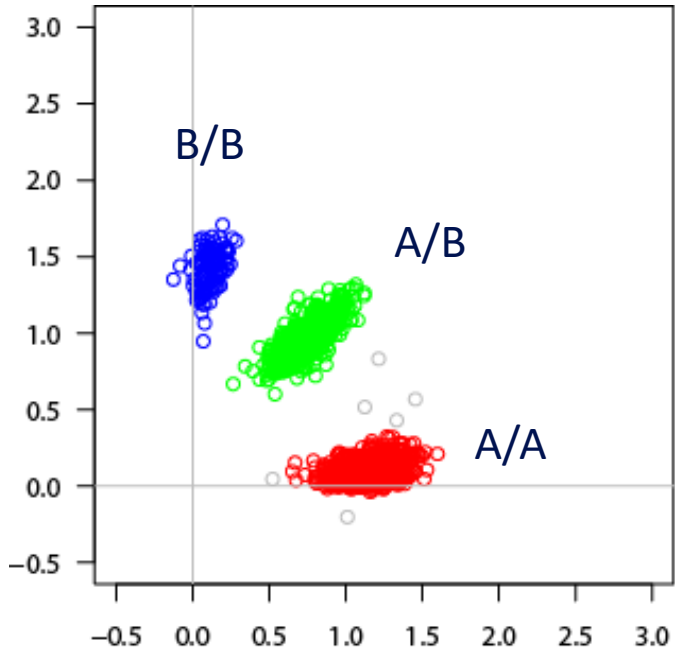


Wash the DNA over and let it hybridise to millions of probes – one for each SNP

Flourescent markers are then attached. A picture is taken of the array.

How a microarray works

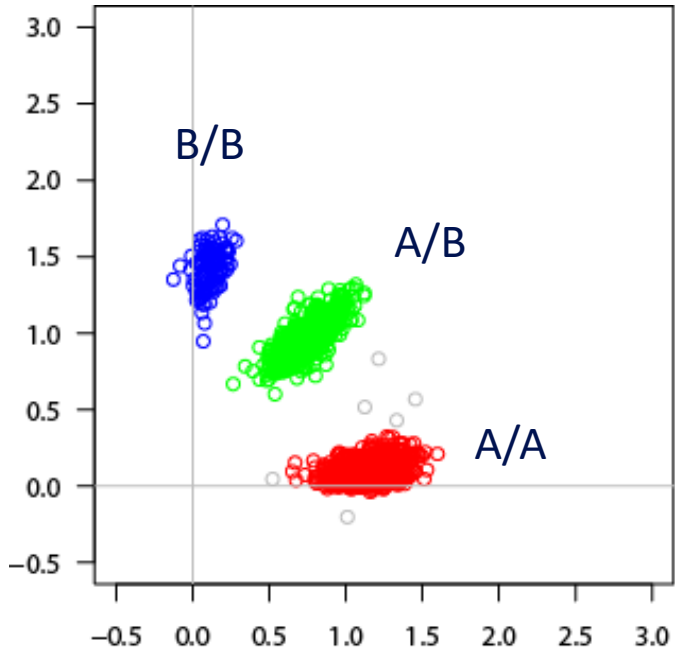
For each SNP, you get back this:



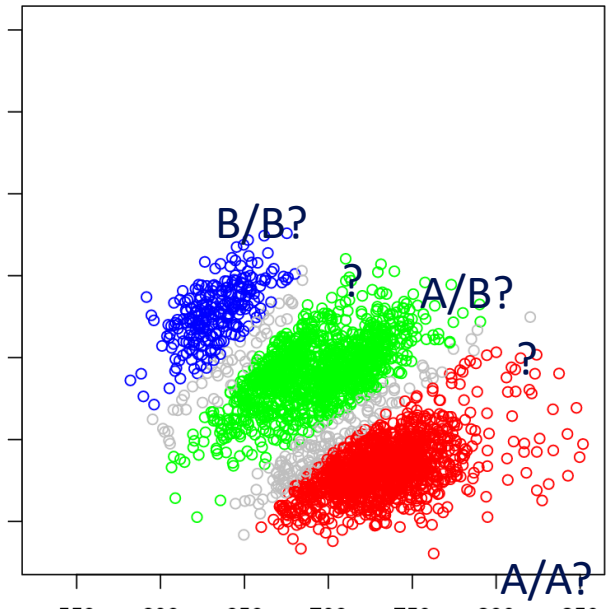
Each dot represents DNA from one individual.
X axis = image intensity for 1st allele probe
Y axis = image intensity for 2nd allele probe

How a microarray works

For each SNP, you get back this:



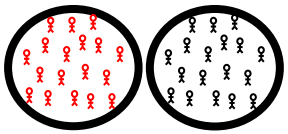
Or this if you're less lucky:



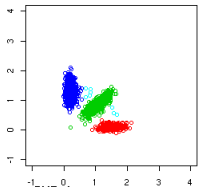
Each dot represents DNA from one individual.
X axis = image intensity for 1st allele probe
Y axis = image intensity for 2nd allele probe

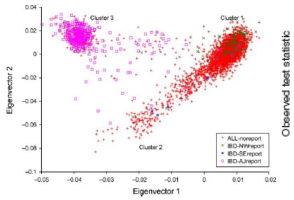
Genome-wide Association Study (GWAS) Recipe

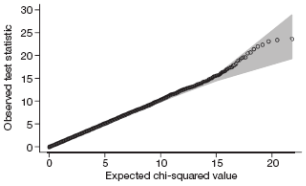
Genotype 100,000s **common** SNPs in 1000s of cases+controls



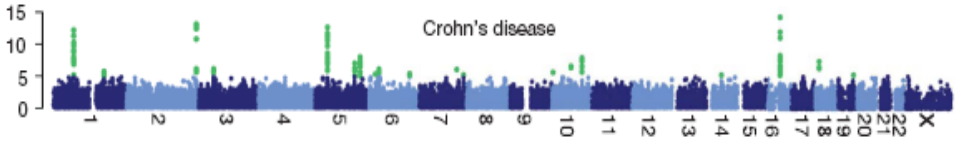
Quality-control analyses: e.g. genotype calling, population biases



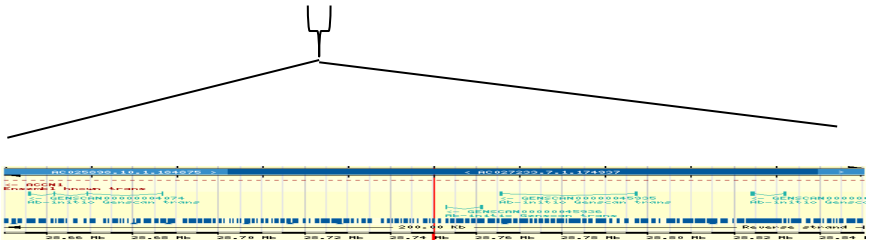




At each SNP test for allele frequency difference between cases& controls (chi^2, logistic regression)

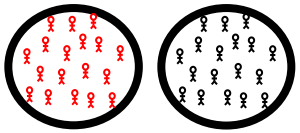


Identify significant associations, nominal p-value (5x 10^{-8})



Assess genomic info: genes, SNP density, regulatory regions, etc

Genotype selected SNPs in different case+control samples of same population: replication/meta-analysis



More on this in Session 2

2005-7: first GWAS emerging....

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*]
Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1]
Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6]
Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3]
Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7†]

SCIENCE VOL 308 15 APRIL 2005

Age-related Macular Degeneration (AMD):
100K SNPs in 96 cases and 50 controls!!

Rs380390, CC/CG vs. GG:
OR = 4.6 (2.0-11.0), $P=4.1 \times 10^{-8}$

2005-7: first GWAS emerging....

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2*] Emily Y. Chew,[3*] Jen-Yue Tsai,[4*] Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

SCIENCE VOL 308 15 APRIL 2005

Age-related Macular Degeneration (AMD):
100K SNPs in 96 cases and 50 controls!!

Rs380390, CC/CG vs. GG:
OR = 4.6 (2.0-11.0), P=4.1 x 10^{-8}

A genome-wide association study identifies novel risk loci for type 2 diabetes

Robert Sladek[1,2,4], Ghislain Rocheleau[1*], Johan Rung[4*], Christian Dina[5*], Lishuang Shen[1], David Serre[1], Philippe Boutin[5], Daniel Vincent[4], Alexandre Belisle[4], Samy Hadjadj[6], Beverley Balkau[7], Barbara Heude[7], Guillaume Charpentier[8], Thomas J. Hudson[4,9], Alexandre Montpetit[4], Alexey V. Pshezhetsky[10], Marc Prentki[10,11], Barry I. Posner[2,12], David J. Balding[13], David Meyre[5], Constantin Polychronakos[1,3] & Philippe Froguel[5,14]

NATURE|Vol 445|22 February 2007

A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene

Richard H. Duerr,[1,2] Kent D. Taylor,[3,4] Steven R. Brant,[5,6] John D. Rioux,[7,8] Mark S. Silverberg,[9] Mark J. Daly,[8,10] A. Hillary Steinhart,[9] Clara Abraham,[11] Miguel Regueiro,[1] Anne Griffiths,[12] Themistocles Dassopoulos,[5] Alain Bitton,[13] Huiying Yang,[3,4] Stephan Targan,[4,14] Lisa Wu Datta,[5] Emily O. Kistner,[15] L. Philip Schumm,[15] Annette T. Lee,[16] Peter K. Gregersen,[16] M. Michael Barmada,[2] Jerome I. Rotter,[3,4] Dan L. Nicolae,[11,17] Judy H. Cho[18*]

SCIENCE VOL 314 1 DECEMBER 2006

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

NATURE|Vol 447|7 June 2007

~2,000 cases for each of 7 diseases + ~3,000 shared controls
Bipolar disorder, CAD, Crohn's, Hypertension, RA, Type I/II Diabetes

ORs ranging from 1.3 – 2.0 (5-15: MHC)

2007: Wellcome Trust Case-Control Consortium

Genome-Wide Association Across Major Human Diseases

DESIGN

Collaboration amongst 26 UK disease investigators
2000 cases each from 7 diseases

GENOTYPING

Affymetrix 500k SNPs + 175k Perlegen highly selected SNPs

CASES

1. Type 1 Diabetes
2. Type 2 Diabetes
3. Crohn's Disease
4. Coronary Heart Disease
5. Hypertension
6. Bipolar Disorder
7. Rheumatoid Arthritis

CONTROLS

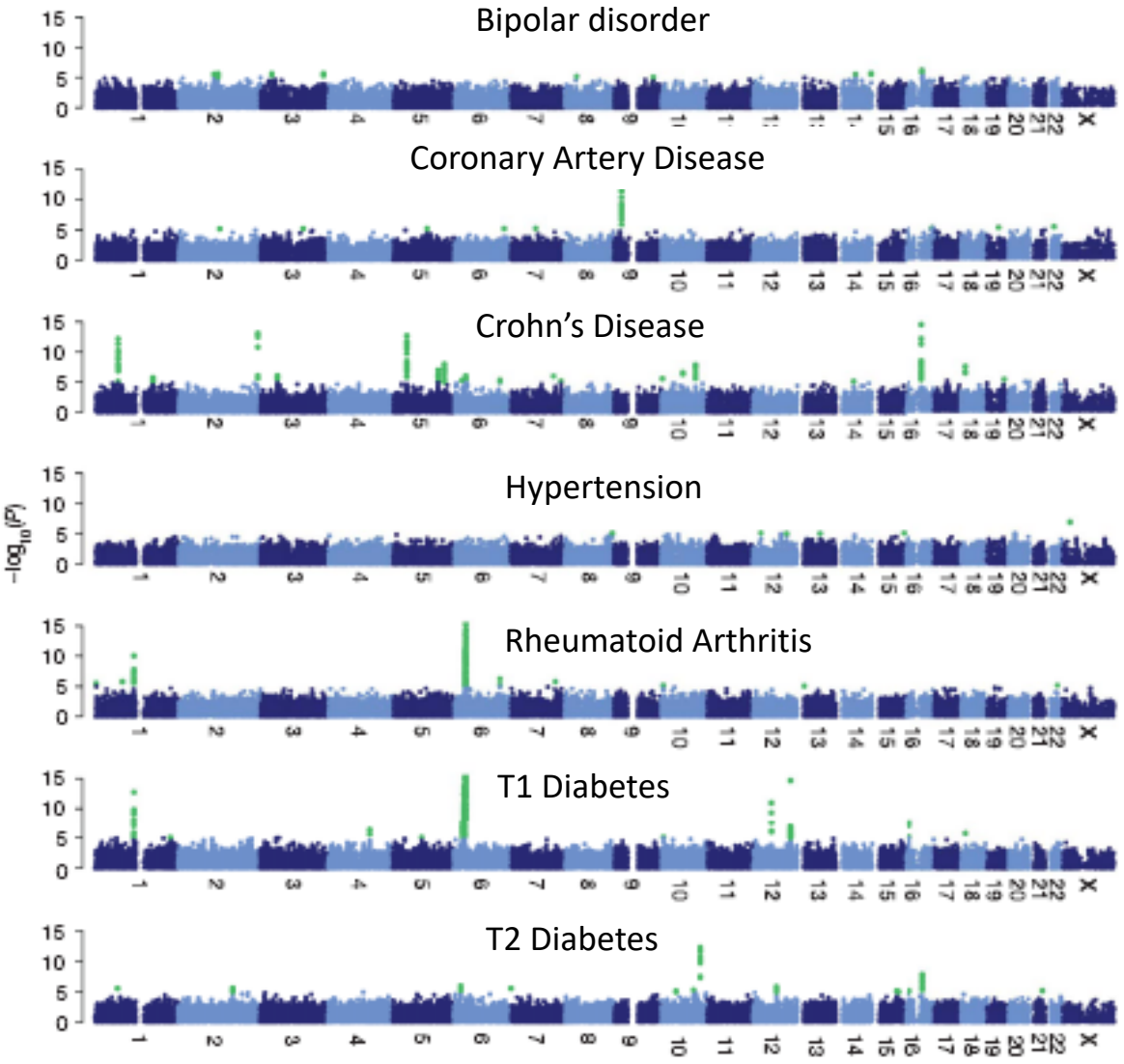
1. UK Controls A (n=1500: 1958 BC)
2. UK Controls B (n=1500: NBS)

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

NATURE|Vol 447|7 June 2007

Wellcome Trust Case-Control Consortium



Significant hits

16p12

9p21

1p31 5p13 10q24
2q37 5q33 16q12
3p21 10q21 18p11

-

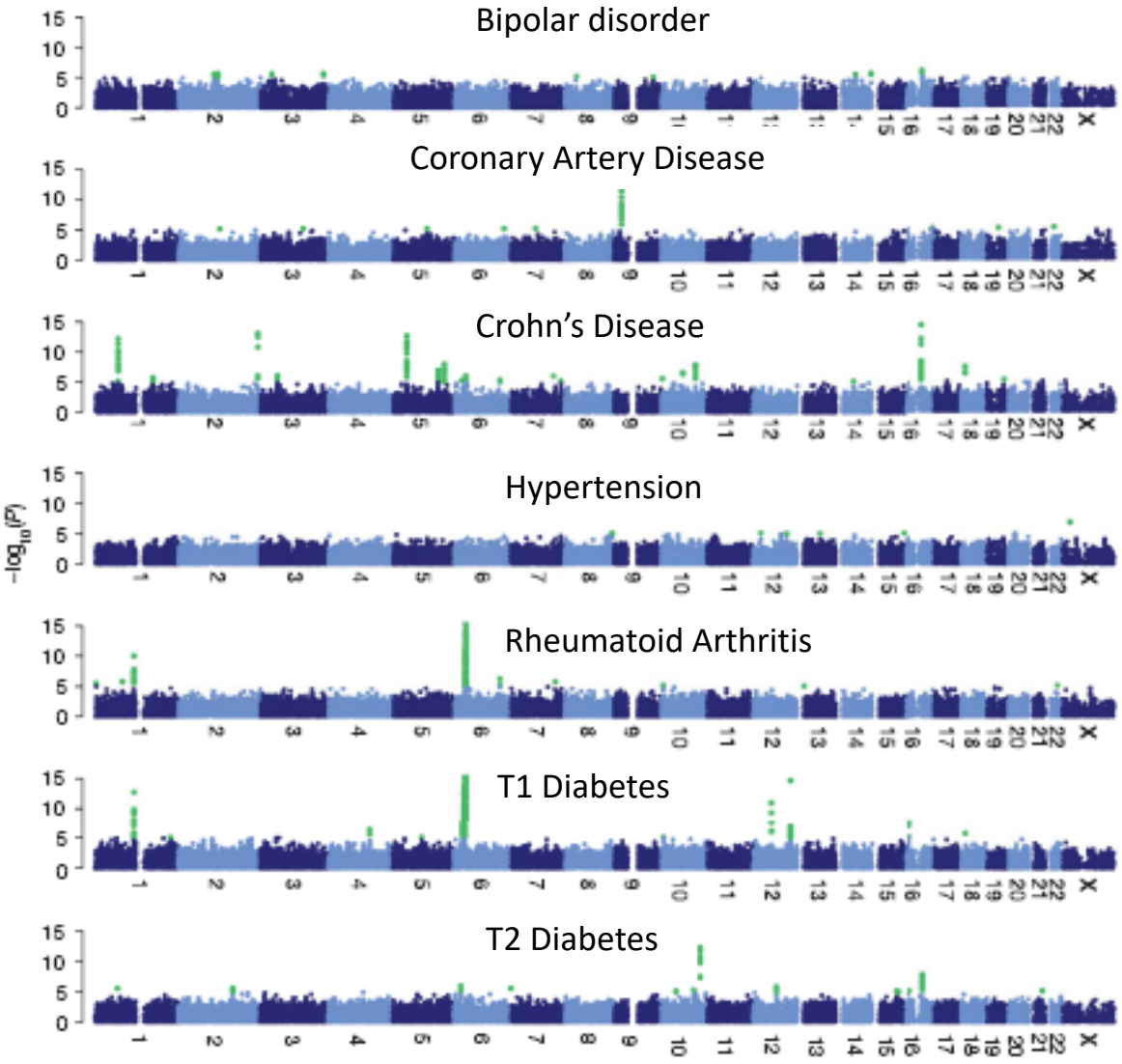
6 (MHC)
1p13
7q32 (females only!)

6 (MHC) 12q13
1p13 12p13
4q27 12q24 16p13

6p22 10q25
16q12

Wellcome Trust Case-Control Consortium

Replicated!



Significant hits

16p12

9p21

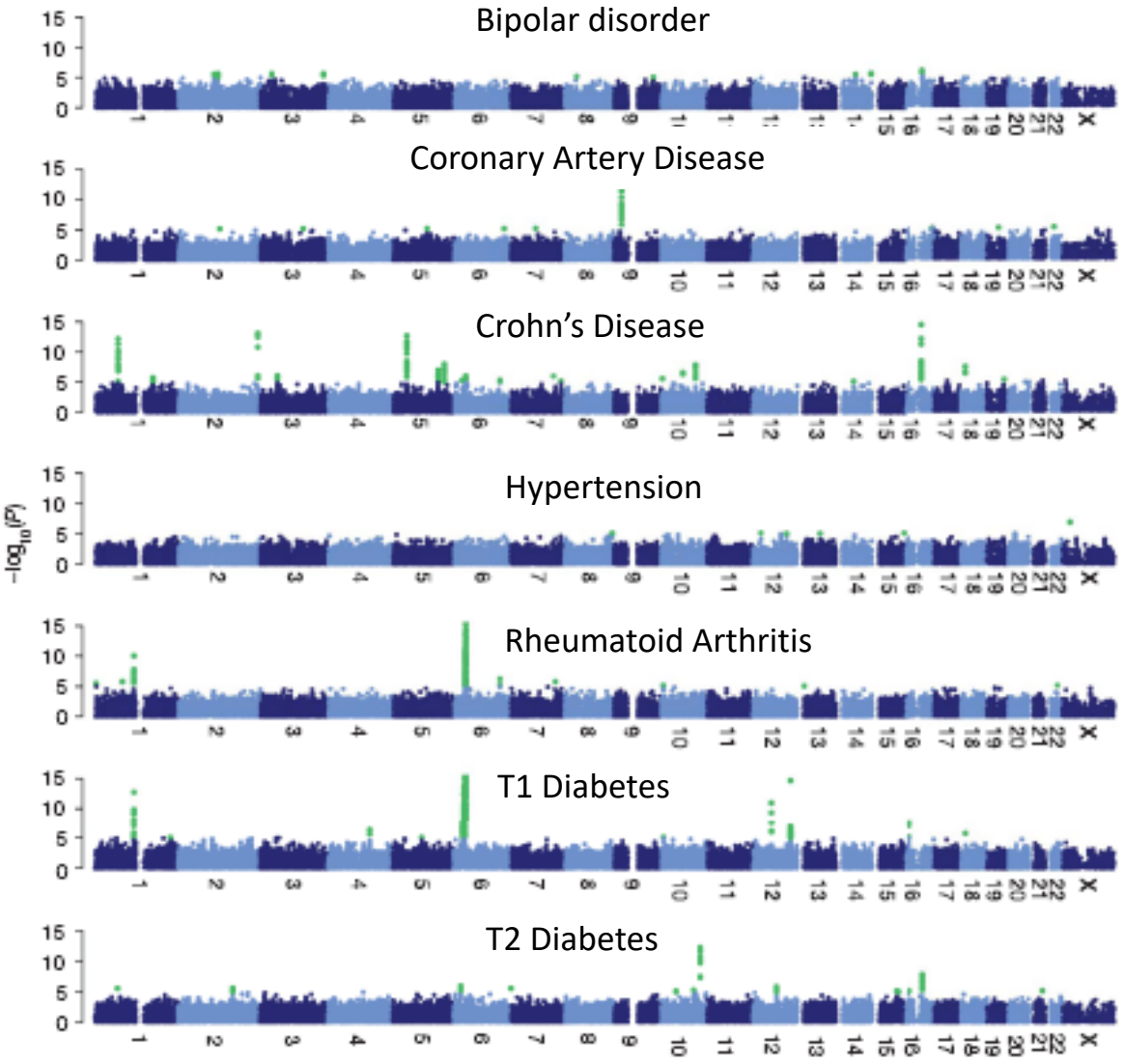
1p31 5p13 10q24
2q37 5q33 16q12
3p21 10q21 18p11

6 (MHC)
1p13
7q32 (females only)

6 (MHC) 12q13
1p13 12p13
4q27 12q24 16p13

6p22 10q25
16q12

Wellcome Trust Case-Control Consortium



	Significant hits	Heritability
Bipolar disorder	16p12	0.6-0.8
Coronary Artery Disease	9p21	0.5
Crohn's Disease	1p31 5p13 10q24 2q37 5q33 16q12 3p21 10q21 18p11	0.8
Hypertension	-	0.3-0.6
Rheumatoid Arthritis	6 (MHC) 1p13 7q32 (females only!)	0.5-0.7
T1 Diabetes	6 (MHC) 12q13 1p13 12p13 4q27 12q24 16p13	>0.85
T2 Diabetes	6p22 10q25 16q12	0.3-0.6

NHGRI GWA study Catalog: published Genome-Wide Associations at $P \leq 5 \times 10^{-8}$

>7,796 publications and >159,202 associations

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease
- Other trait

National Human Genome Research Institute

EMBL-EBI

Complex diseases: Rare variants?

- Low-frequency (MAF 0.5-5%) and rare (MAF<0.5%) variation may contribute to the heritability of complex traits:
    - IFIH1 and type 1 diabetes;
    - MYH6 and sick sinus syndrome.
- Assaying rare variants: Exome genotyping arrays or re-sequencing
- Rare variants are expected to have larger effects on complex traits that common variants.
- Single variant tests lack power to detect association.
- Statistical methods focus on the accumulation of minor alleles at rare variants (mutational load) within the same "genomic unit".
  - Burden tests: assumes same direction of effect of all rare variants
  - Dispersion tests: allow for different direction of effect

Summary

It is clinically useful and interesting to look for the genetic variants contributing to human traits

A combination of theory, understanding of population genetics, and technology has made it possible to carry out GWAS analysis.

GWAS are an appropriate tool to identify common variants of relatively small risk (<1.5) for complex traits/disease.

- Requires tens of thousands of samples
- Good coverage of the SNP markers