Biometrical Model and the Genome and its secrets

Benjamin Neale, PhD

Boulder Workshop

Content warning: eugenicists

Sick individuals and sick populations Rose 1985

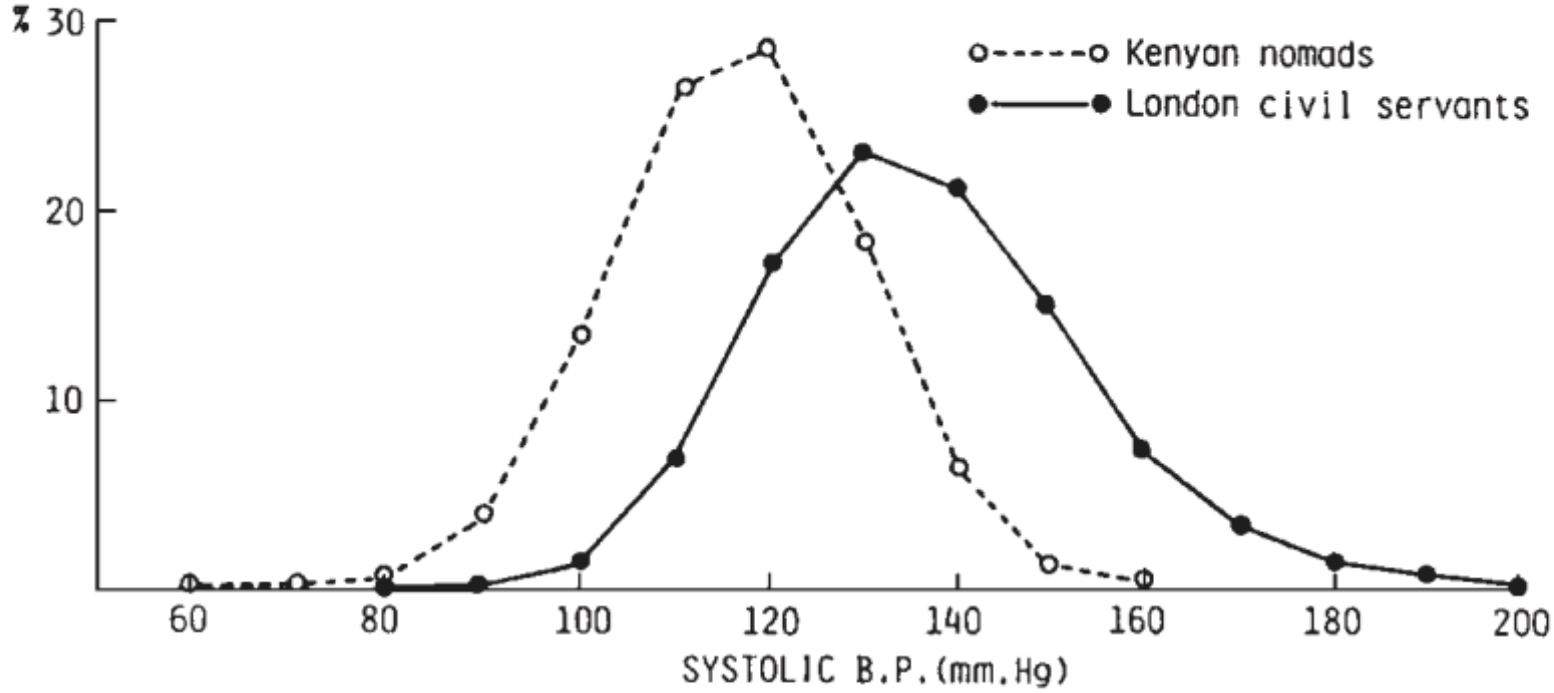
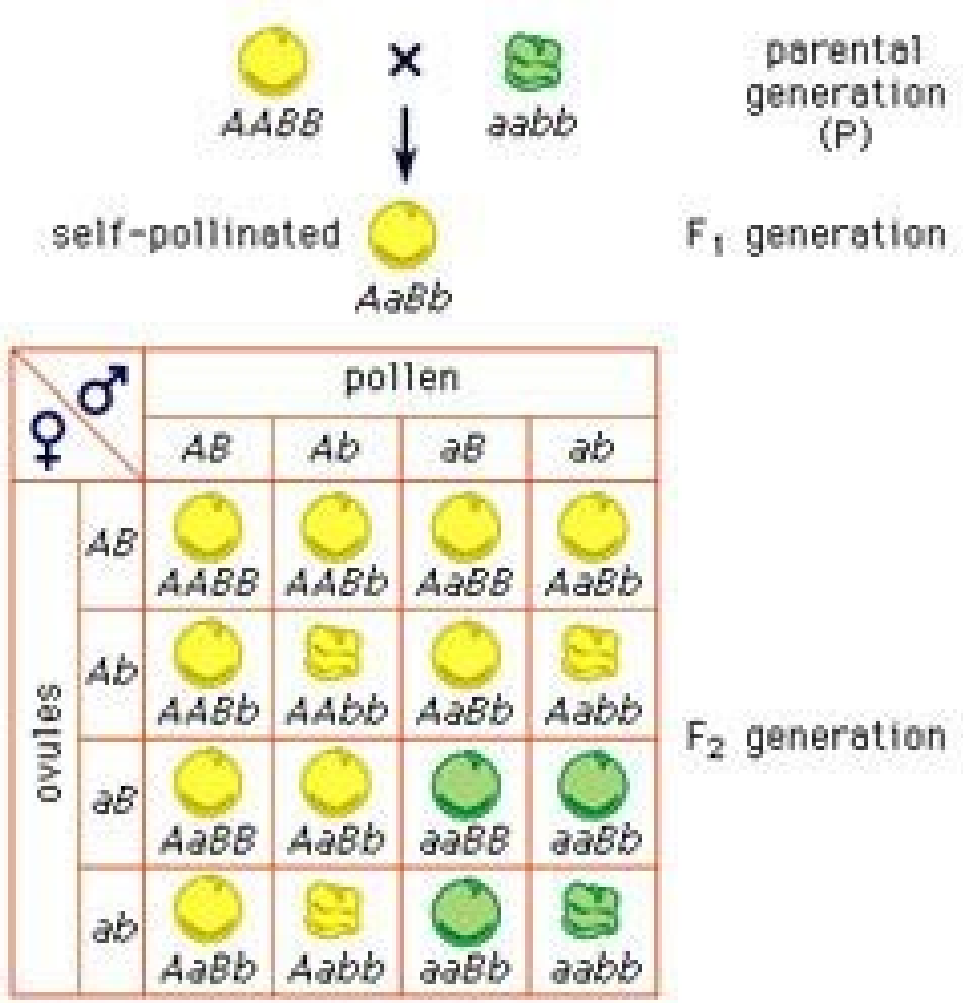
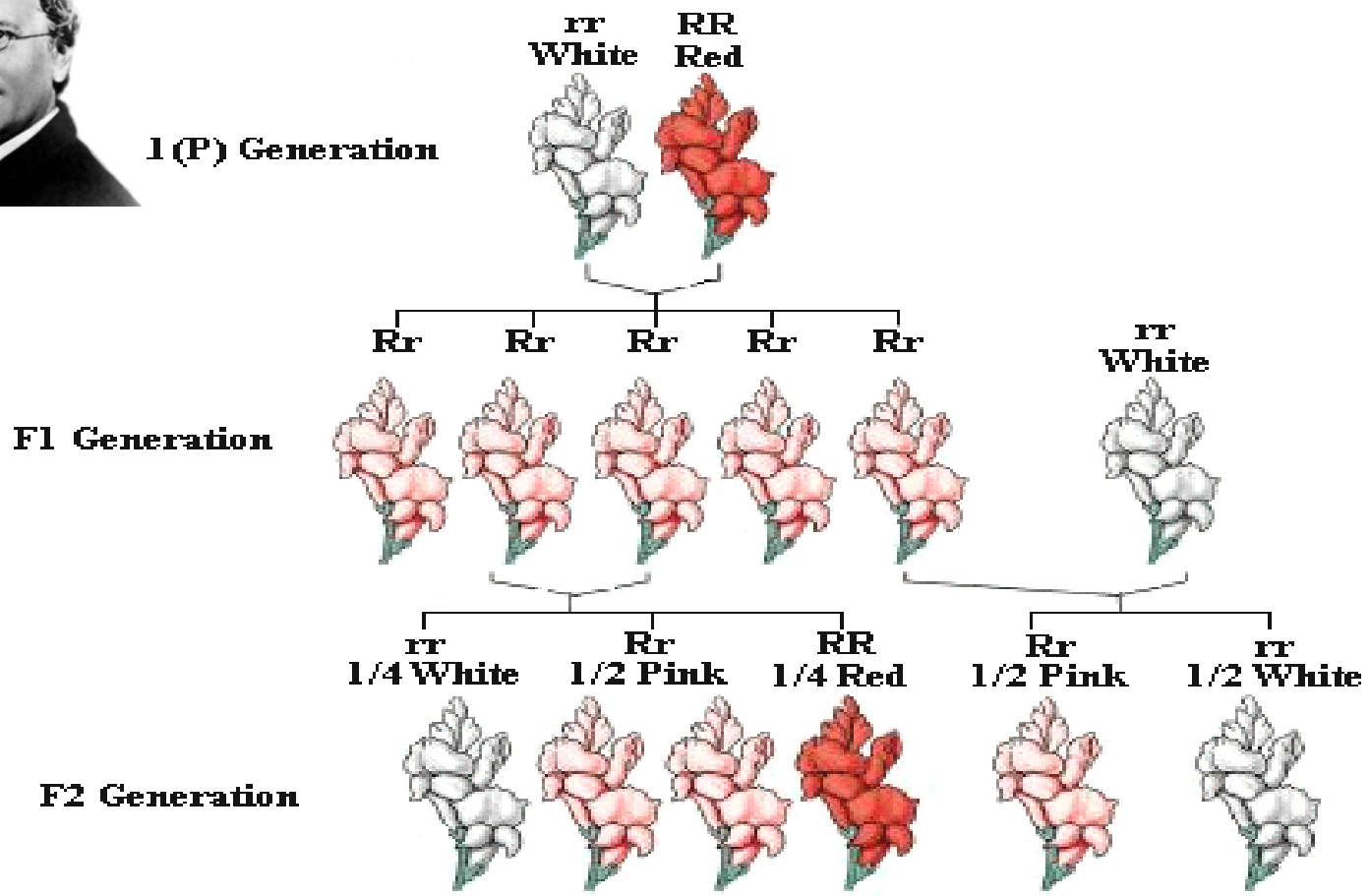


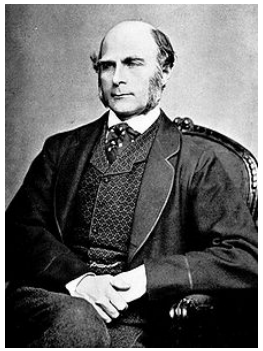
Figure 2 Distributions of systolic blood pressure in middle-aged men in two populations[2,3]

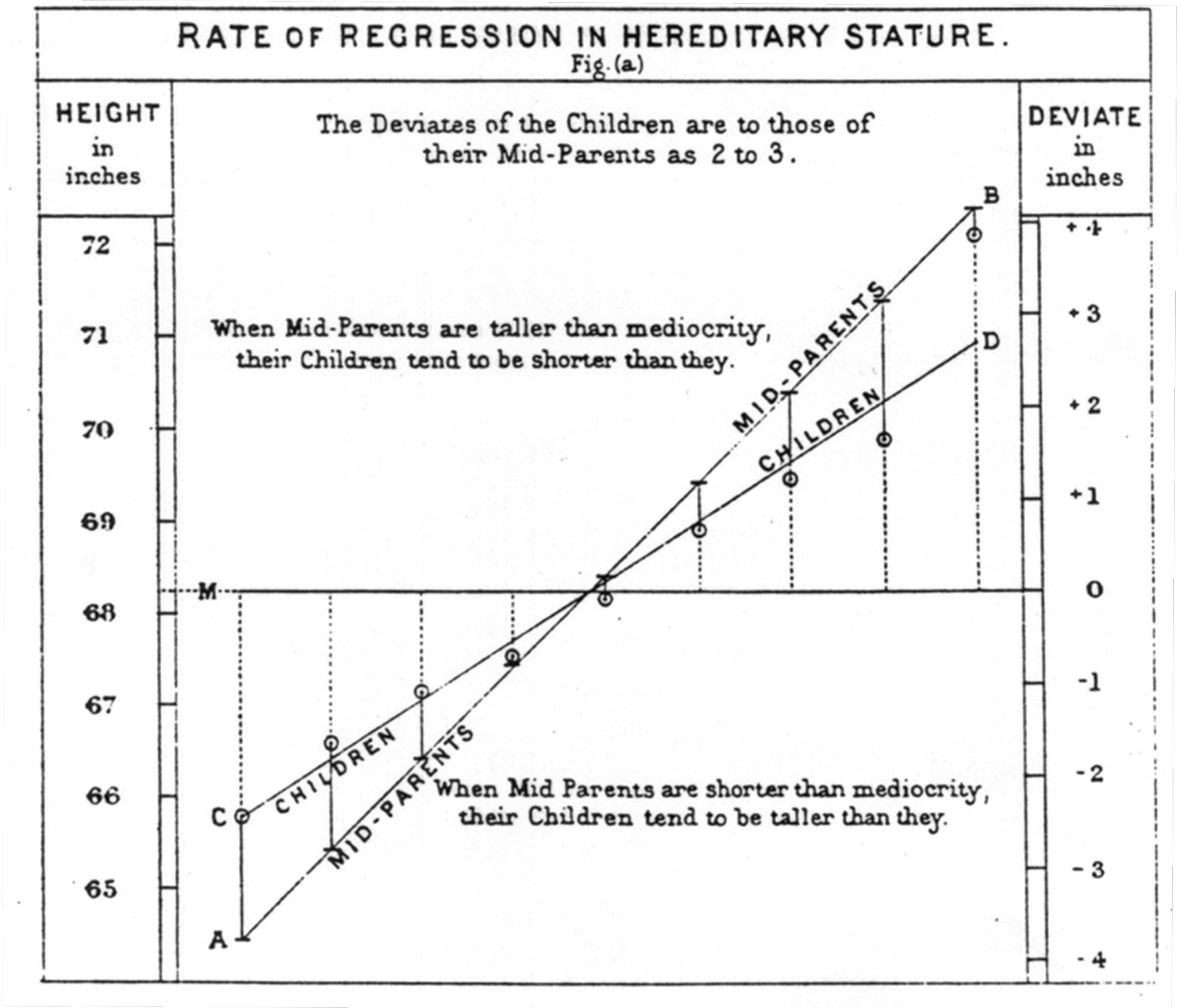
Mendelian Genetics



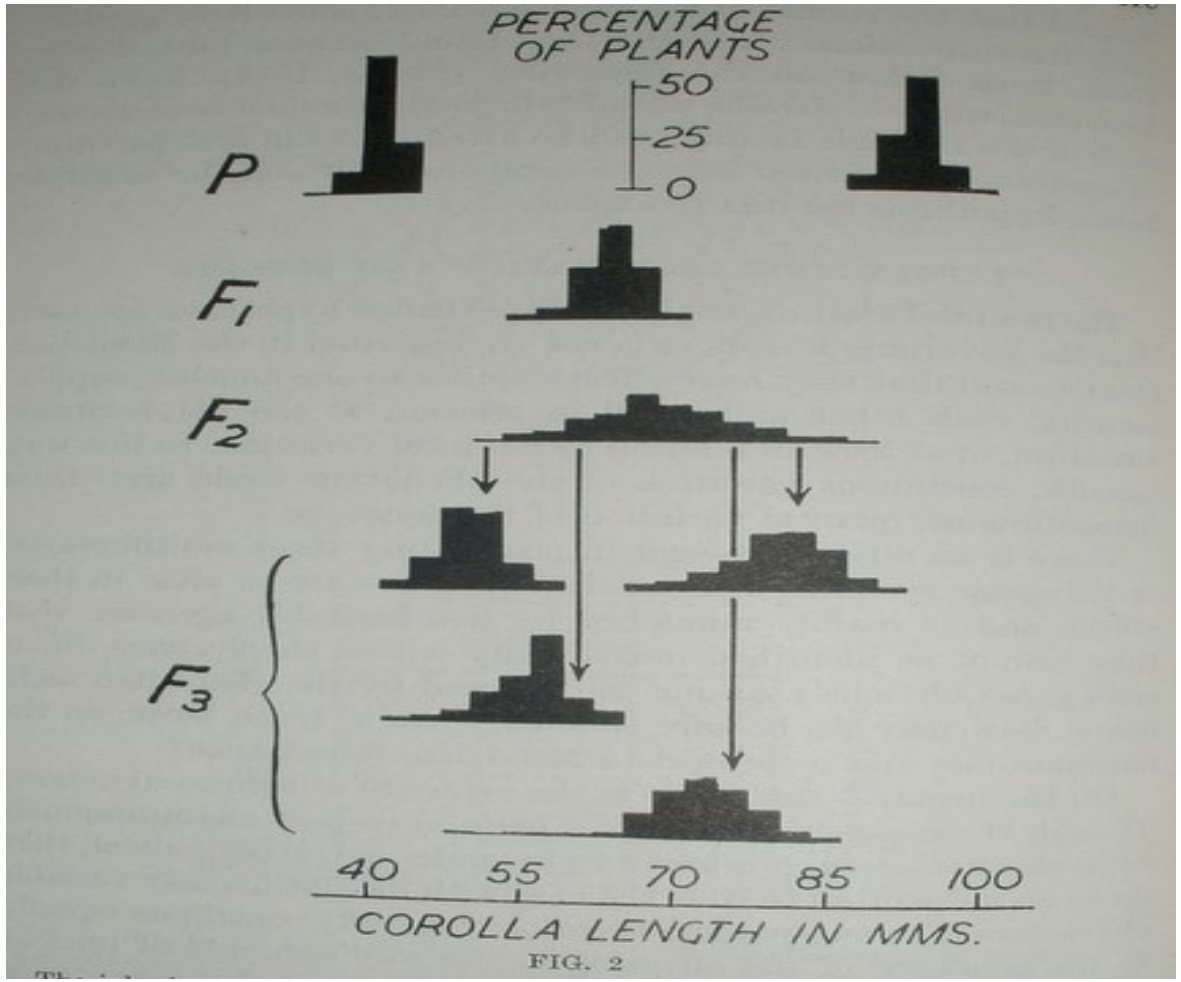
Co-dominance






RATE OF REGRESSION IN HEREDITARY STATURE.
Fig. (a)
HEIGHT in inches
DEVIATE in inches
The Deviates of the Children are to those of their Mid-Parents as 2 to 3.
When Mid-Parents are taller than mediocrity, their Children tend to be shorter than they.
When Mid Parents are shorter than mediocrity, their Children tend to be taller than they.
MID-PARENTS
CHILDREN
CHILDREN
MID-PARENTS
A
B
C
D
M
72
71
70
69
68
67
66
65
+4
+3
+2
+1
0
-1
-2
-3
-4

East 1915: Inheritance of Corolla Length in *Nicotiana longiflora*



Neo-Darwinist Reconciliation

XV.—**The Correlation between Relatives on the Supposition** of Mendelian Inheritance. By **R. A. Fisher**, B.A. *Communicated by* Professor J. Arthur Thomson. (With Four Figures in Text.)

(MS. received June 15, 1918. Read July 8, 1918. Issued separately October 1, 1918.)

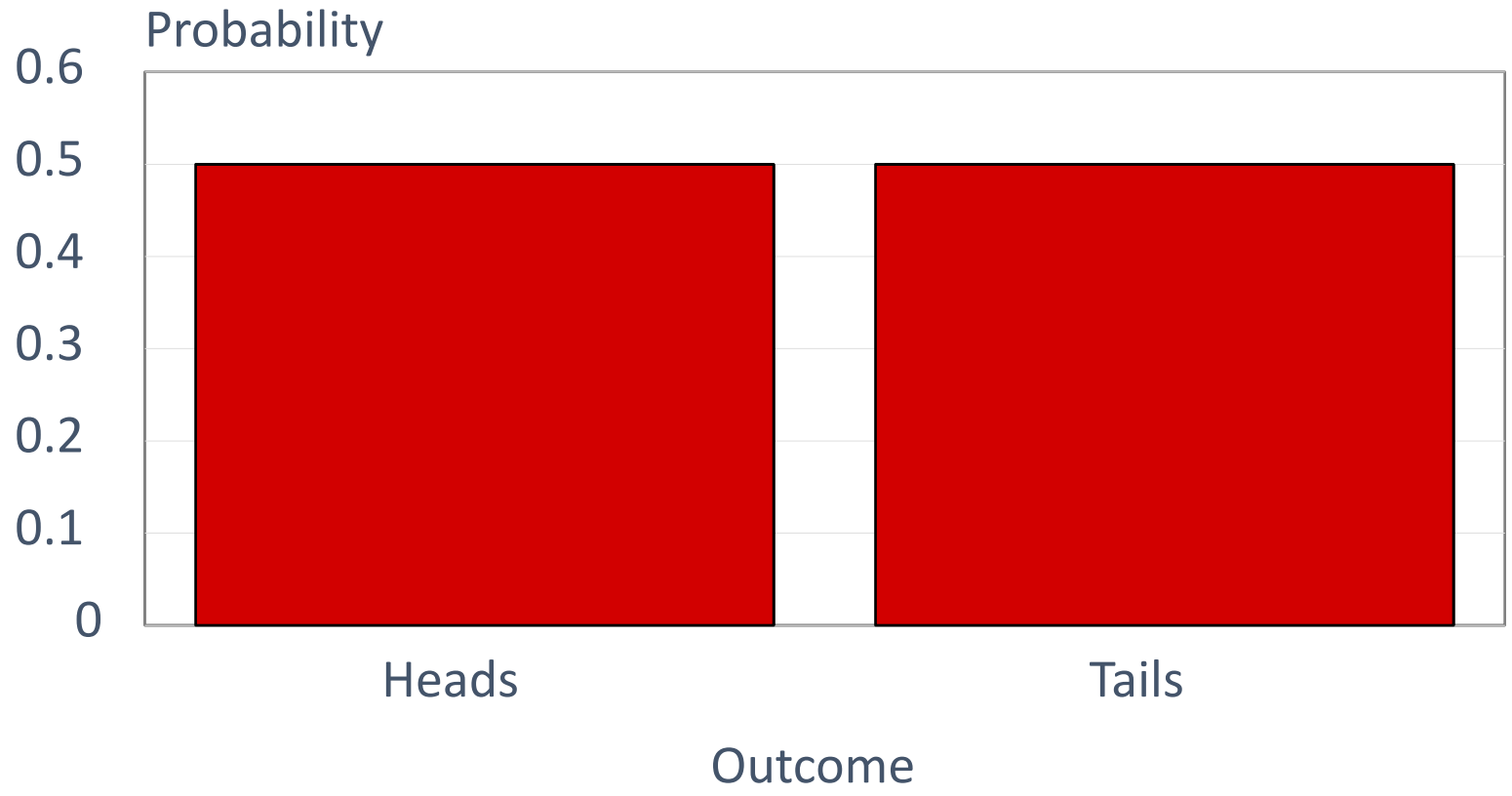
CONTENTS.

Several attempts have already been made to interpret the well-established results of biometry in accordance with the Mendelian scheme of inheritance. It is here attempted to ascertain the biometrical properties of a population of a more general type than has hitherto been examined, inheritance in which follows this

Ronald Fisher (1918)

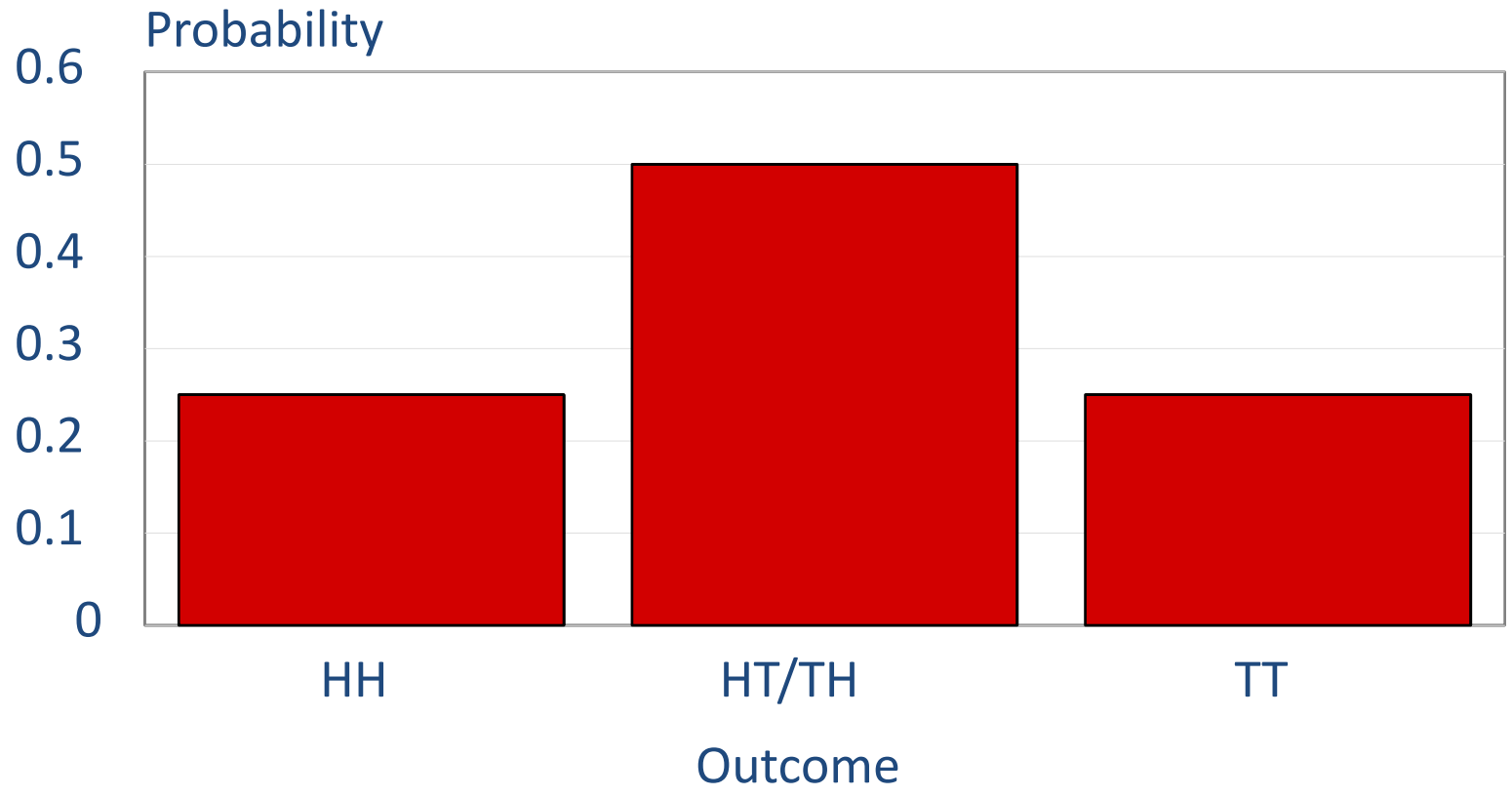
One Coin toss

2 outcomes



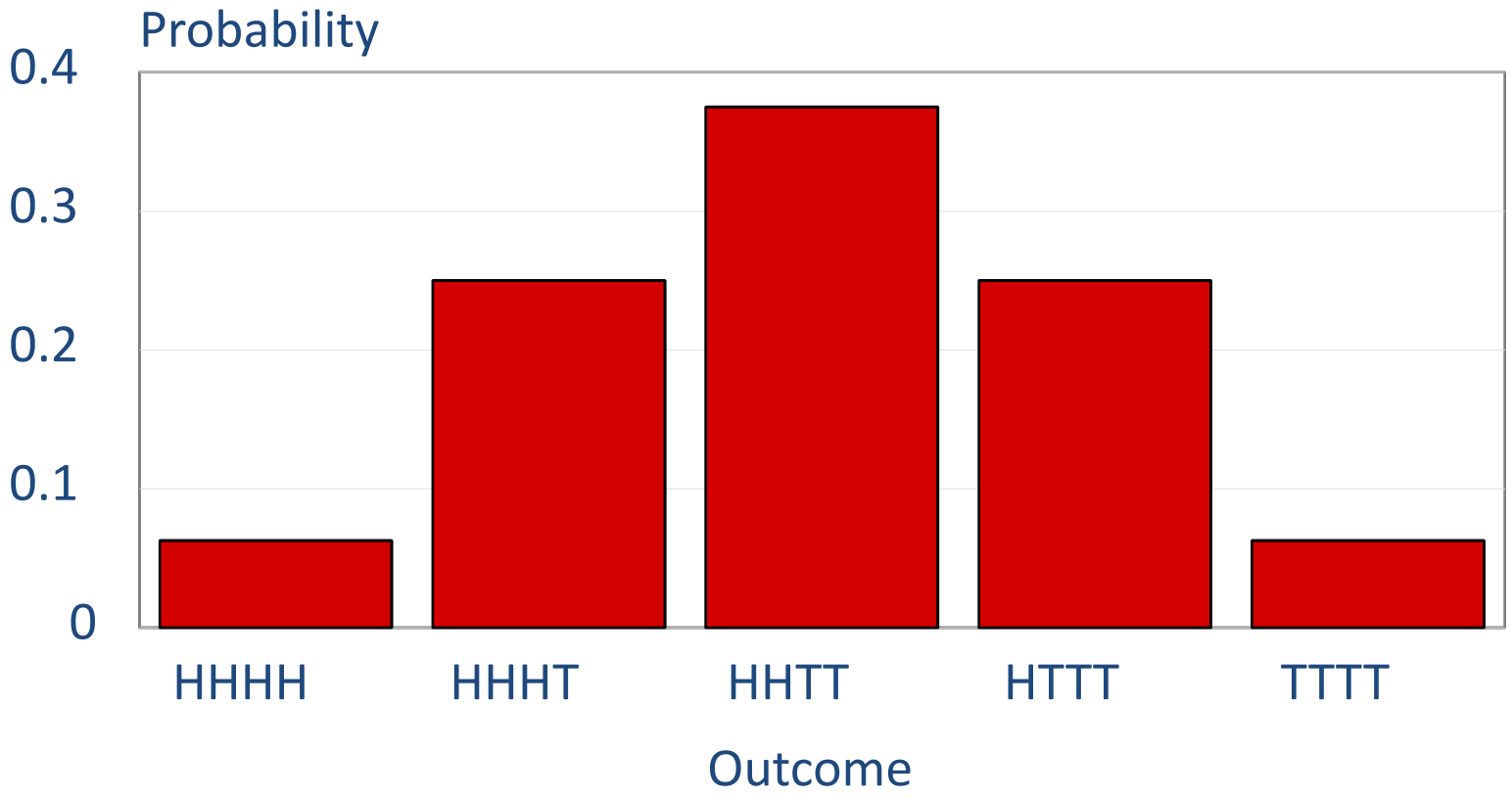
Two Coin toss

3 outcomes



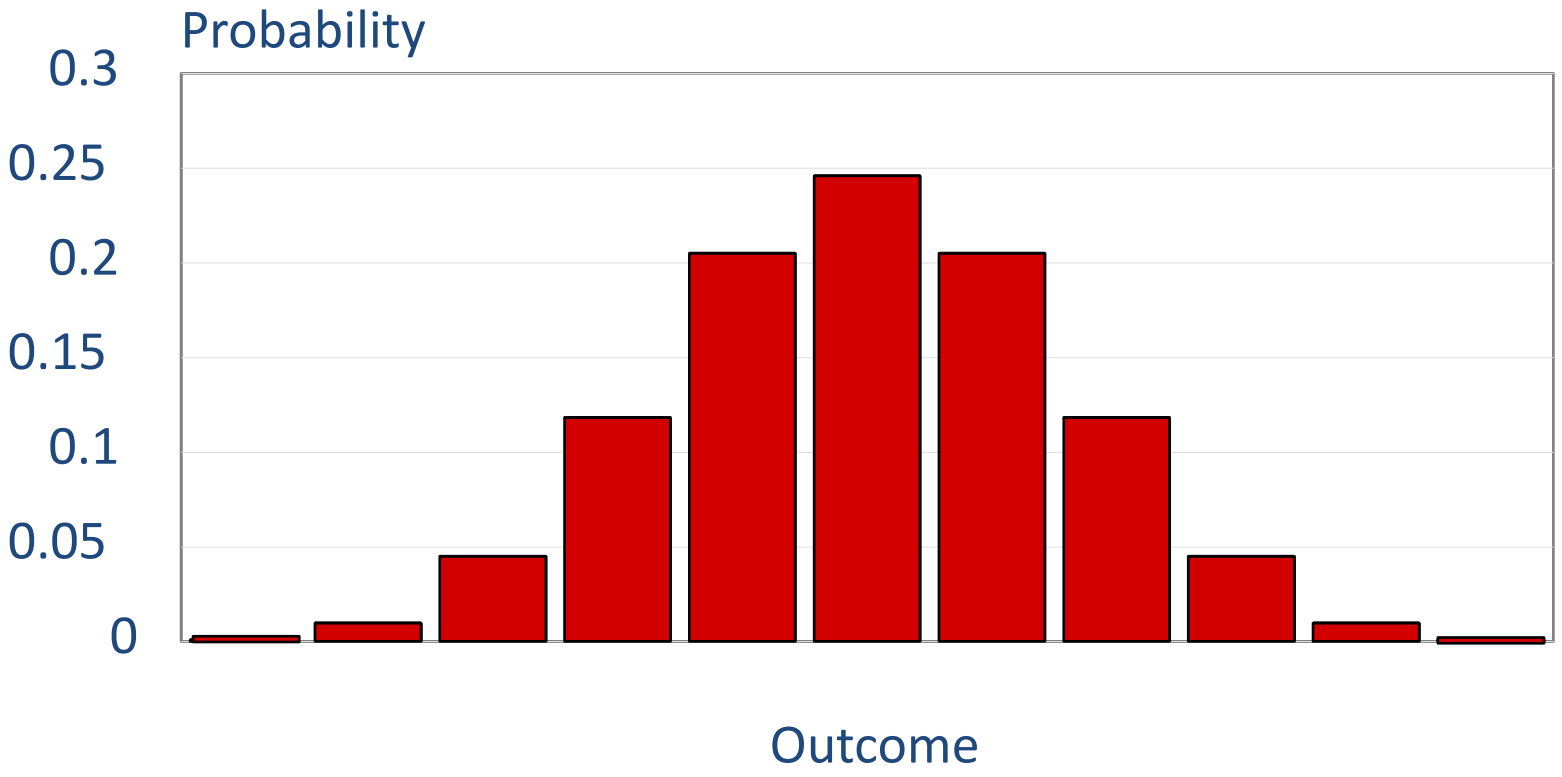
Four Coin toss

5 outcomes

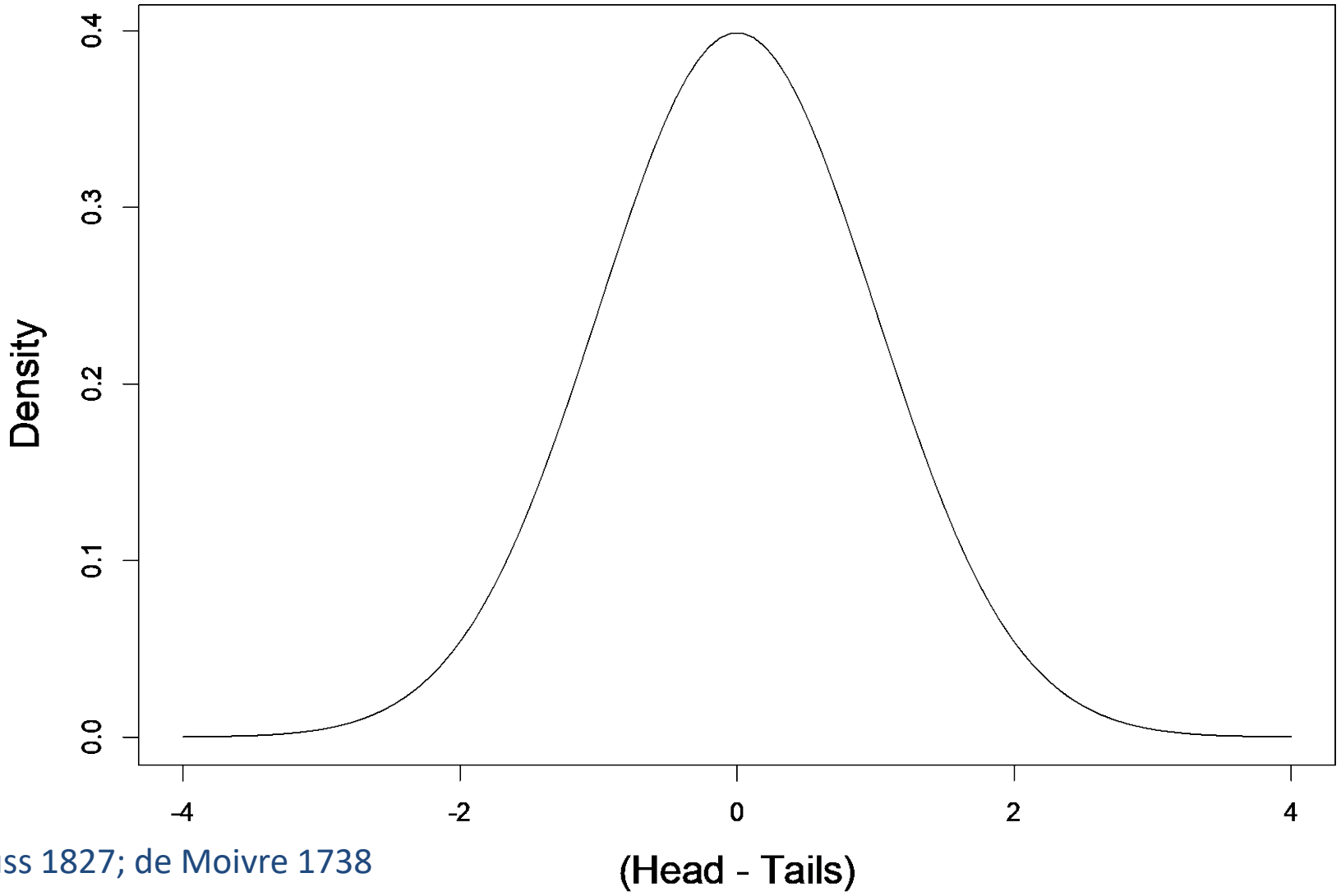


Ten Coin toss

11 outcomes

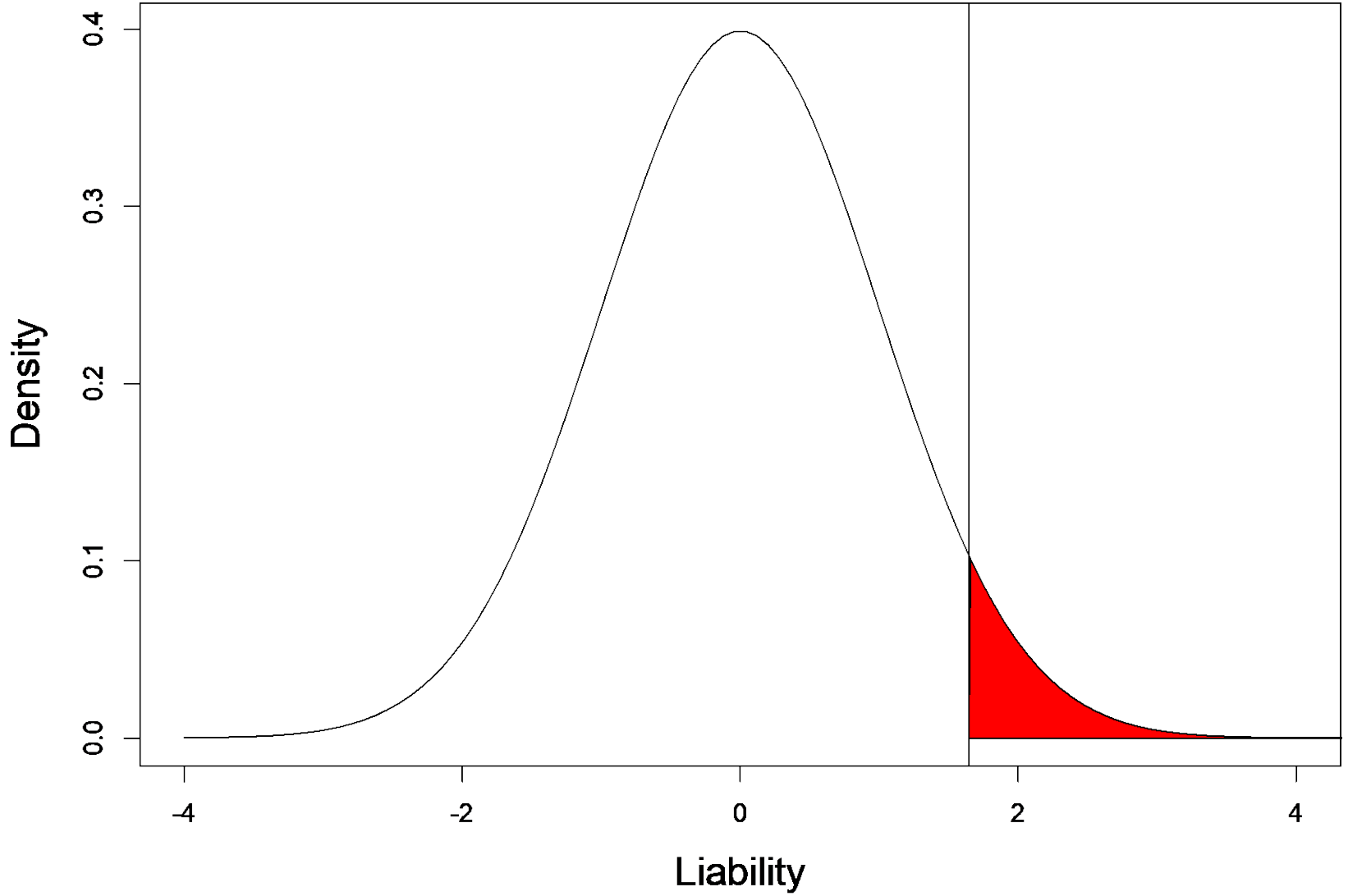


Infinite Outcomes



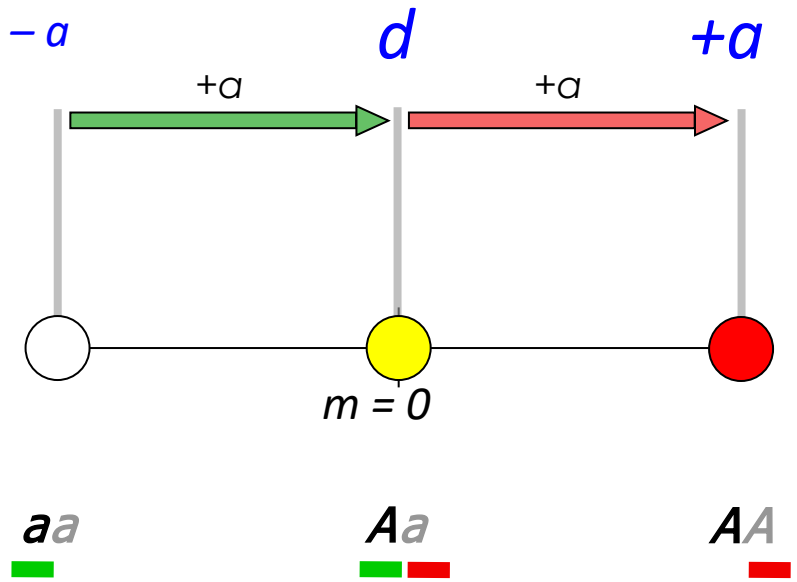
Gauss 1827; de Moivre 1738

Liability threshold model
Pearson and Lee (1901)



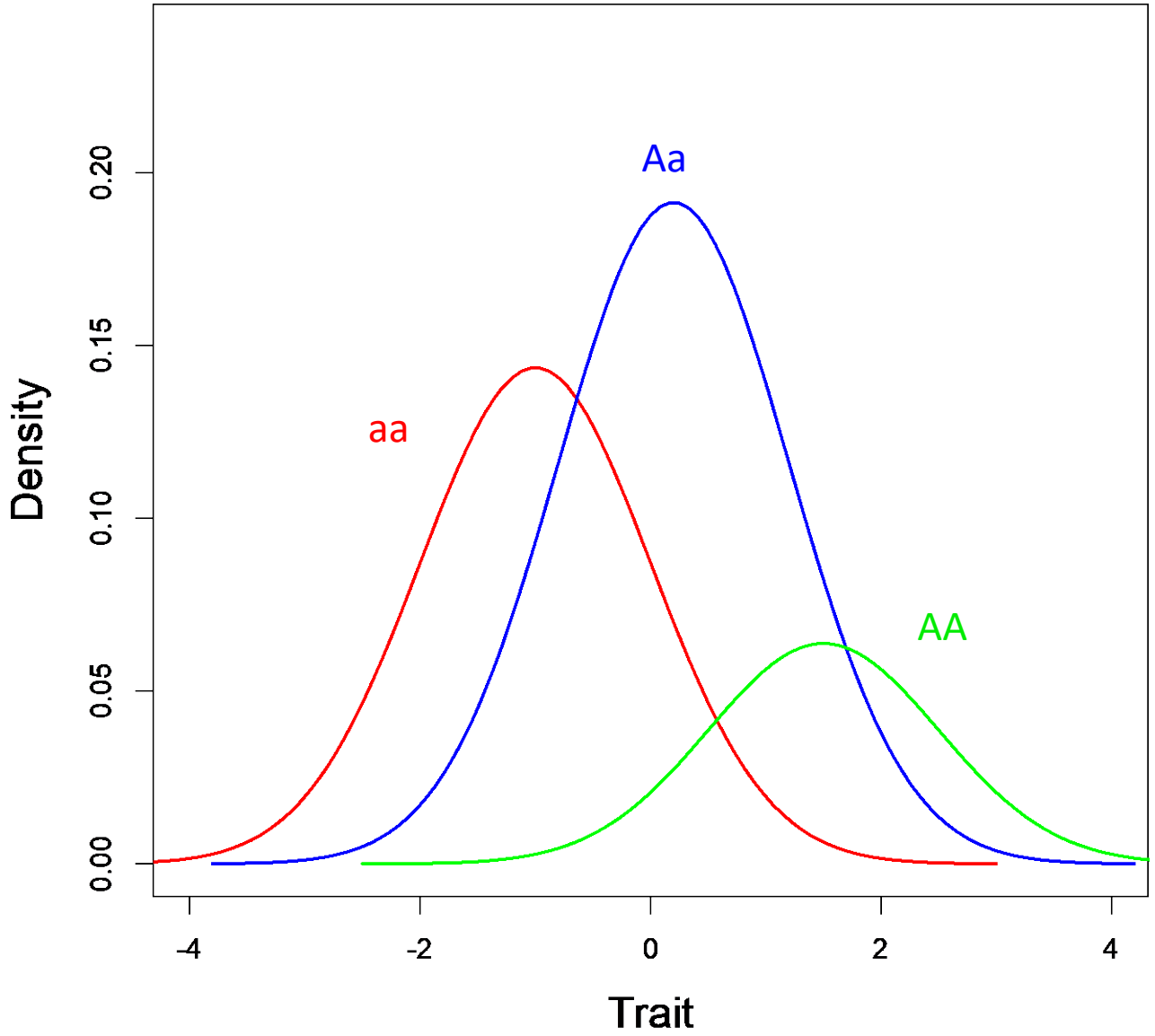
Genotype with an additive effect

d = 0 *(no dominance)*

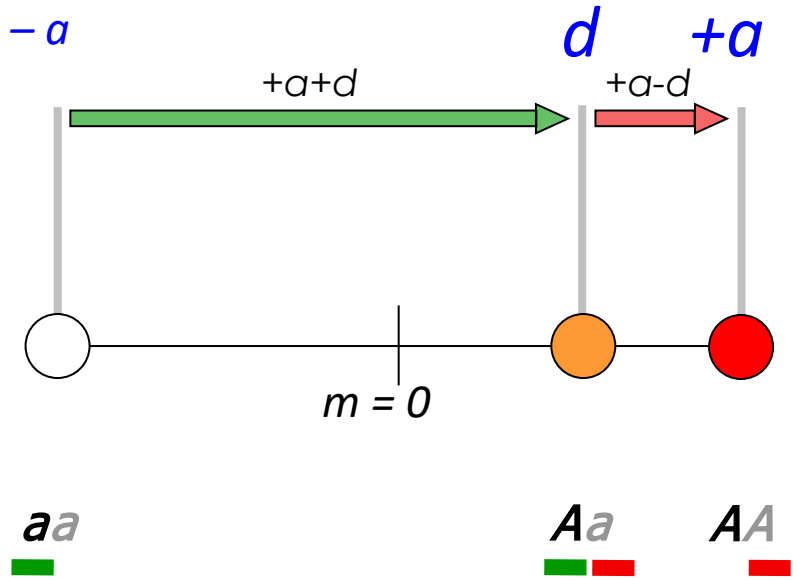


Additive model

Source of variation



Genotype with an additive and dominance effects $d > 0$ *(dominance)*



Dominant model

How much mean and variance?

1. Defining the Mean (*X*)

e.g. cholesterol levels in the population

$$\mu = \sum_i x_i f(x_i)$$

Genotypes	***AA***	***Aa***	***aa***
Effect, x	***a***	***d***	***-a***
Frequencies, *f*(x)	p^2	$2pq$	q^2

Mean (X) $= a(p^2) + d(2pq) - a(q^2) = a(p-q) + 2pqd$

How much mean and variance?

2. Contribution of the QTL to the Variance (*X*)

$$Var = \sum_i (x_i - \mu)^2 f(x_i)$$

Genotypes	***AA***	***Aa***	***aa***
Effect, x	***a***	***d***	***-a***
Frequencies, *f*(x)	p^2	$2pq$	q^2

$$Var\ (X) = (a-m)^2p^2 + (d-m)^2 2pq + (-a-m)^2q^2$$

$$= V_{QTL}$$

Heritability of *X* at this locus = V_{QTL} / V_{Total}

How much mean and variance?

$$Var\,(X) = (a-m)^2p^2 + (d-m)^2 2pq + (-a-m)^2q^2$$

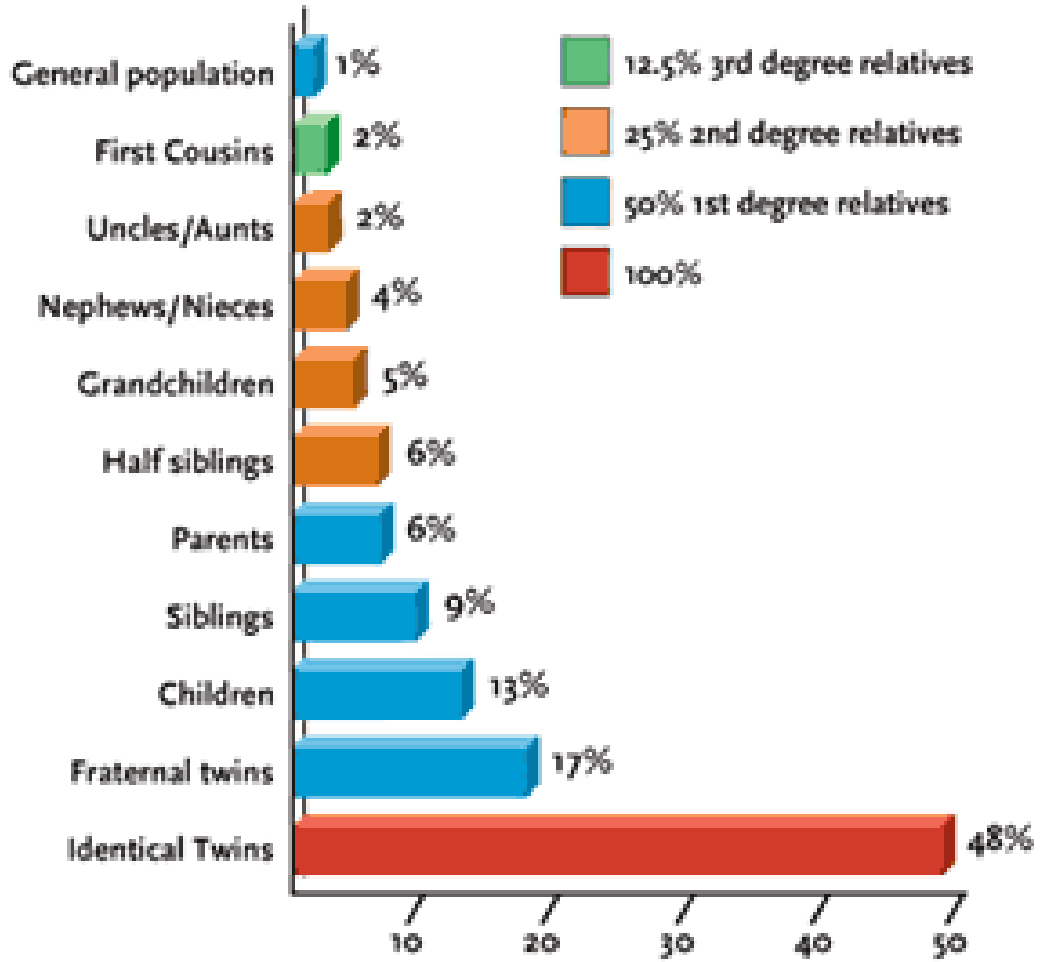
$$m = a(p-q) + 2pqd$$

$$= \underbrace{2pq[a+(q-p)d]^2}_{V_{A_{QTL}}} + \underbrace{(2pqd)^2}_{V_{D_{QTL}}}$$

Additive effects: the main effects of individual alleles

Dominance effects: represent the deviation from additive effects

Twin and family studies – probability of having schizophrenia conditional on relative



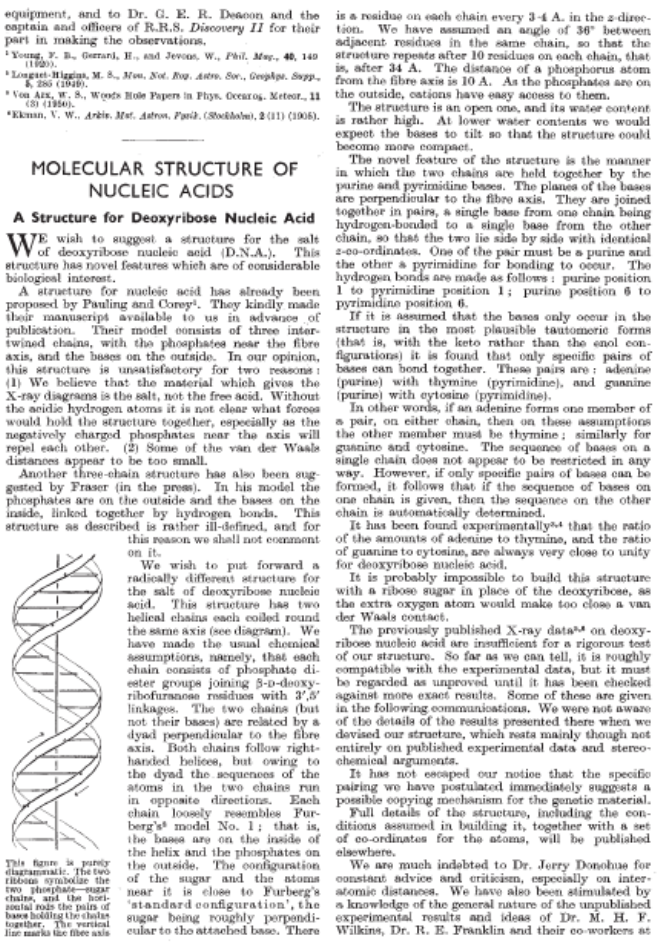
Near continuous fall off of risk – proportional to amount of shared genome

Gottesman 2001 Genes and More

What about the genome?

DNA

- **Structure**



No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerard, R., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., **11** (3) (1950).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

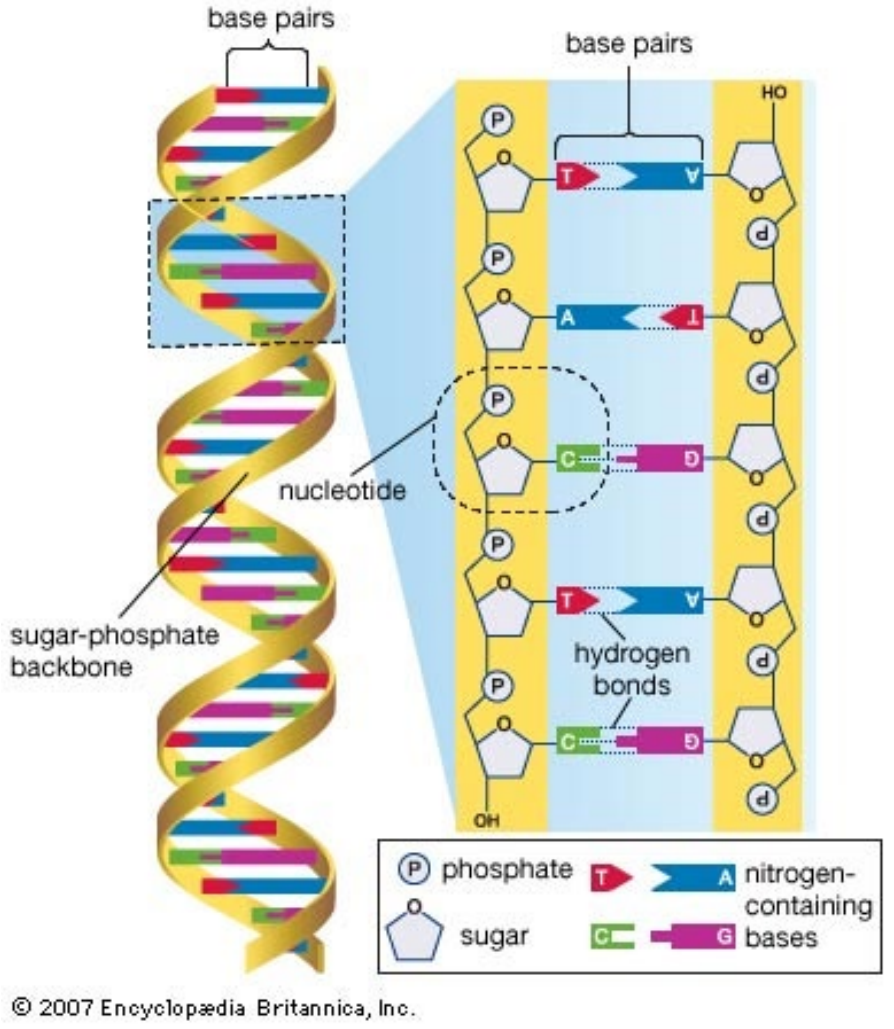
The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

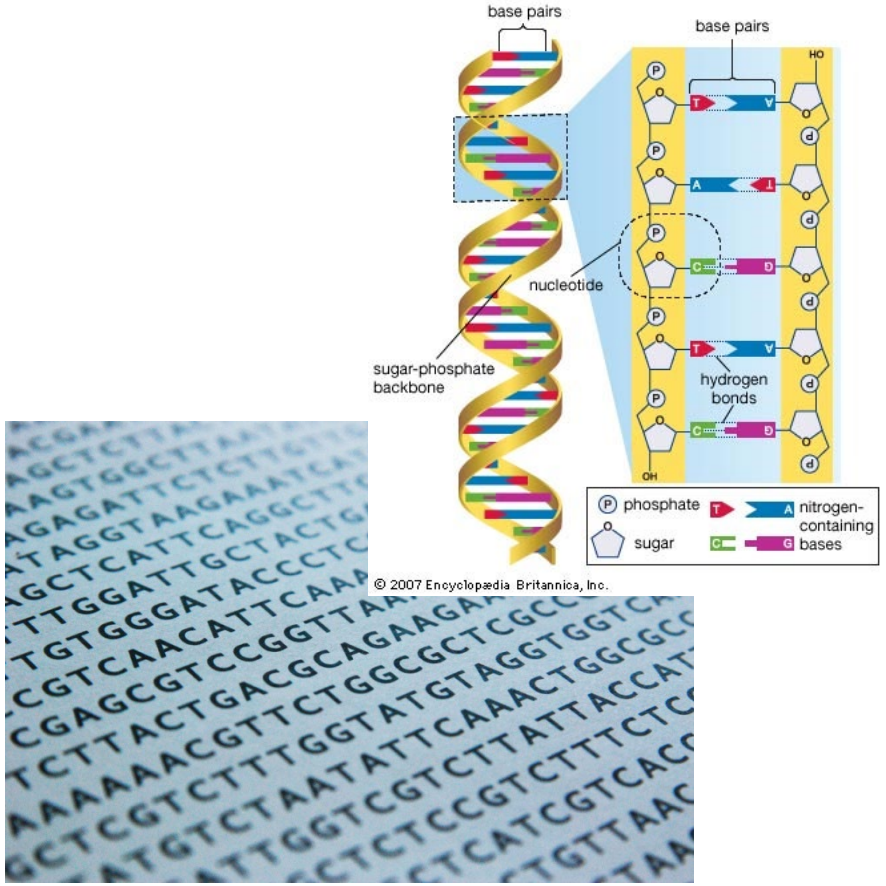
This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis



1953 Watson & Crick

Bases: **A**, **C**, **G**, **T**

- **Sequence**

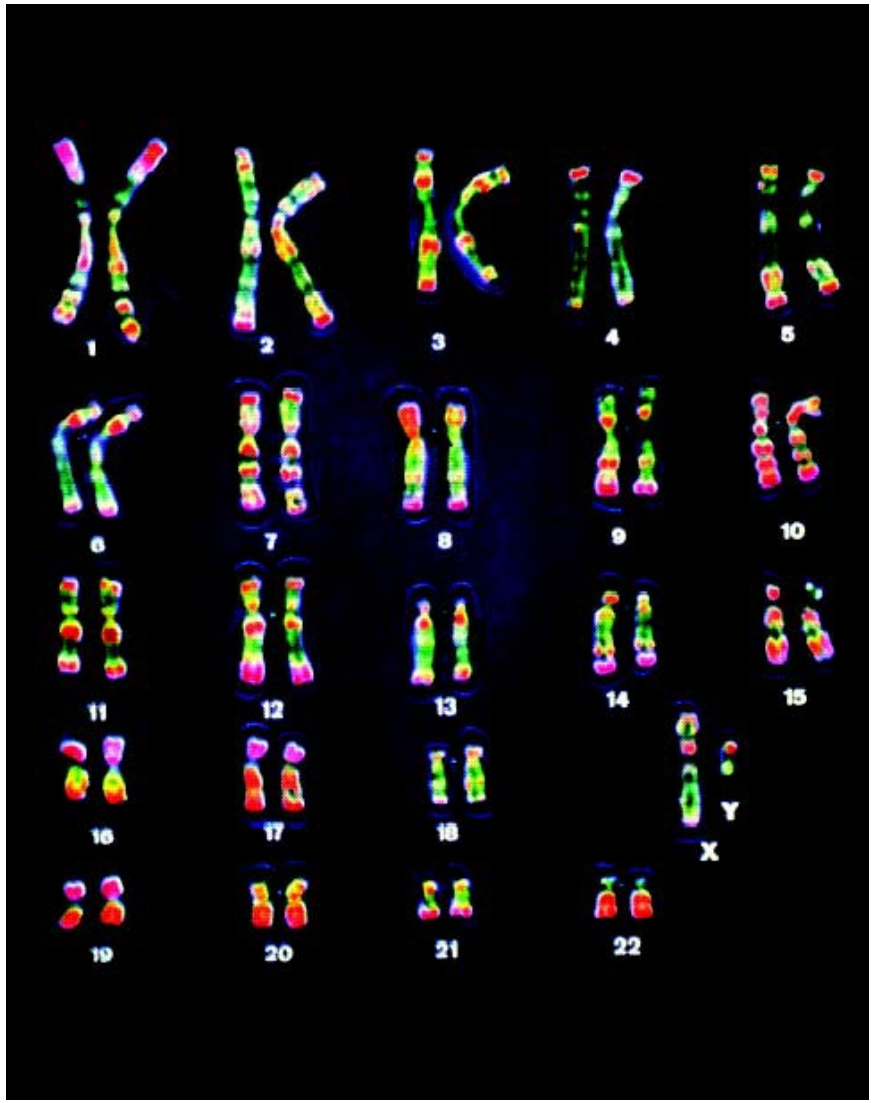


- *10 years*
- *USD $3 billion*
- *Sequence DNA of one single person*

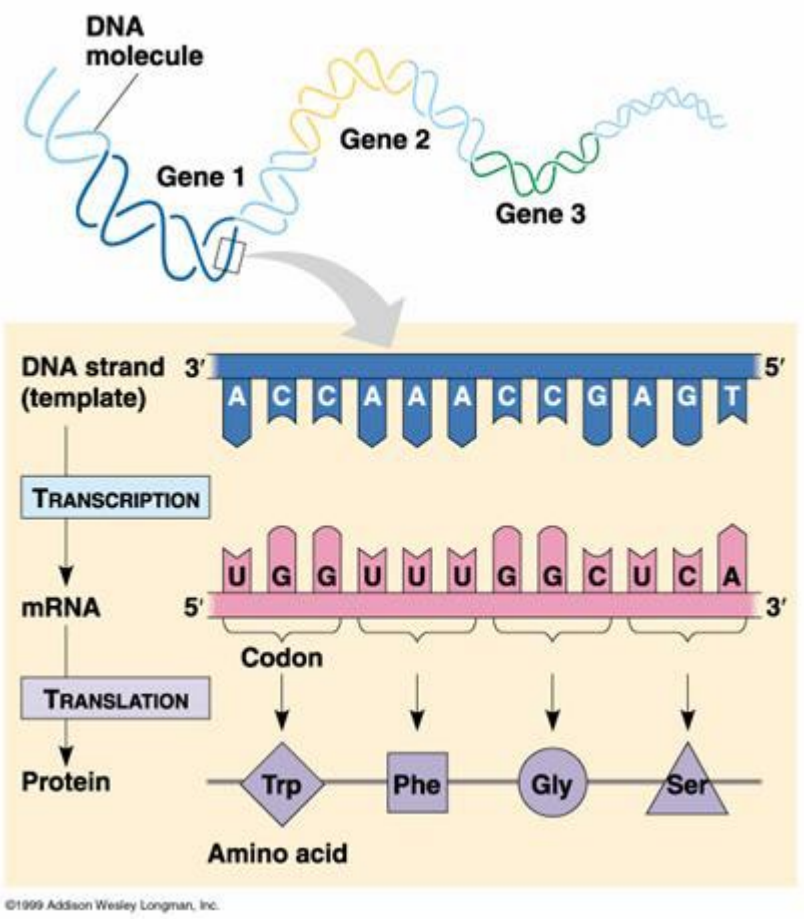
~ 3,000,000,000 nucleotides

2001

Organization



- 23 pairs of chromosomes
- "pairs": one copy from father, one mother
- *~20,000 genes*



Variation

2010

- *Sequenced DNA of >1,000 individuals*
- *5 years*
- *Less than $5,000 per individual*

~3,000,000,000 bases
~30,000,000 different

↓

Contribute to making us different:

how we look
behave,
diseases,
etc

DNA

30,000,000 nucleotides where the base can differ between people

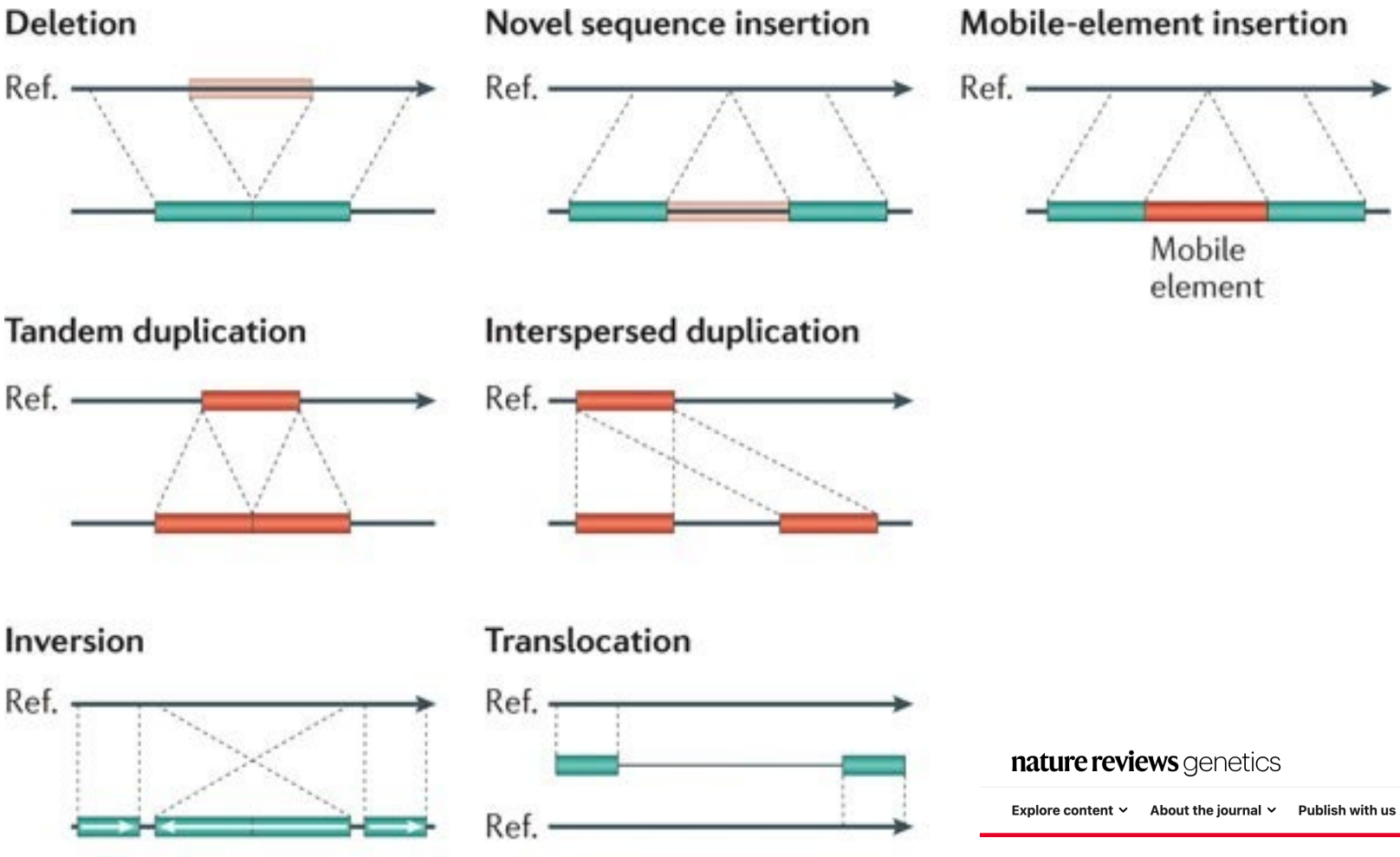
- *Single Nucleotide Polymorphism, SNP*

	Chrom.	DNA sequence	Genotype SNP 1	Genotype SNP 2
Person 1	Mat	GTAACTTGGGATCT**A**GACCA**G**ATAGAT	**A A**	**G G**
	Pat	GTAACTTGGGATCT**A**GACCA**G**ATAGAT		
Person 2	Mat	GTAACTTGGGATCT**A**GACCA**G**ATAGAT	**A C**	**G G**
	Pat	GTAACTTGGGATCT**C**GACCA**G**ATAGAT		
Person 3	Mat	GTAACTTGGGATCT**C**GACCA**G**ATAGAT	**C C**	**G T**
	Pat	GTAACTTGGGATCT**C**GACCA**T**ATAGAT		

SNP 1 SNP 2

- Mutation that arose at some point in evolution
- Typically, each SNP has two alleles (bases)
- Each SNP is eventually given an "rs" number rs214621

Other kinds of variation



nature reviews genetics

Explore content | About the journal | Publish with us | Subscribe

nature › nature reviews genetics › review articles › article

Published: 01 March 2011

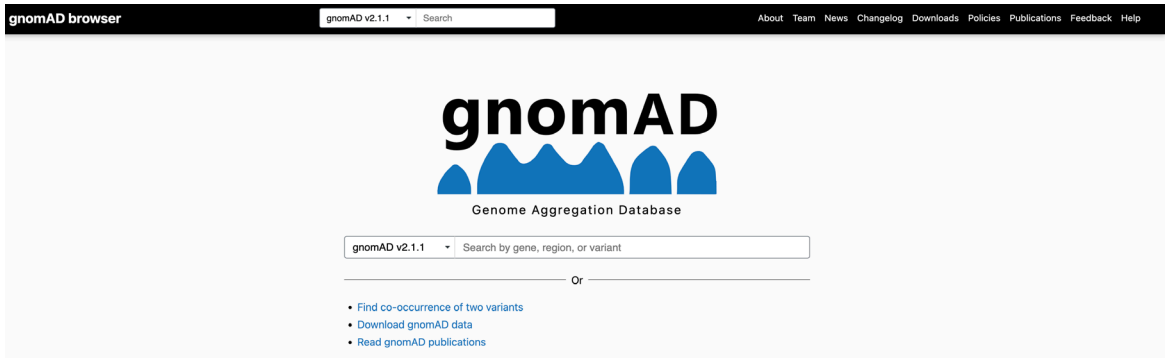
Genome structural variation discovery and genotyping

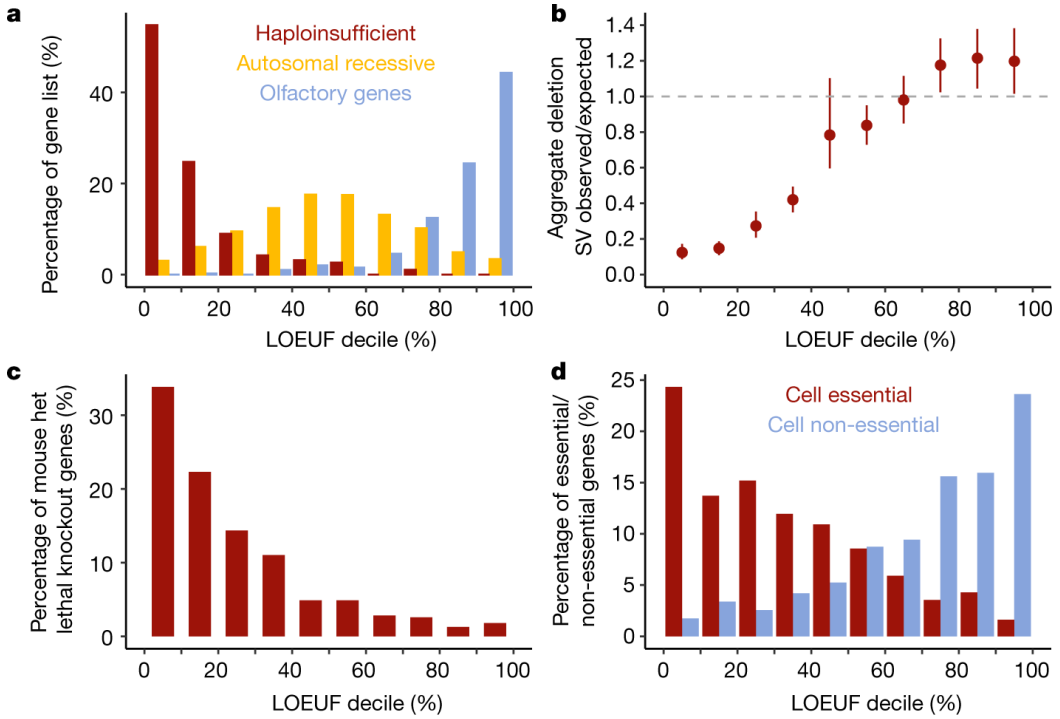
Can Alkan, Bradley P. Coe & Evan E. Eichler

Nature Reviews Genetics **12**, 363–376 (2011) | Cite this article

35k Accesses | **929** Citations | **45** Altmetric | Metrics

Surveys of variation





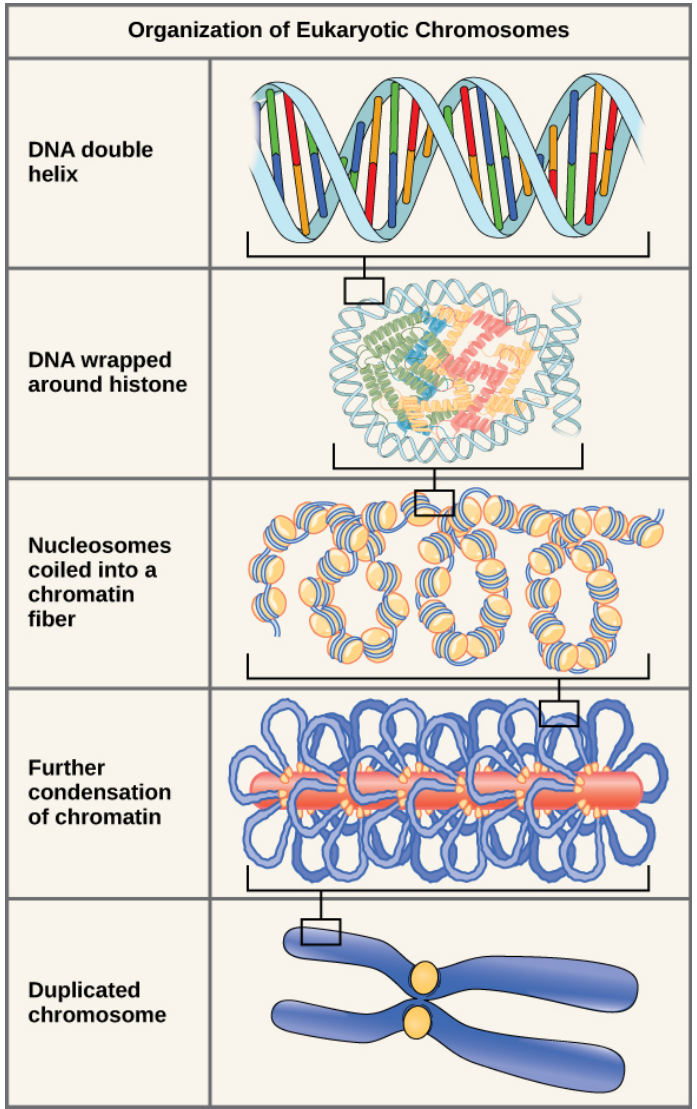
2020

The Genome Aggregation Database (gnomAD) has aggregated 15,708 whole genomes and 125,748 exomes

Ways to assay genetic variation

	Arrays	Exomes	Genomes
			
Upside	Hits + epi	Gene identification	Comprehensive capture
Downside	Hit interpretation	Limited Scope	Cost (small N)

Other interesting things about our genome



The genome is dynamic

DNA is chemically modified

-e.g. methylation

It acquires somatic mutations

-particularly in response to mutagens

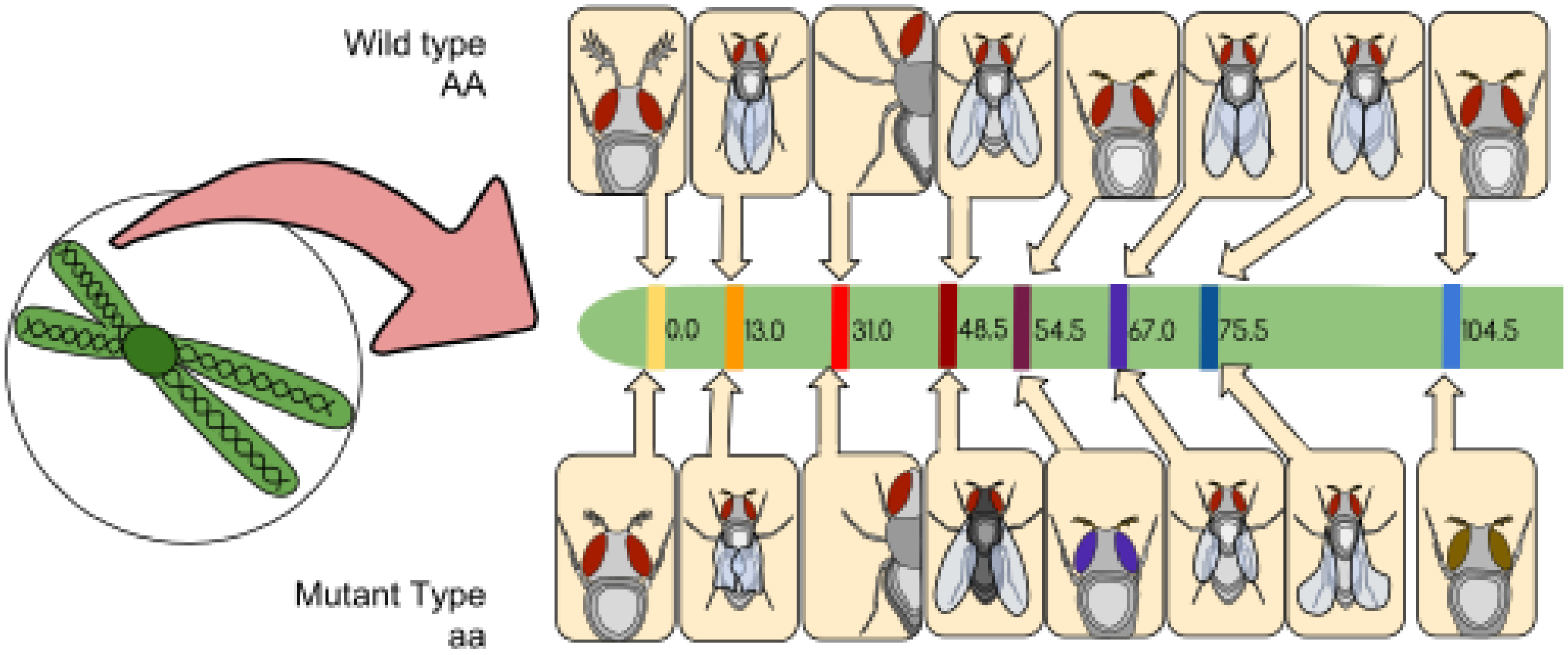
Example of epigenetic assay – ATAC-Seq

Open Chromatin -> genes in the region might be expressed

Dynamic process in the cell – changes in response to stimulus

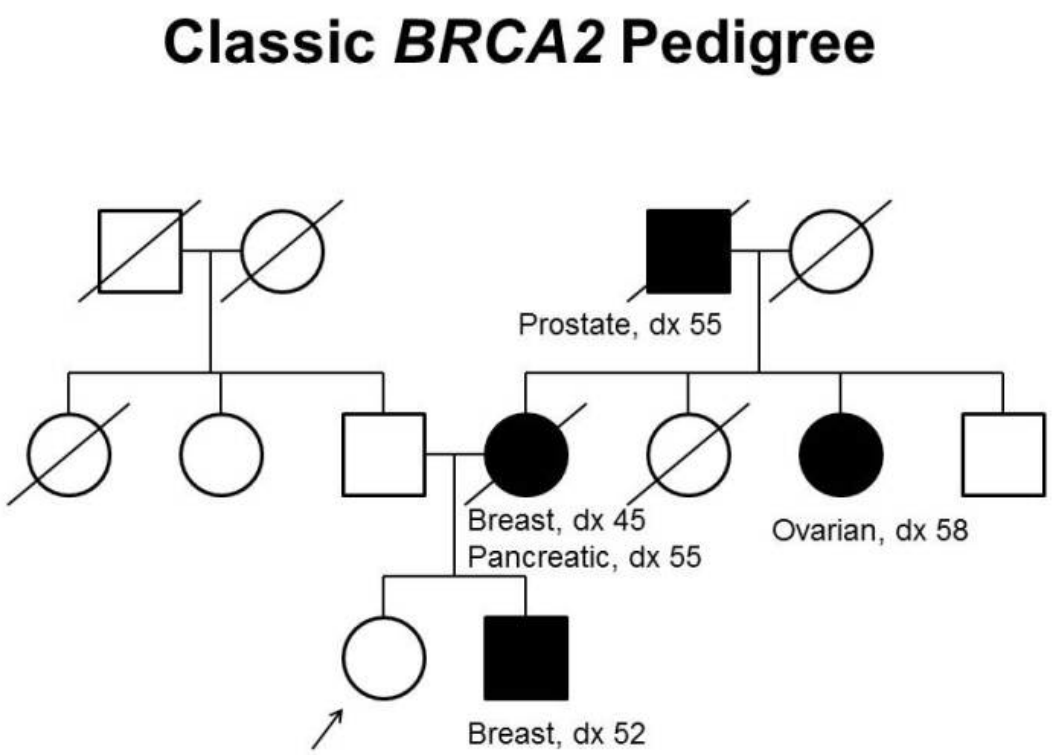
Finding and quantifying the impact of genetic variation on traits

Linkage analysis – fruit flies and simple traits



h/t Wikipedia entry on genetic linkage

Building linkage maps



Basic approach
Collect families
Genotype ‘microsatellites’ - variable length polymorphisms
Trace the inherited chunks of chromosomes
Find identity-by-descent

https://www.ncbi.nlm.nih.gov/books/NBK65767/figure/CDR0000062855__2586/

Linkage found the gene for many single gene disorders

Table 1: Examples of Human Diseases, Modes of Inheritance, and Associated Genes

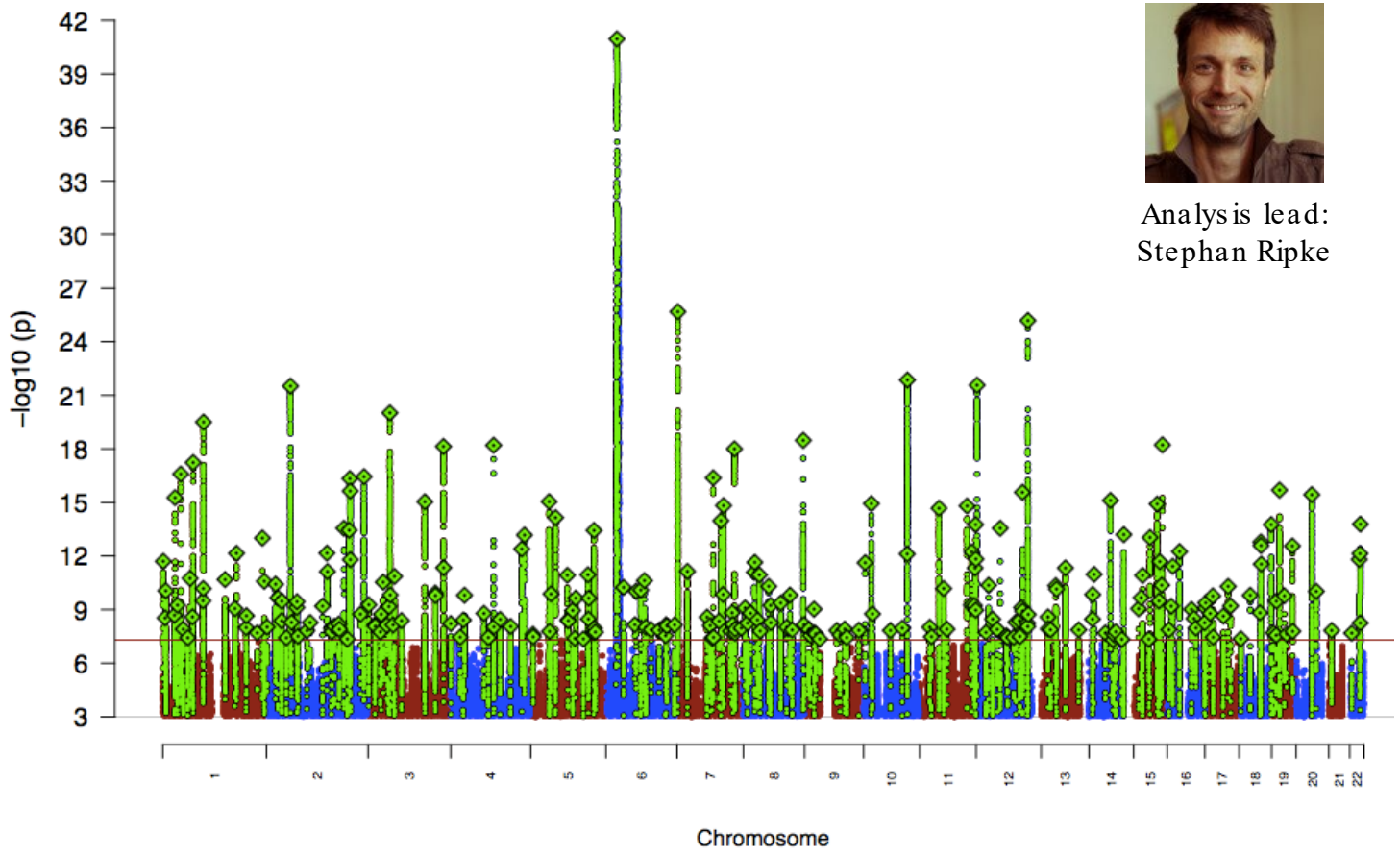
Disease	Type of Inheritance	Gene Responsible
Phenylketonuria (PKU)	Autosomal recessive	Phenylalanine hydroxylase (*PAH*)
Cystic fibrosis		
Sickle-cell anemia		
Albinism, oculocutan		
Huntington's disease		
Myotonic dystrophy t		
Hypercholesterolemi dominant, type B		
Neurofibromatosis, t		
Polycystic kidney disease 1 and 2	Autosomal dominant	Polycystic kidney disease 1 (*PKD1*) and polycystic kidney disease 2 (*PKD2*), respectively
Hemophilia A	X-linked recessive	Coagulation factor VIII (*F8*)
Muscular dystrophy, Duchenne type	X-linked recessive	Dystrophin (*DMD*)
Hypophosphatemic rickets, X-linked dominant	X-linked dominant	Phosphate-regulating endopeptidase homologue, X-linked (*PHEX*)
Rett's syndrome	X-linked dominant	Methyl-CpG-binding protein 2 (*MECP2*)
Spermatogenic failure, nonobstructive, Y-linked	Y-linked	Ubiquitin-specific peptidase 9Y, Y-linked (*USP9Y*)

However, linkage basically did not work for most complex traits with a handful of counterexamples including:
APOE for Alzheimer's Disease
BRCA for breast cancer
NOD2 in Crohn's Disease

ch as BRCA example
tance in families

Common variant discovery in schizophrenia

PGC schizophrenia working group



Analysis lead:
Stephan Ripke

PGC-SCZ chair:
Mick O'Donovan

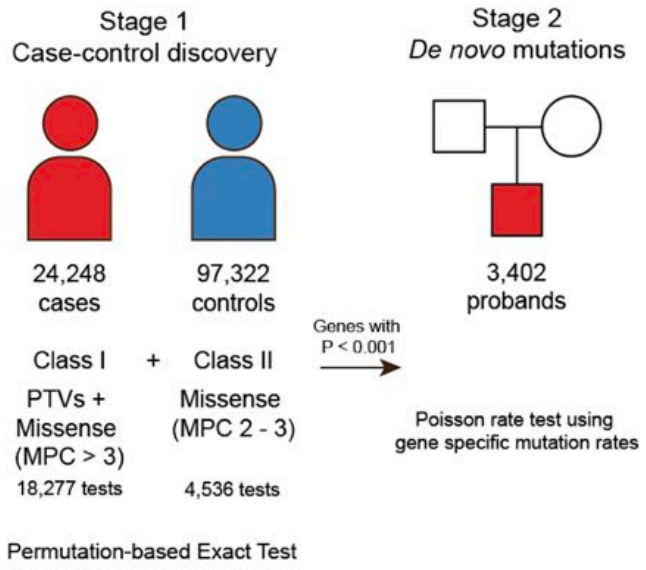
PGC Lead PI:
Pat Sullivan

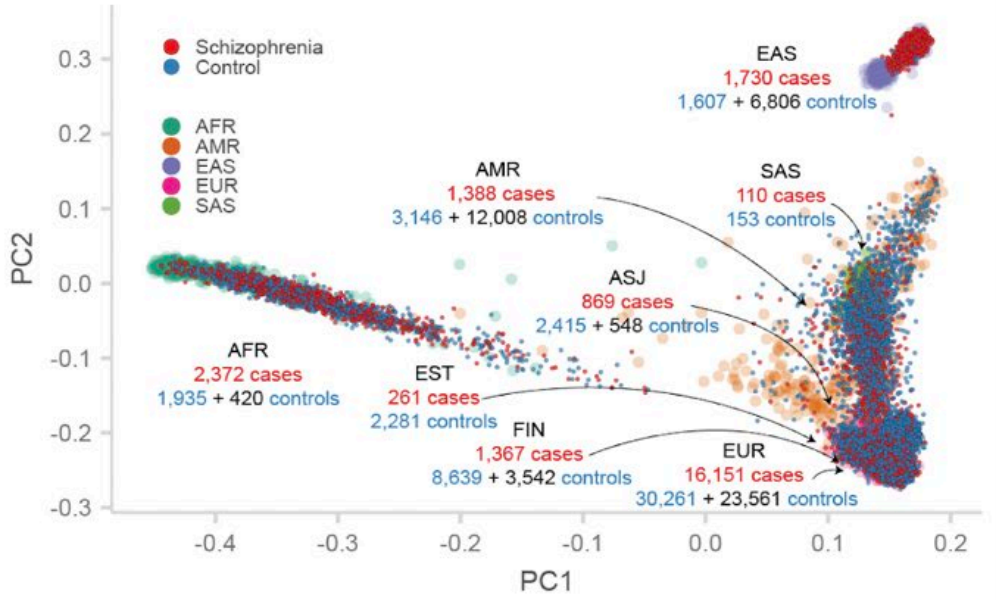
69,369 with schizophrenia
236,642 without Dx

270 loci

Schizophrenia exome meta-analysis (SCHEMA) data and definitions

TJ Singh



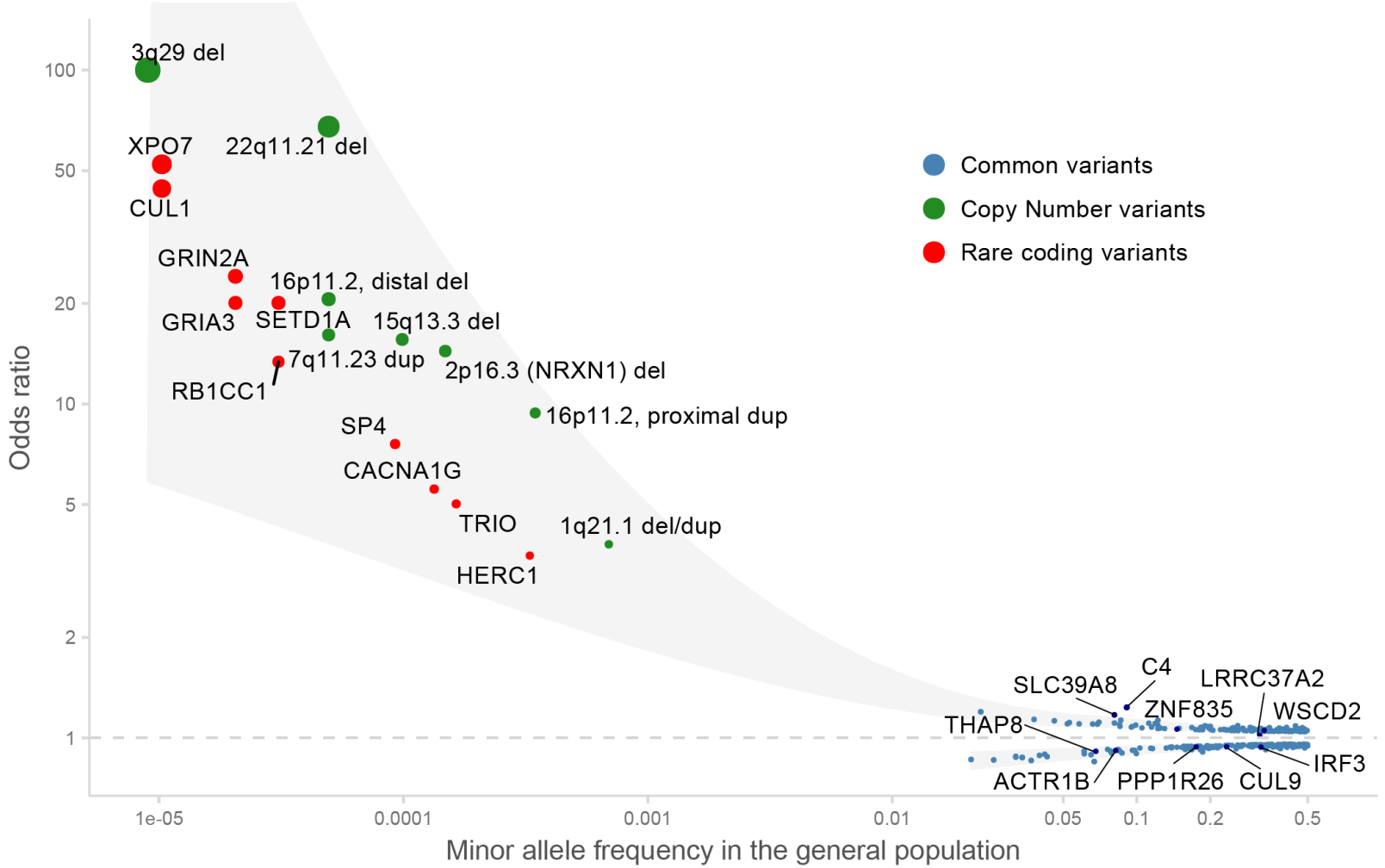


Article

Rare coding variants in ten genes confer substantial risk for schizophrenia

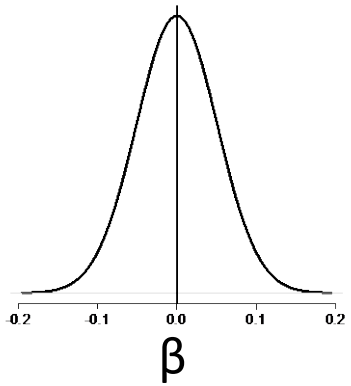
Nature, 2022

Known genetic architecture of schizophrenia



Polygenes!

Many small genetic effects ⟶ Can we develop a little further?



We can assume a distribution of SNP effects and now generate estimates of heritability

REPORT

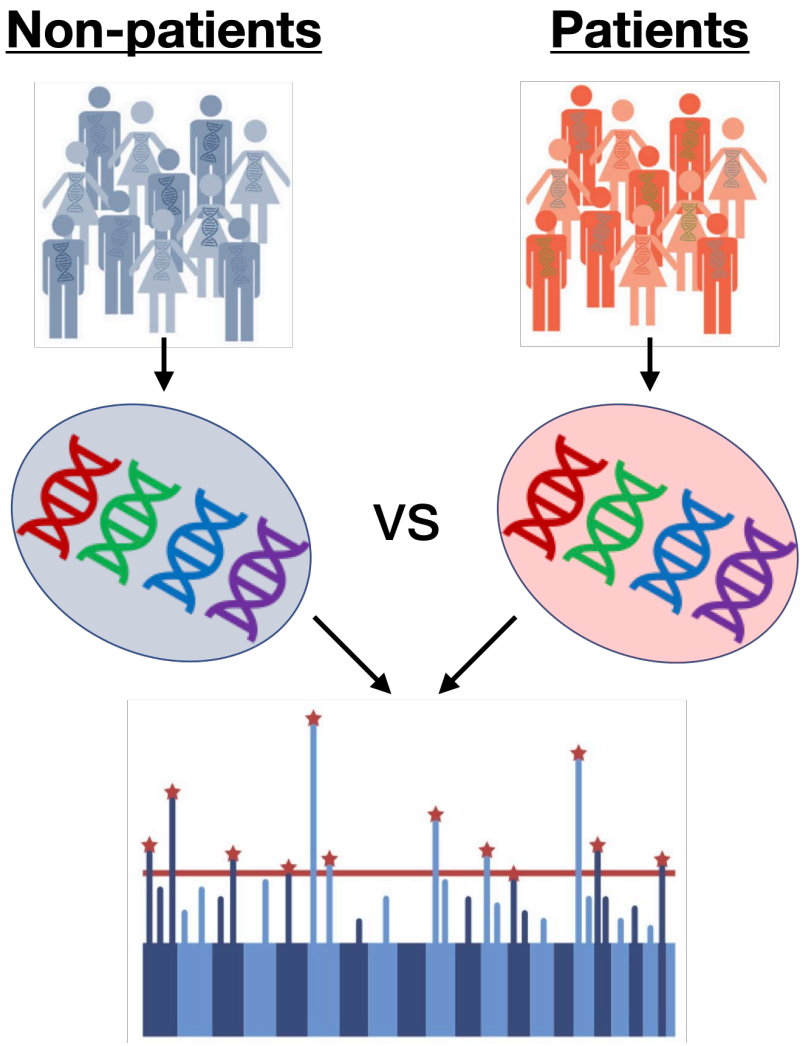
GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

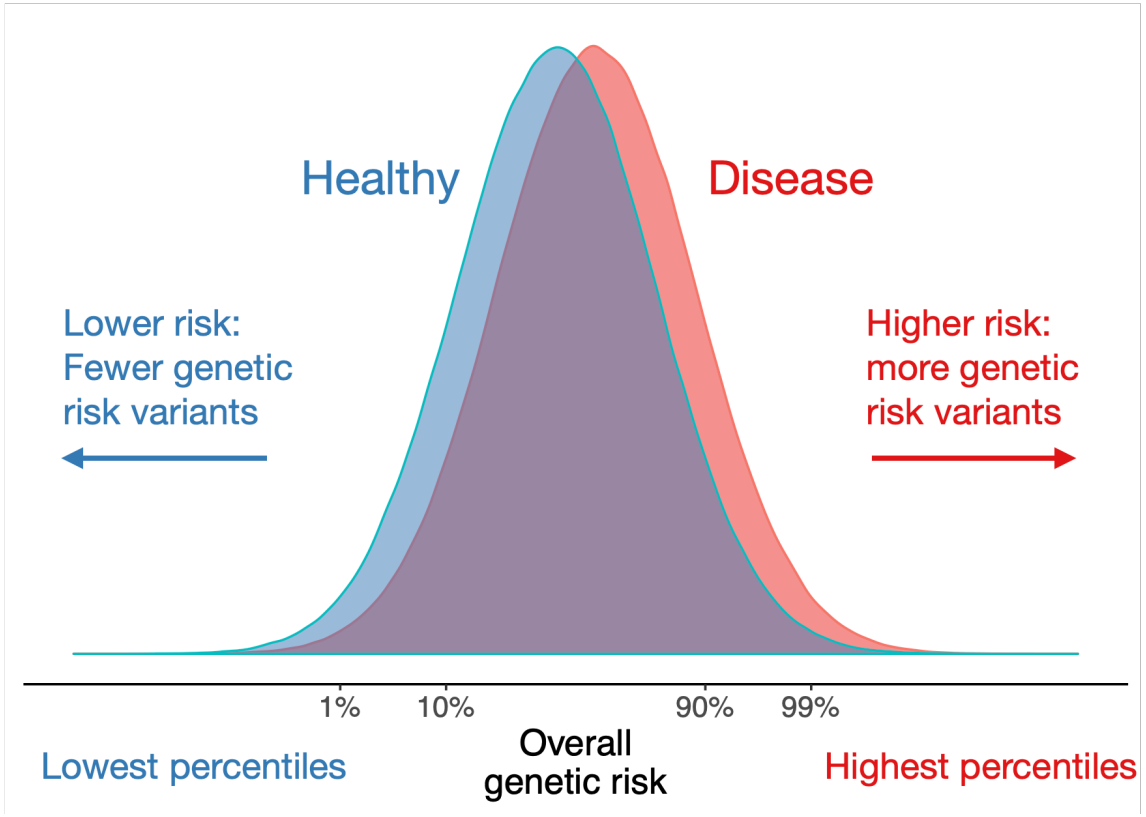
LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan[1–3], Po-Ru Loh[1,4], Hilary K Finucane[4,5], Stephan Ripke[2,3], Jian Yang[6], Schizophrenia Working Group of the Psychiatric Genomics Consortium[7], Nick Patterson[1], Mark J Daly[1–3], Alkes L Price[1,4,8] & Benjamin M Neale[1–3]

The rise of the polygenic score

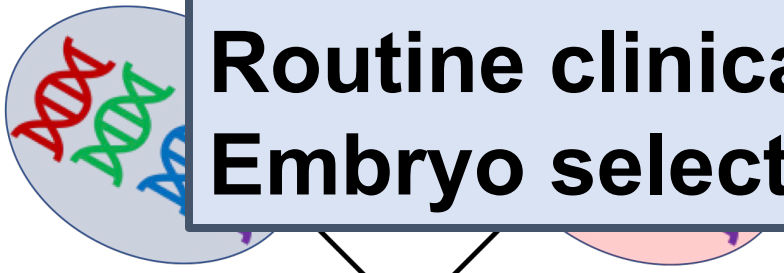


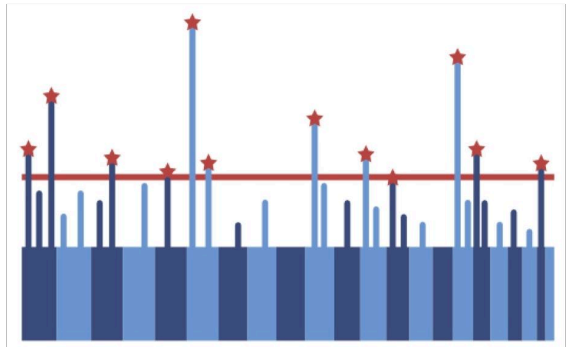
Polygenic risk score: risk prediction of an individual's phenotype from DNA



The rise of the polygenic score

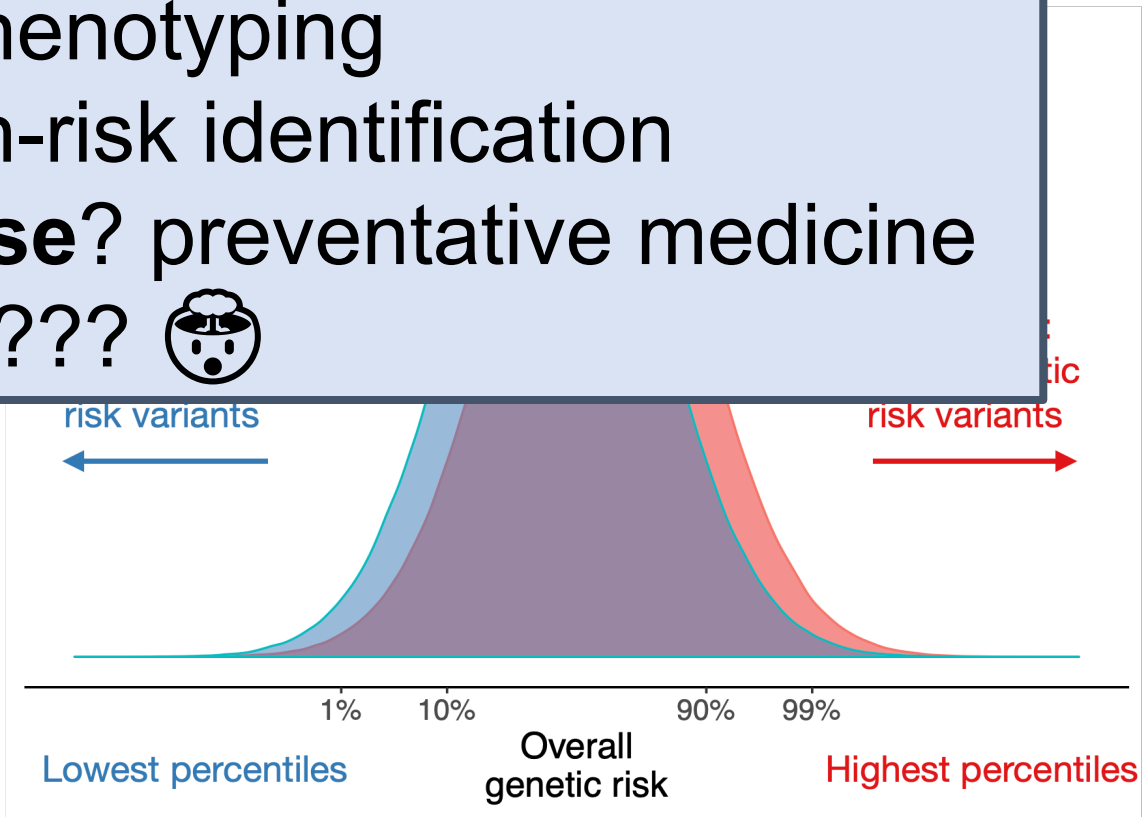
Non-patients **Patients**





Polygenic risk score: risk prediction of an individual's phenotype from DNA

Research: deep phenotyping
Clinical trials: high-risk identification
Routine clinical use? preventative medicine
Embryo selection??? 😲



Some questions that you might learn how to answer over the course

Can we quantify the impact genetic variation has on trait variation?	How do we analyze unrelated individuals?	Is it possible to estimate individual genetic risk?
What genes and variants matter?	How do we analyze family data?	Can we find causal relationships using genetics?
How do we protect against artifacts and confounds in genetic analysis?	What do associated variants do biologically?	How do we analyze rare genetic variation?