Finding genes for complex traits: the GWAS revolution

Nick Martin
Queensland Institute of Medical Research
Brisbane

21st Anniversary Workshop
Leuven
August 11, 2008

22nd International Statistical Genetics Methods Workshop

- Michael Neale (director) 
- Nick Martin (symposium) 
- Dorret Boomsma 
- John Hewitt 
- Stacey Cherny 
- Lindon Eaves 
- Jeff Lessem 
- Danielle Posthuma 
- Ben Neale 
- Meike Bartels 
- David Evans 
- Sarah Medland 

- Hermine Maes (host) 
- Martine Thomis (host) 
- John Rice 
- Peter Visscher 
- Jaakko Kaprio 
- Danielle Dick 
- Kate Morley 
- William Valdar 
- Matt Keller 
- Tim York 
- William Stewart 
- Irene Rebollo 

Previous Workshop Attendance

	Year	Place	Type	#Fac	#Stud
TC1	87	L	I	10	24
TC2	89	L	I	11	41
TC3	90	B	I	11	28
TC4	91	L	I	14	49
			A	12	55
TC5	93	B	I	13	49
TC6	94	B	I	16	43
TC7	95	H	I	10	29
TC8	96	B	I	10	49
TC9	97	B	I	10	55
TC10	98	B	I	12	57
TC11	98	L	I	10	55
			A	13	62
TC12	99	B	A	12	37

	Year	Place	Type	#Fac	#Stud
TC13	00	B	I	12	63
TC14	01	B	A	18	65
TC15	02	B	I	18	95
TC16	03	B	A	15	82
TCE1	*03*	*E*	*I*	*15*	*65*
TC17	04	B	I	18	90
TCE2	*04*	*E*	*A*	*16*	*64*
TC18	05	B	A	18	64
TCE3	*05*	*E*	*A*	*13*	*55*
TC19	06	B	I	15	93
TCE4	*06*	*E*	*A*	*12*	*48*
TC20	07	B	A	20	55
TC21	08	B	I	19	95
TC22	08	L	A	27	57

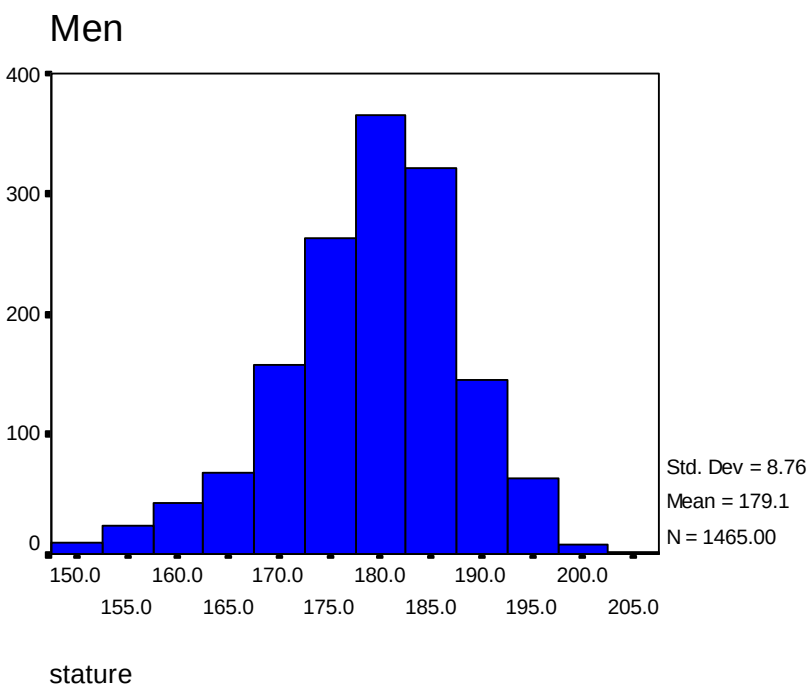
L: Leuven, B: Boulder, H: Helsinki, E: Egmond

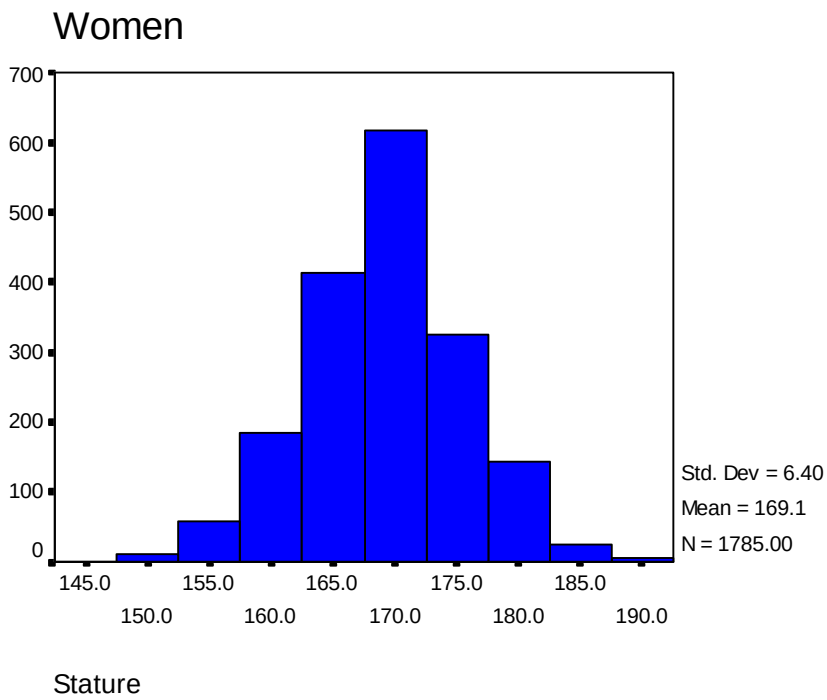
I: Introductory, A: Advanced

Number of Individuals

requency	1	2	3	4->6	7->9	.0->12	.3->21	22-23	27-28	
Faculty	13	6	3	10	9	3	4	5		**53**
Students	647	185	46	28	2		# of 'Unique' Students			**908**
including Egmond										
Faculty	14	6	3	10	10	7	4	1	4	**59**
Students	709	222	64	42	3		# of 'Unique' Students			**1040**

Variation (individual differences): Stature (in cm) in Dutch adolescent twins



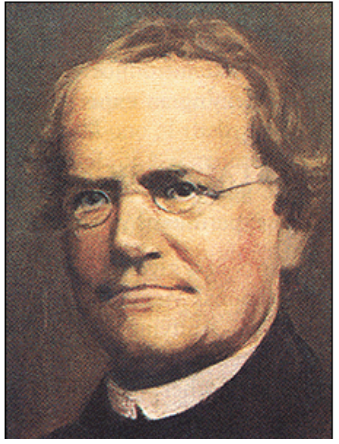


Individual differences in human characteristics, e.g. normal and abnormal behavior

Caused by:

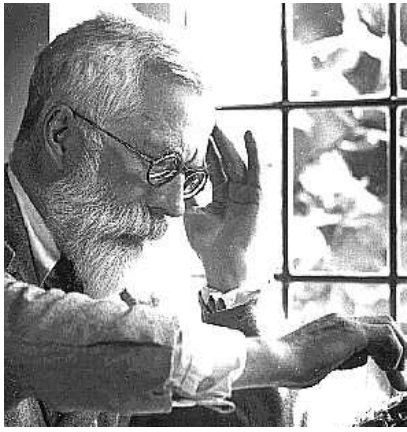
- differences in genotype (G)?
- differences in environment (E)?
- interaction between G and E?

Mendel: Laws of inheritance for monogenic traits:
1 Segregation
2 Independent Assortment



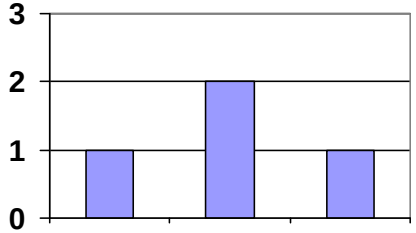
Galton: correlations between family members for continuous traits: Family & Twin Resemblance.

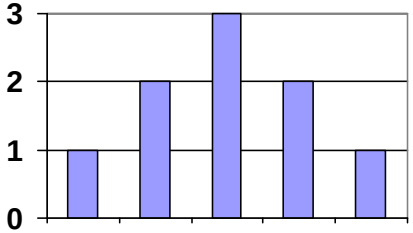
Fisher: traits can be influenced by more than one gene (which each can have small effects). Effects of genes add up and lead to a normal distribution in the population.

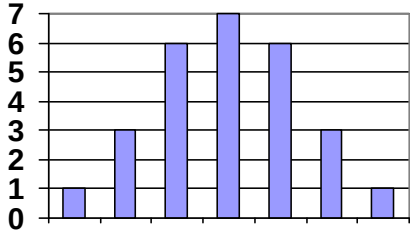


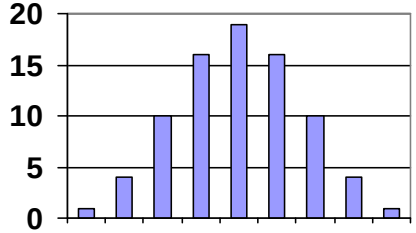
Complex: Polygenic Traits

1 Gene	2 Genes	3 Genes	4 Genes
→ 3 Genotypes → 3 Phenotypes	→ 9 Genotypes → 5 Phenotypes	→ 27 Genotypes → 7 Phenotypes	→ 81 Genotypes → 9 Phenotypes

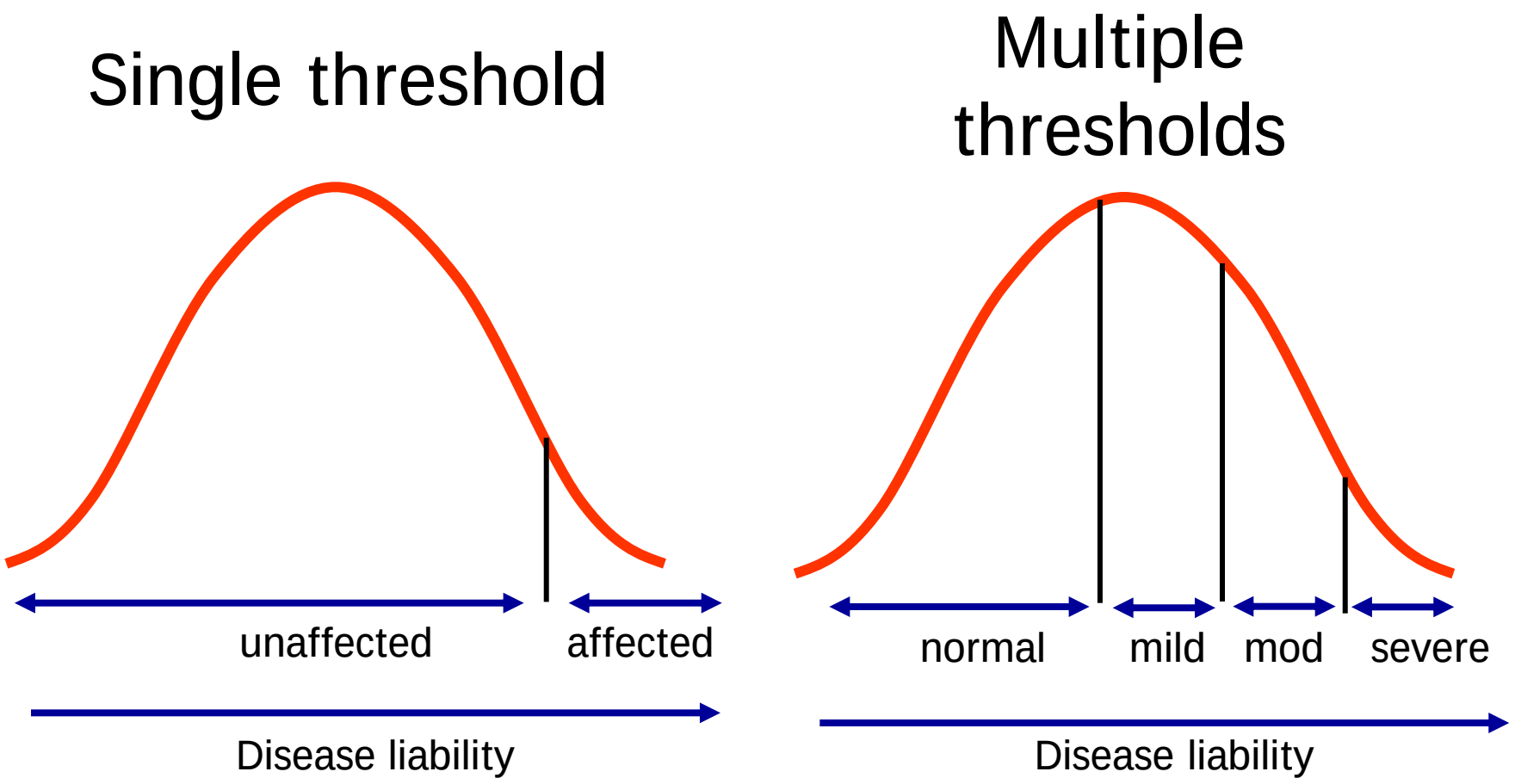






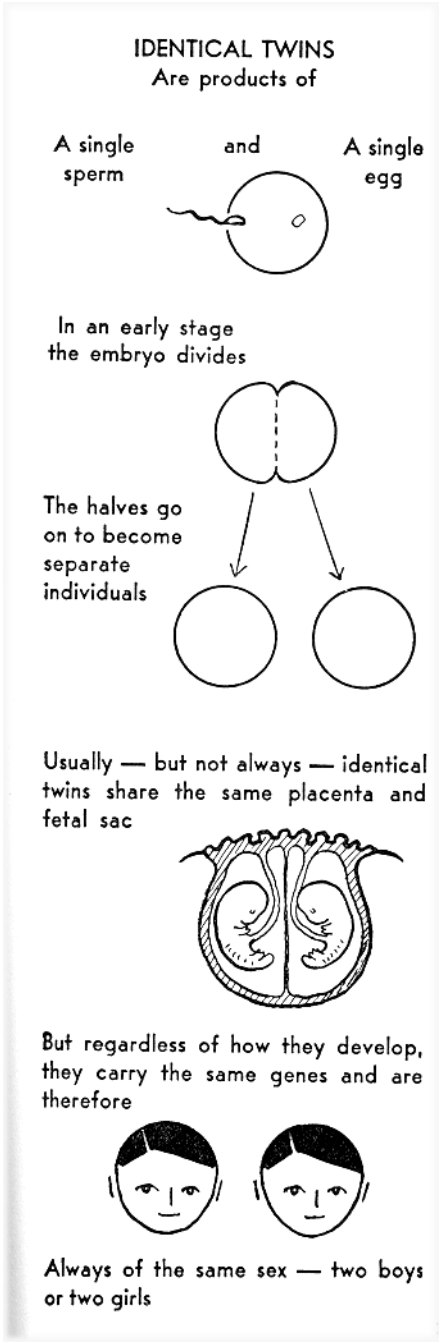


Multifactorial Threshold Model of Disease



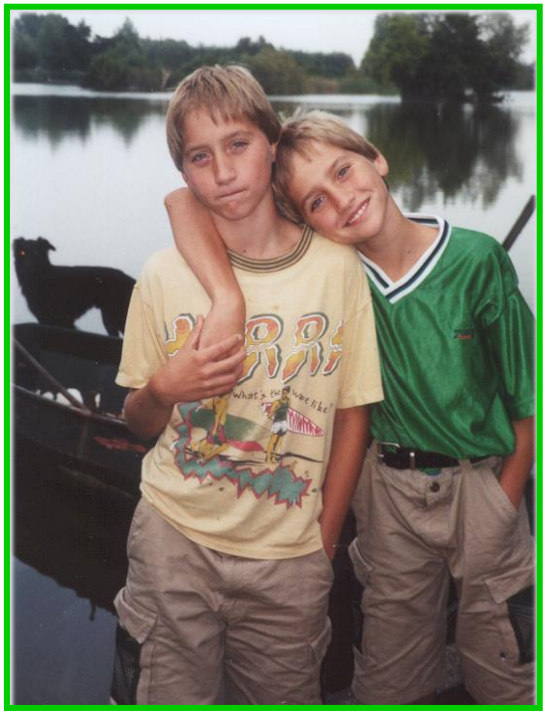
Designs to disentangle G + E

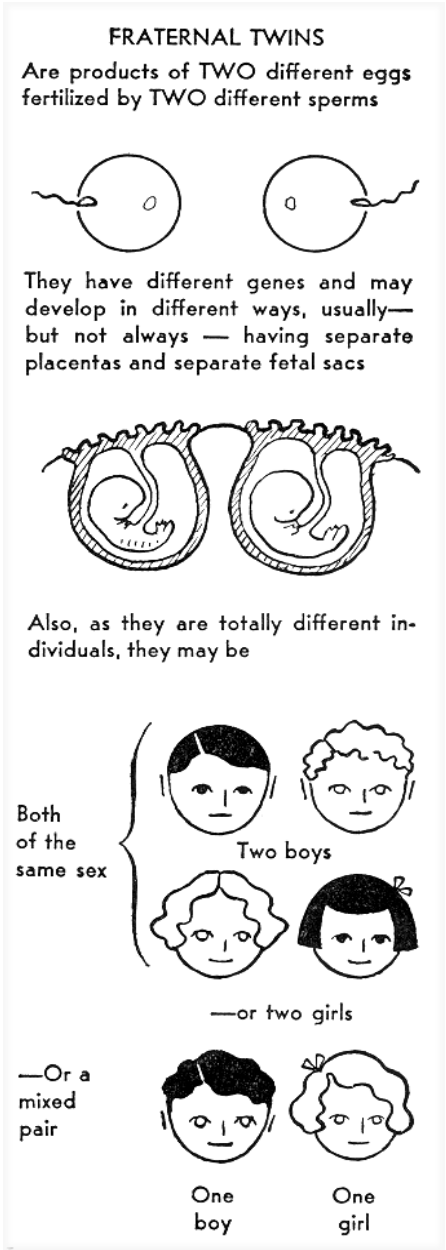
- Family studies – G + C confounded
- MZ twins alone – G + C confounded
- MZ twins reared apart – rare, atypical, selective placement ?
- Adoptions – increasingly rare, atypical, selective placement ?
- MZ and DZ twins reared together
- Extended twin design



‘Identical’ twins

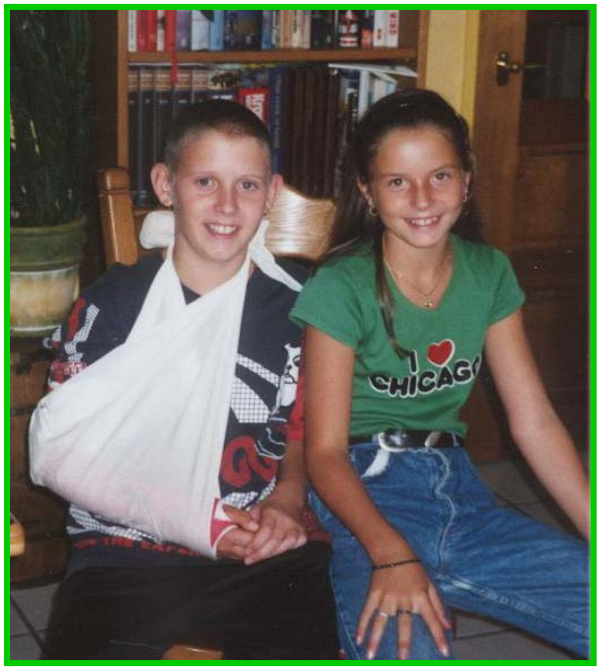
Monozygotic (MZ) twins:
~100% genetically identical



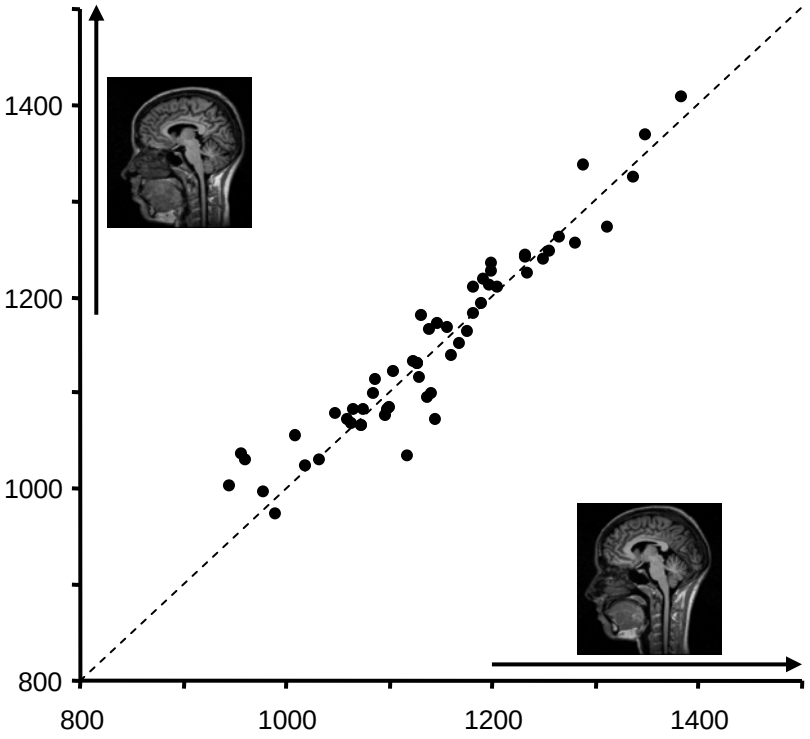


Fraternal twins

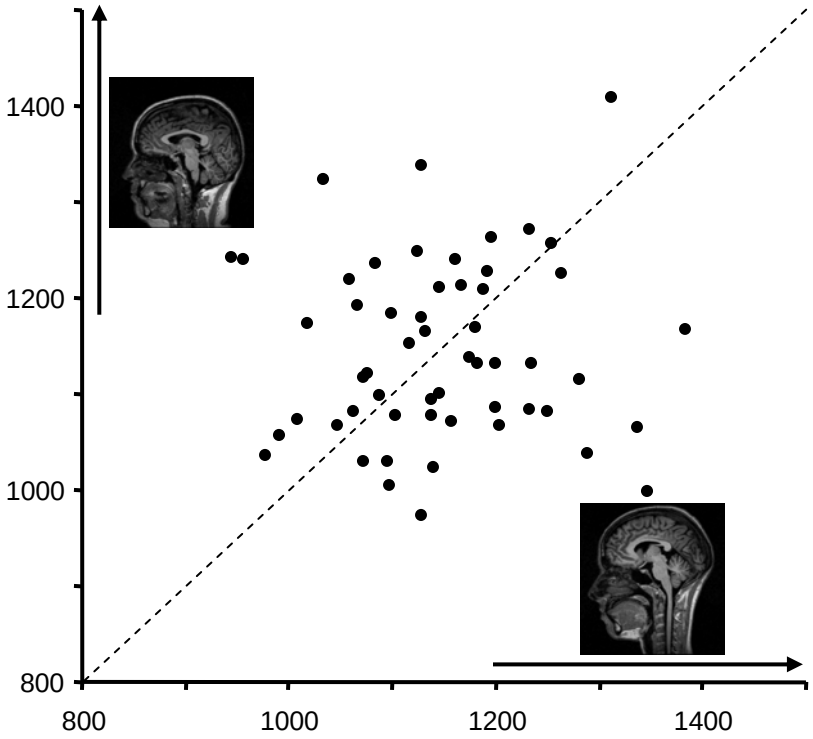
Dizygotic (DZ) twins share ~50% of their segregating genes



Brain volumes: resemblance of MZ and DZ twins



Brain volume MZ twin pairs (milliliter) in twin and co-twin



Brain volume DZ twin pairs (milliliter) in twin and co-twin

Bouchard & McGue: Genetic and environmental influences on human psychological differences (2003)

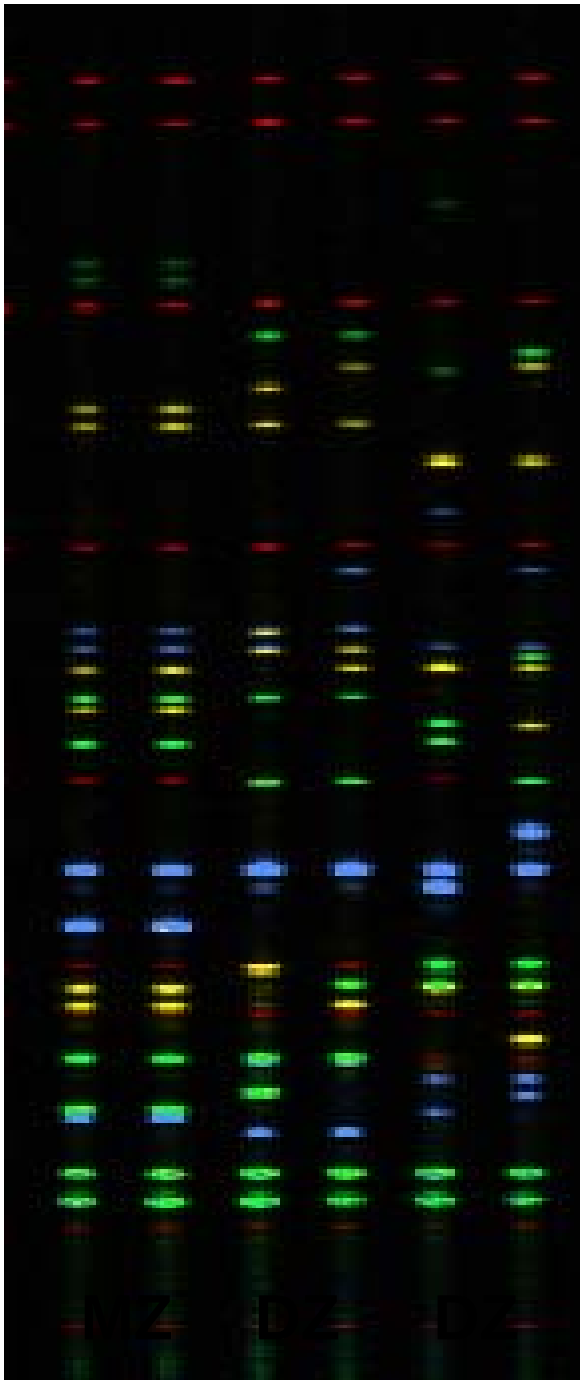
	Intraclass correlations	
	MZT (626 pairs)	MZA (74 pairs)
Positive emotionality	.55	.43
Negative emotionality	.44	.47
Constraint	.56	.58

Classical twin design: Assumptions

Zygosity is known accurately

Twins are representative of the general population

MZs have experienced the same environments as DZs (including prenatal) – the equal environments assumption (EEA)



Zygosity

DZ = opposite sex !

DZ = very unlike in appearance

DZ = different at marker loci

(except for measurement error)

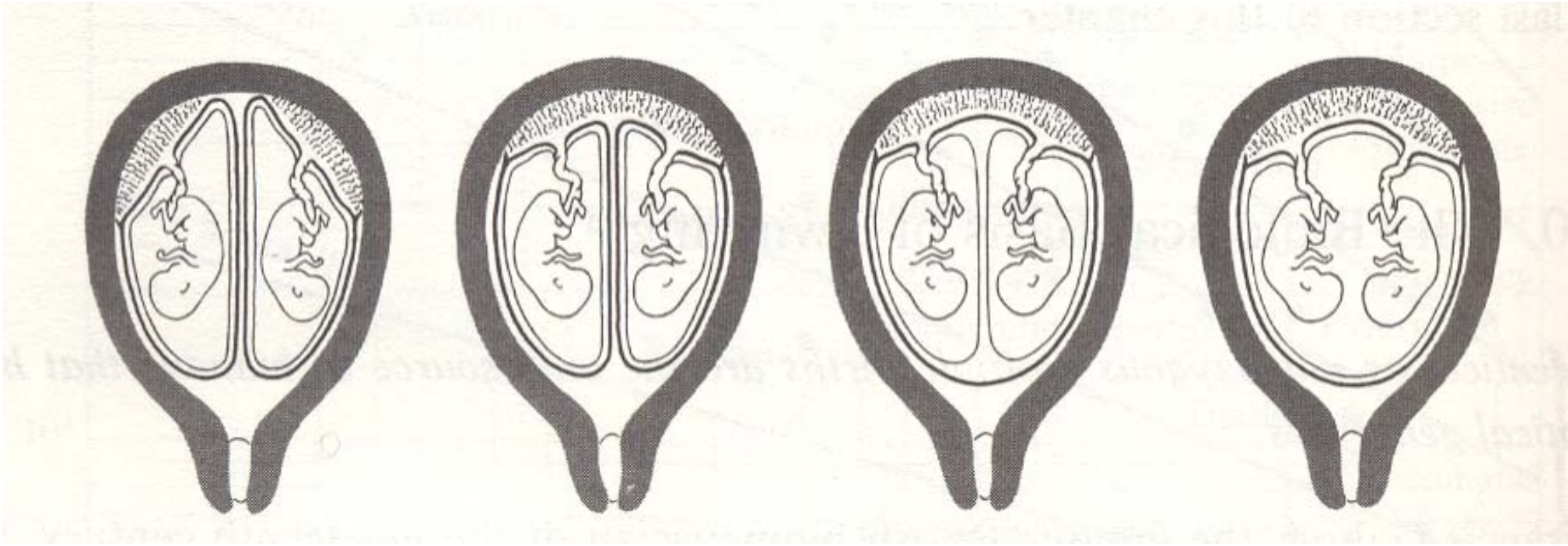
MZ = mono-chorionic

MZ = identical at marker loci

(except for rare mutations)

MZ and DZ twins: determining zygosity using ABI Profiler™ genotyping (9 STR markers + sex)

Placentation and zygosity (EEA?)



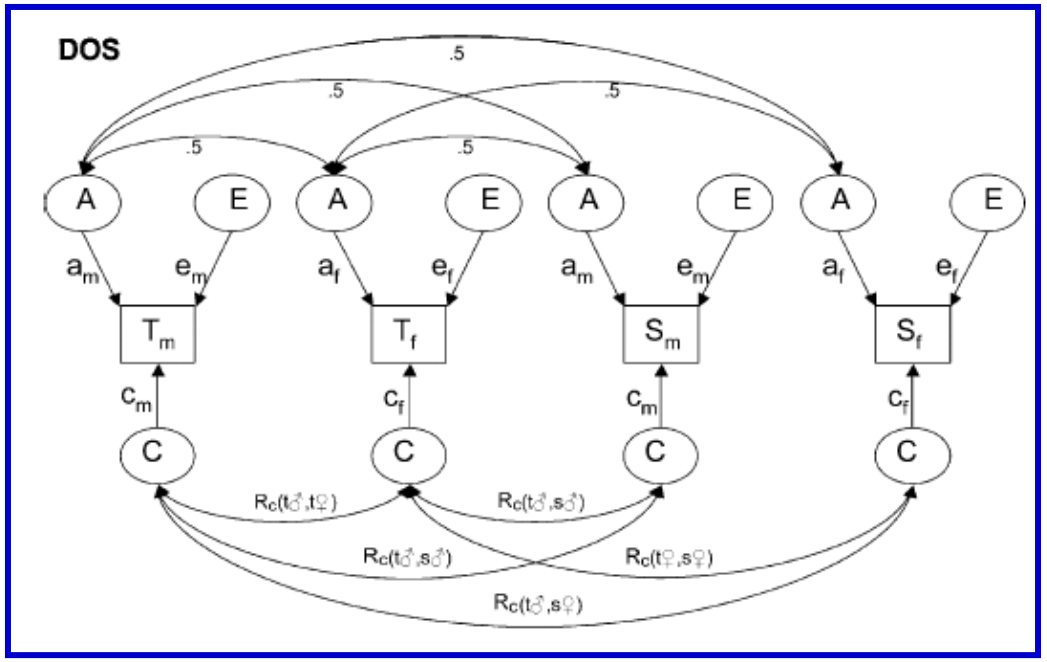
Dichorionic Two placentas	Dichorionic Fused placentas	Monochorionic Diamniotic	Monochorionic Monoamniotic
MZ 19% **DZ 58%**	**MZ 14%** **DZ 42%**	**MZ 63%** **DZ 0%**	**MZ 4%** **DZ 0%**

Representative?

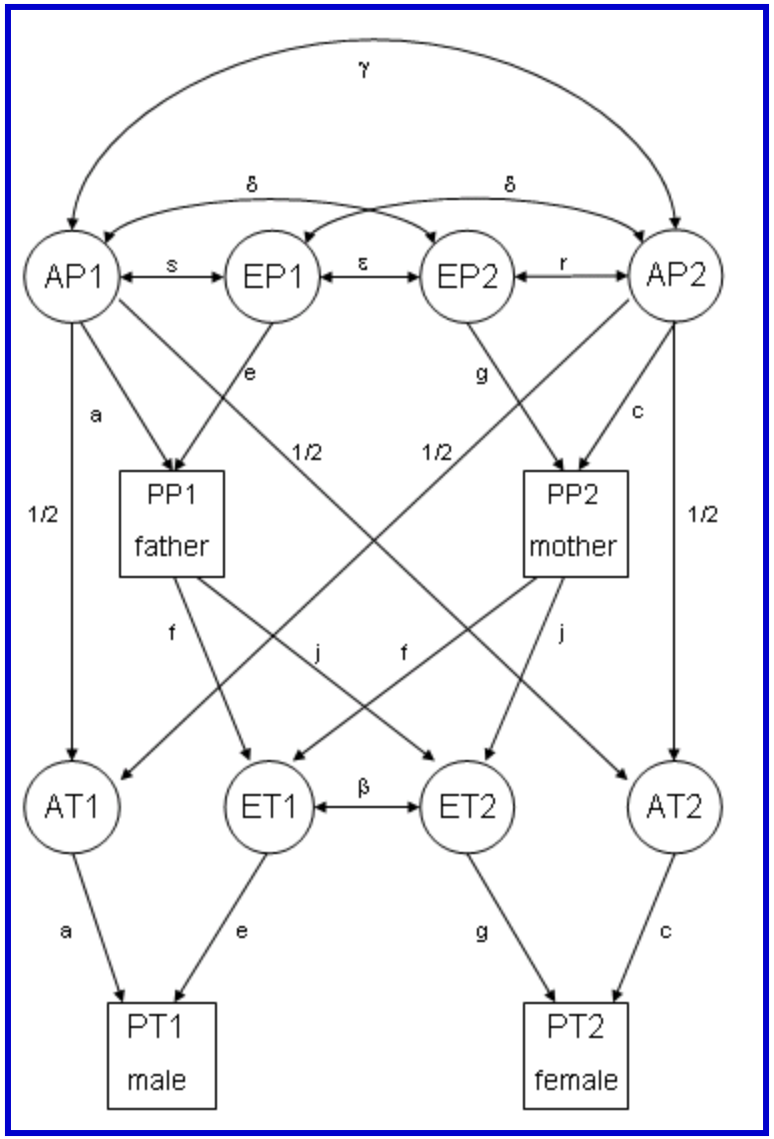
- Test for “twin effects”: Include other family members (e.g. siblings of twins)

- Look at resemblance in twins of mistaken zygosity (parents say DZ, testing says MZ)

Extended twin designs



Twin and sibs: tests of special twin effects;
increased power to detect Common environment, Non-additive genetic effects



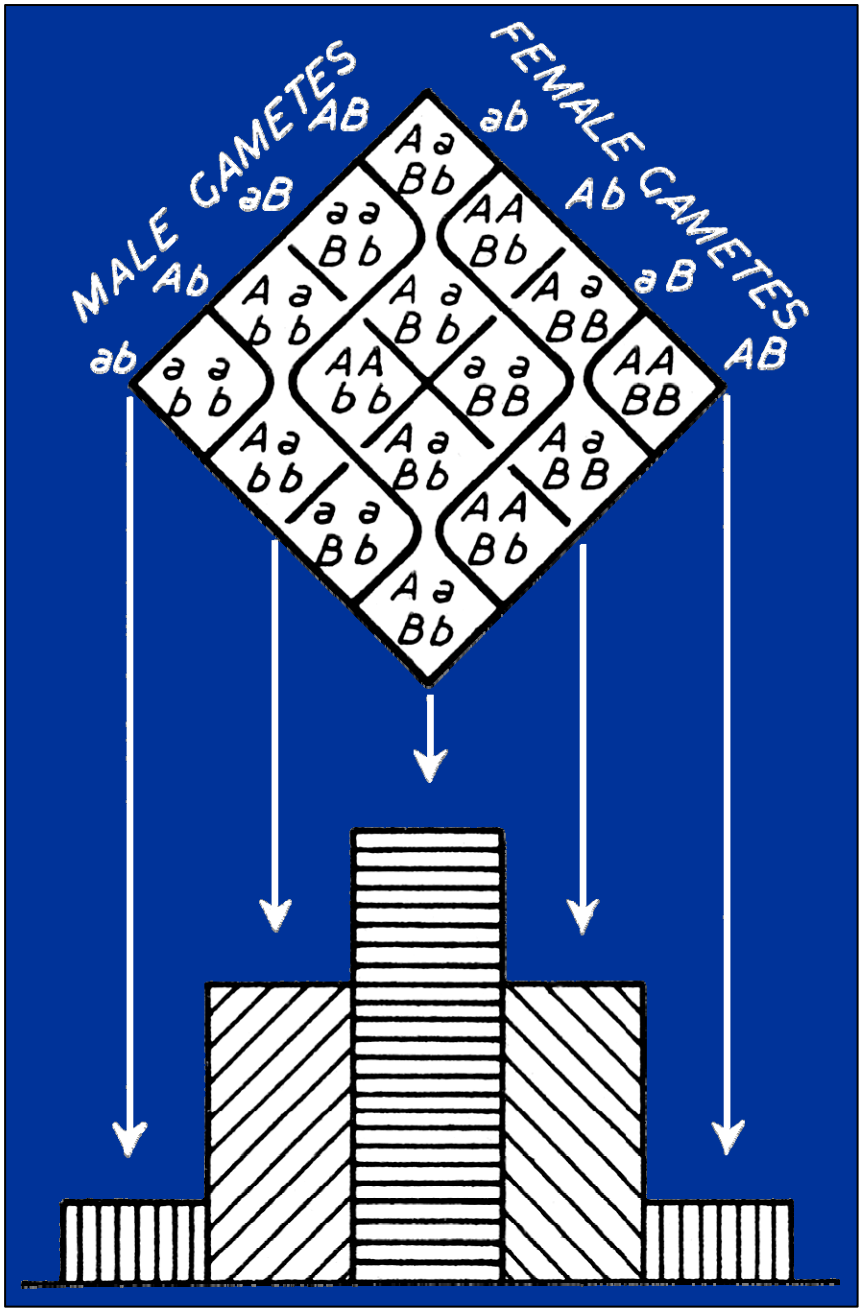
Twin and parents: genetic and cultural transmission, GE correlation, assortment

Resemblance between relatives caused by:

- **shared Genes (G = A + D)**
- **environment Common to family members (C)**

Differences between relatives caused by:

- **non-shared Genes**
- **Unique environment (U or E)**



Punnett square

Genetics explains both the *resemblances* and the *differences* of family members (e.g. sibs).

Distribution of phenotypes in offspring of two heterozygous parents (AaBb).
(2 genes (A & B) with additive allelic effects).

K Mather, Biometrical Genetics, Dover Publ, 1949

what is a gene?

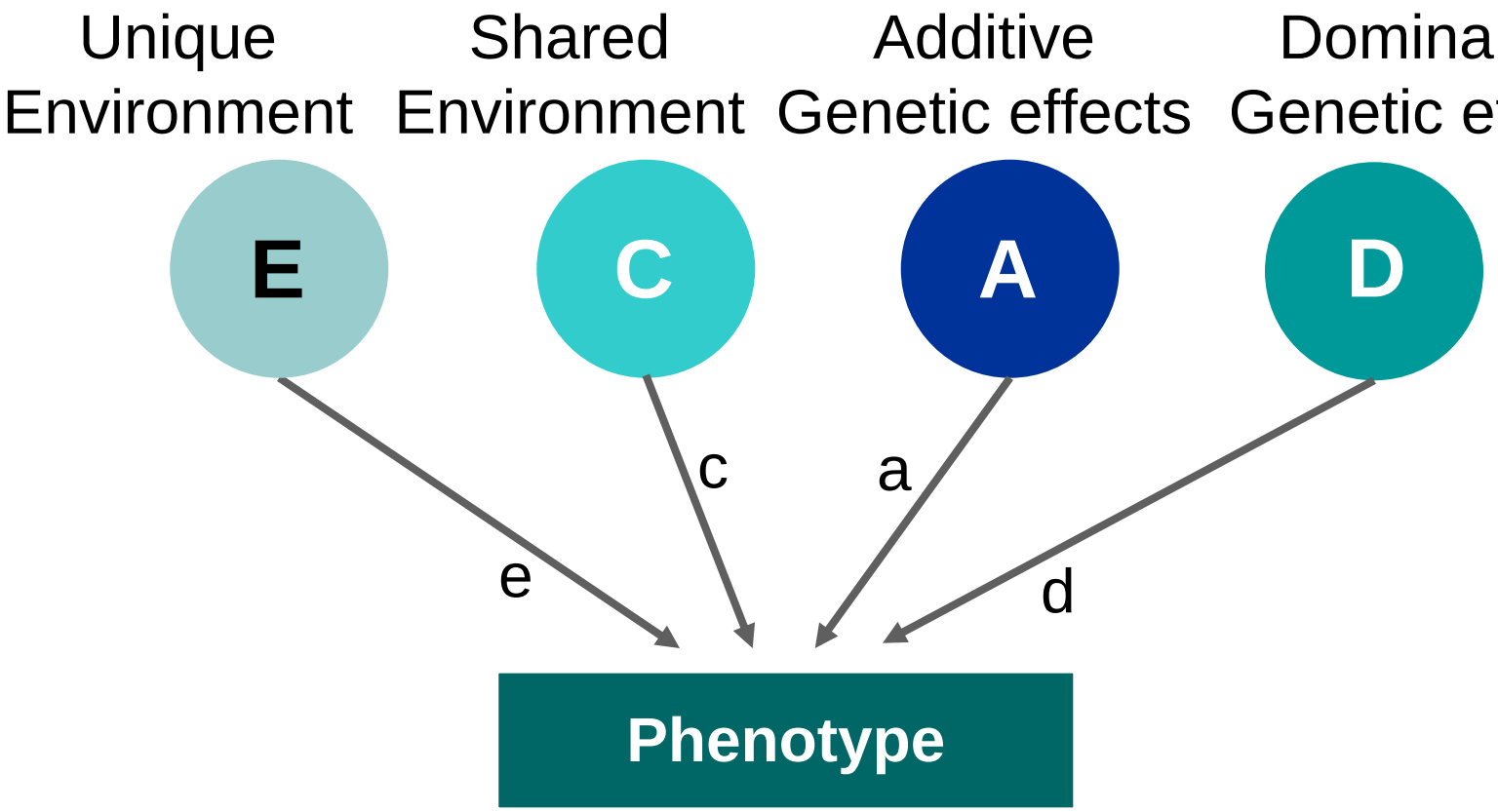
In 2003, estimates from gene-prediction programs suggested there are 24,500 or fewer protein-coding genes.

The Ensembl genome-annotation system estimates them at 23,299. Perhaps the biggest obstacle to gene counting is that the definition of a gene is unclear.

Is a gene:

- a heritable unit corresponding to an observable phenotype
- a packet of genetic information that encodes a protein
- a packet of genetic information that encodes RNA
- must it be translated ?
- are genes genes if they are not expressed ?

TK Attwood: The Babel of Bioinformatics, Science, 290:471, 2000



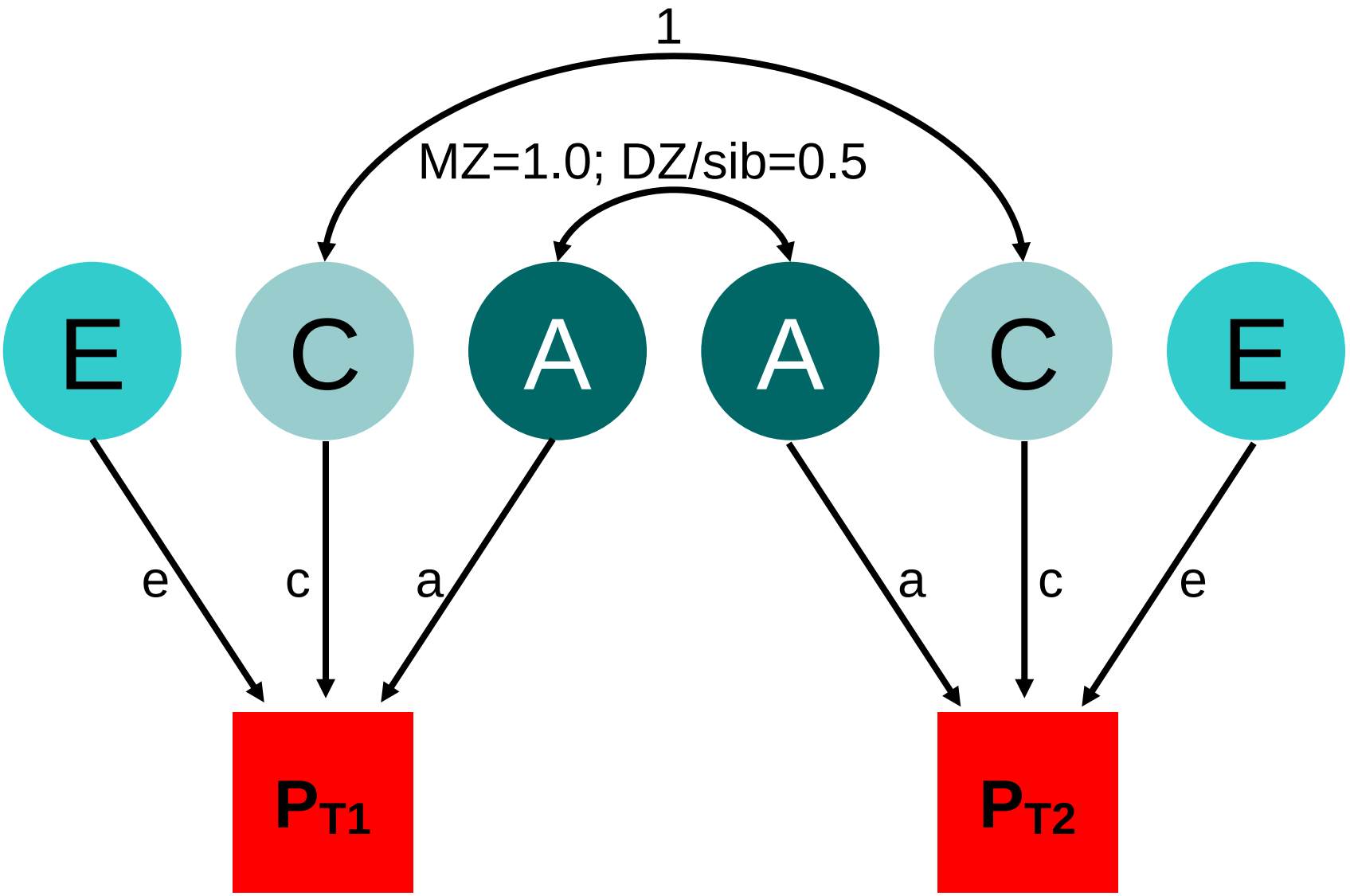
$$P = eE + aA + cC + dD$$

(plus epistasis, assortment, GE interaction, ….)

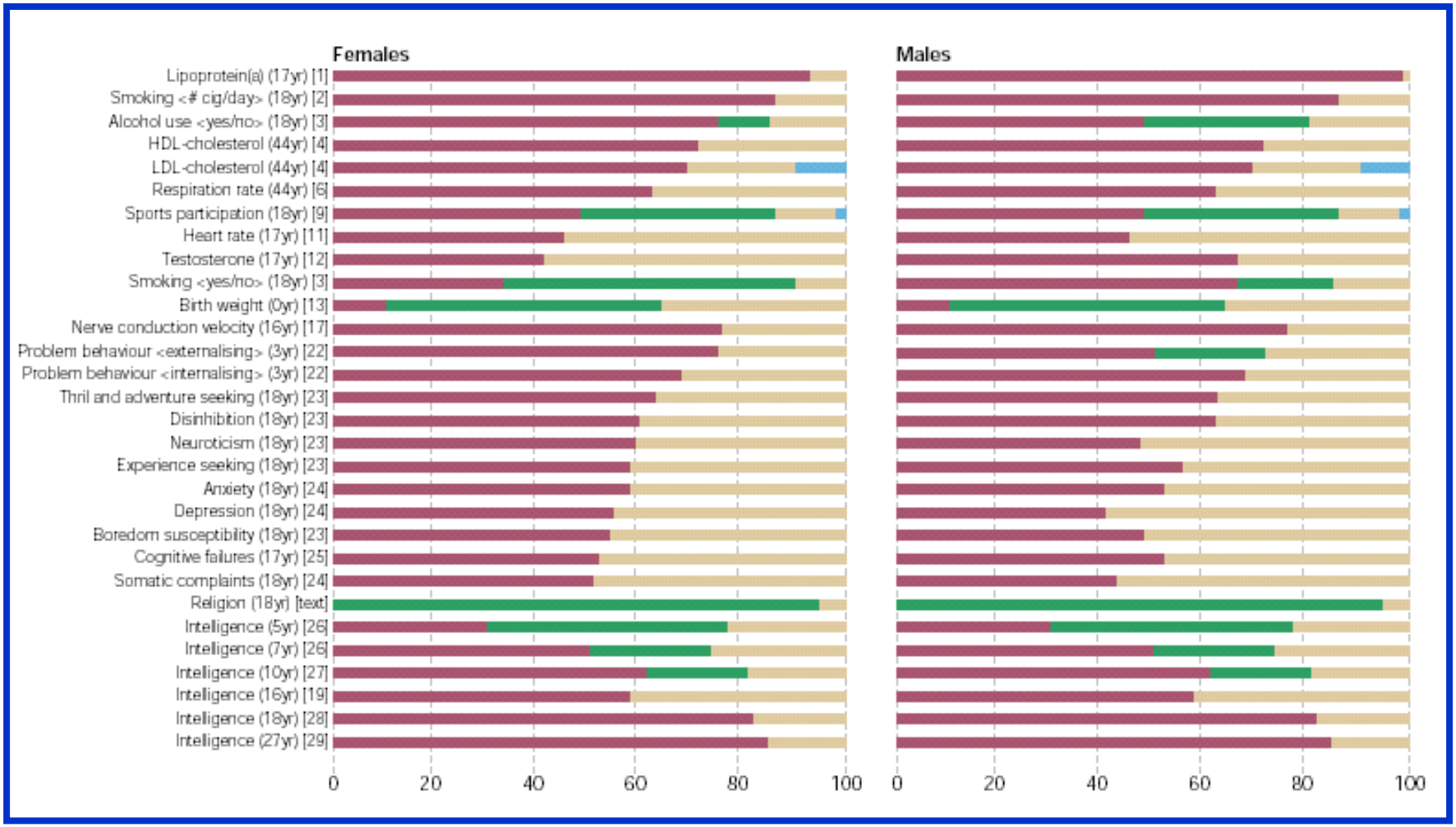
Structural equation modeling

- Both continuous and categorical variables
- Systematic approach to hypothesis testing
- Tests of significance (for effects of G, D, C)
- Can be extended to:
  - More complex questions
  - Multiple variables
  - Other relatives

ACE Model for univariate twin / sib data



Heritability estimates in males and females (ANTR twin data)



Genes

Shared environment

Unique environment

Boomsma et al., 2002, Nat Review Genet

How twin studies changed research agendas

1. Autism – “caused by cold mothers”
   10/11 MZ pairs concordant vs. 2/11 DZs

2. ADHD – “caused by food dyes”
   Twin studies found h^2 ~0.8

3. Multiple sclerosis - “caused by a virus”
   MZ concordance 26%, DZ concordance 2%

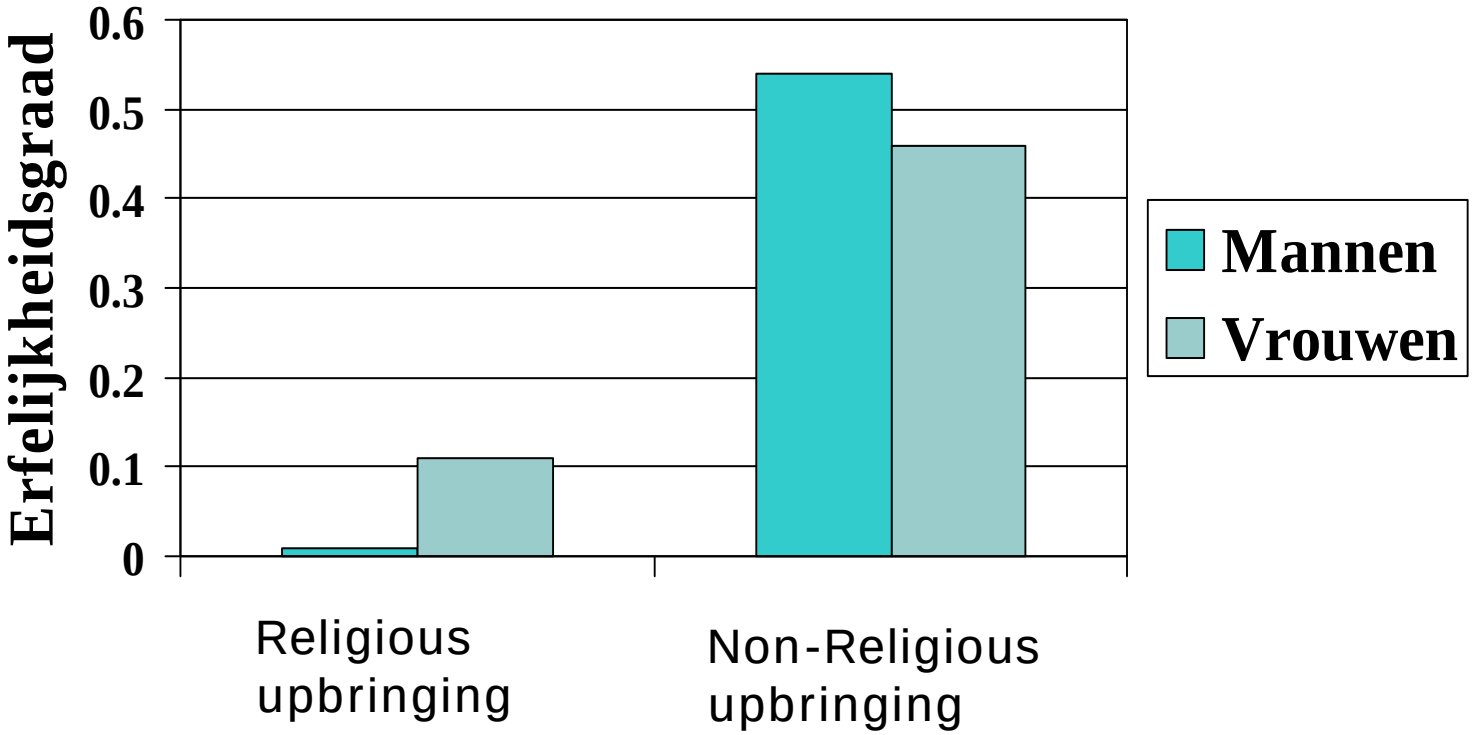
Martin, Boomsma & Machin (1997) *Nature Genetics* 17: 387

Types of Twin Studies I

Classical MZ -DZ comparison:

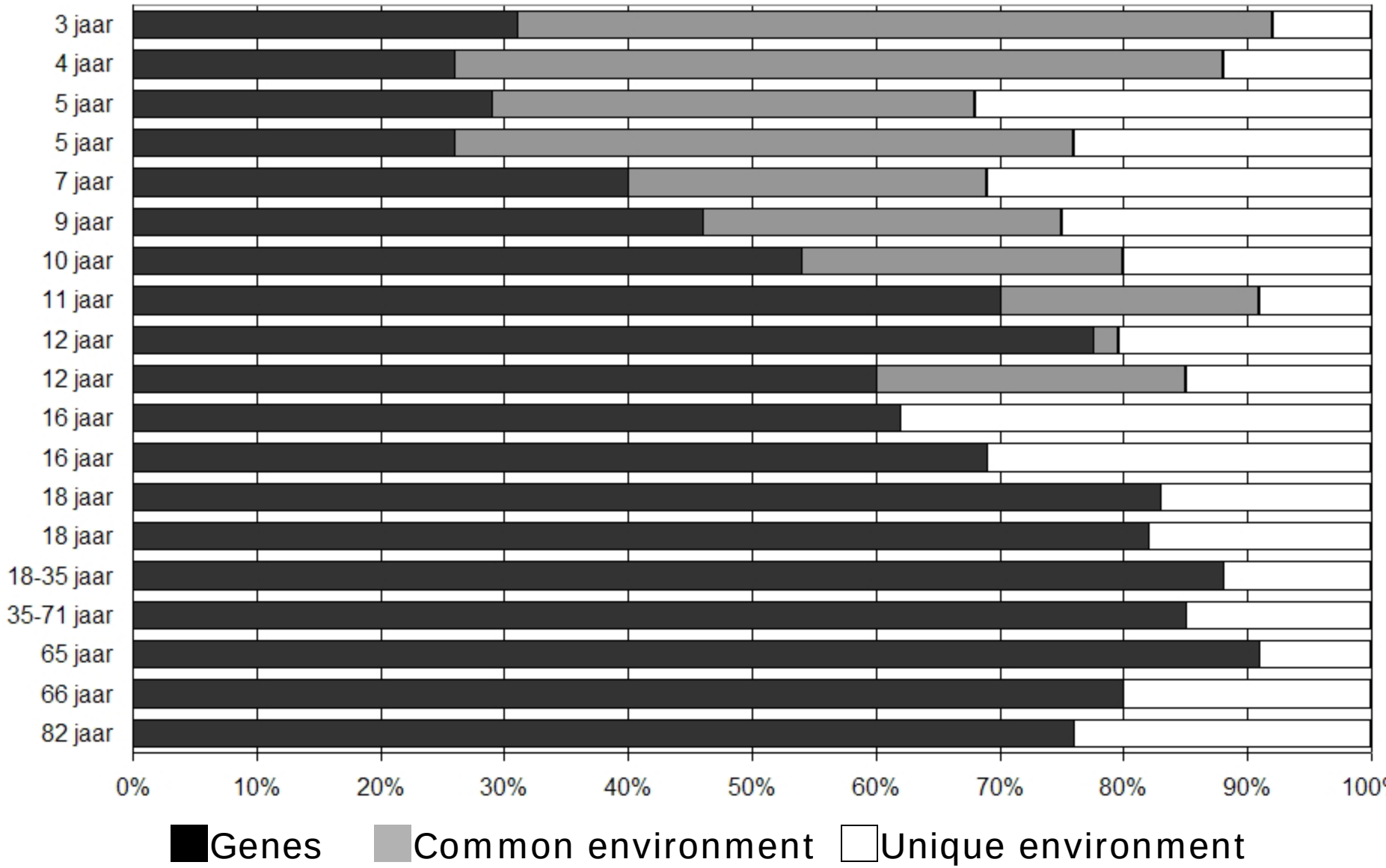
- age differences in heritability
- sex differences in heritability
- genotype x environment interaction
- causal models
- multivariate genetic analyses

Genotype x Environment interaction: Heritability of Disinhibition as a function of religious upbringing

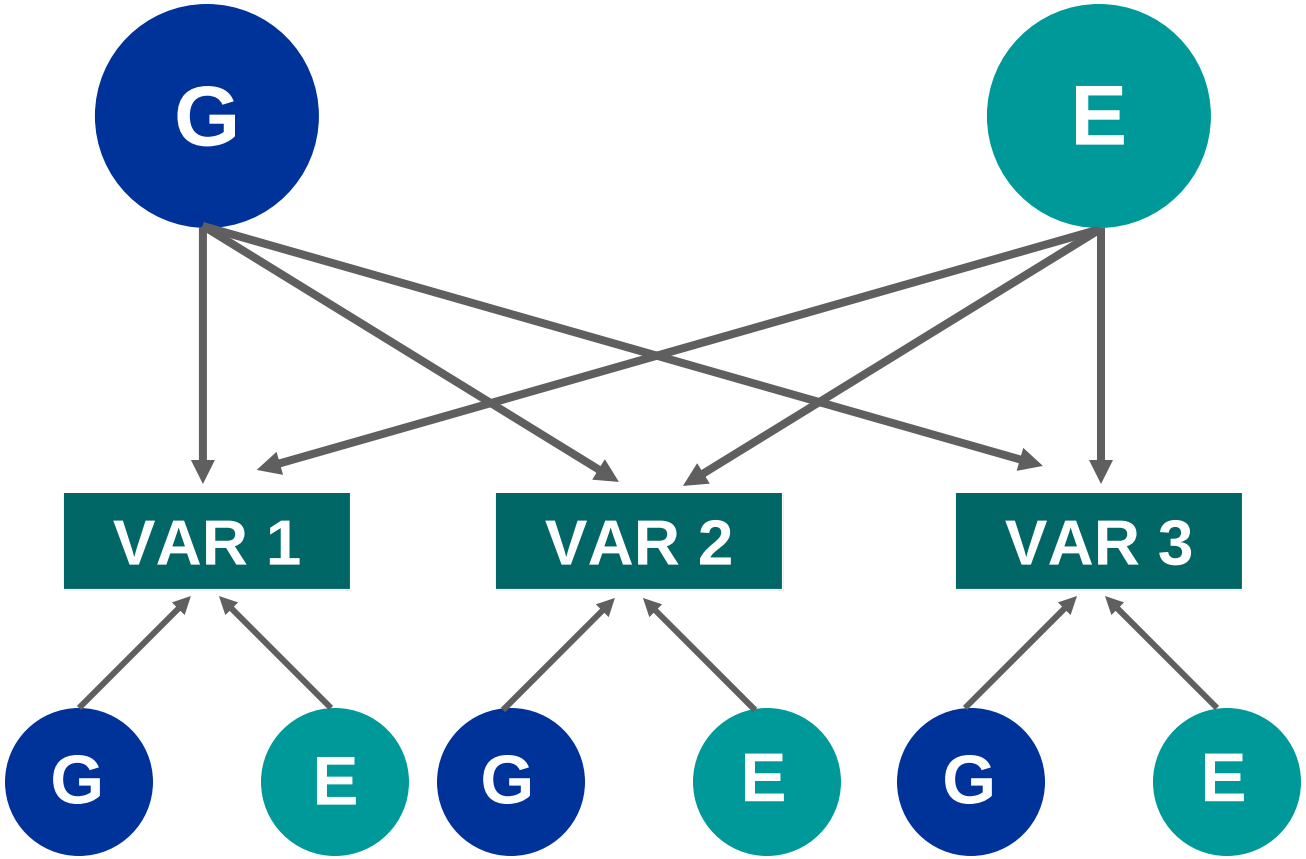


D.I. Boomsma et al. (1999) Twin Research 2, 115-125

IQ heritability (gene x age interaction)



Multivariate analysis: Genetic factor model: do the same latent factors influence multiple traits ?



Types of Twin Studies II

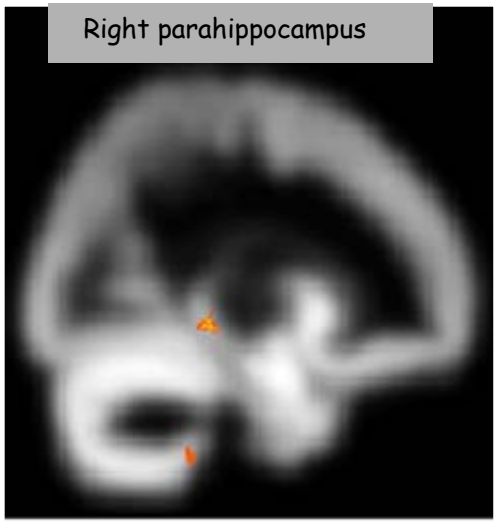
- Co-twin control study
- Extended twin study including:
  - parents: assortative mating
    cultural transmission
  - siblings: social interaction
  - MZ offspring: maternal effects

Monozygotic Twins Discordant for a trait: Identical genomes; differences caused by Environment?

- Different chromosome constitutions because of post-zygotic non-disjunction: e.g. MZ male-female 46,XY - 45,XO
- Differential *methylation* (imprinted genes)
- CNV (copy number variation)
- Skewed X chromosome inactivation in female MZ twins
- Differential trinucleotide repeat expansion
- Post-zygotic mutation
- *Prenatal* differences
- *Postnatal* environmental differences
- The interest is not MZs *per se,* but what discordance tells us about the causes of 'sporadic' disease

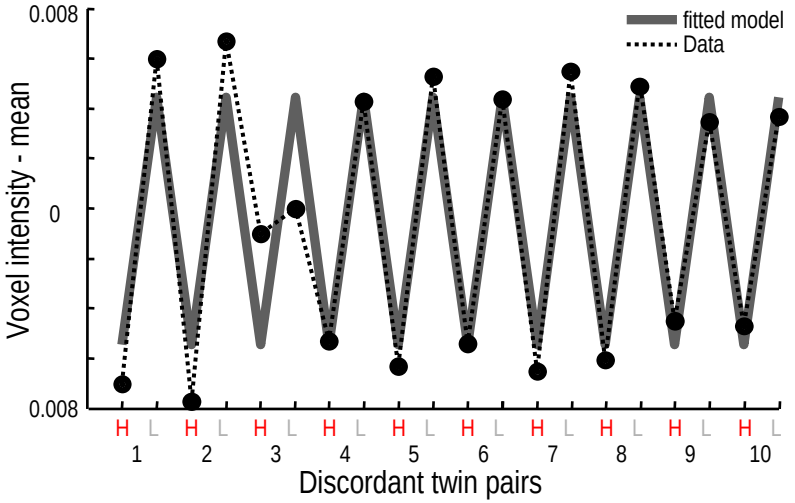
Martin N, Boomsma DI, Machin G. (1997) *Nature Genetics*

MZ twins discordant for depression risk: Gray Matter high risk twin < GM low risk twin



(maximum t =8.08, p < 0.0001 at x=24,y=-34, z=-6 in MNI space)

p<0.001, min 50 voxels



Right parahippocampus is smaller in the high risk twin from discordant MZ pairs (De Geus et al., 2007)

New trends

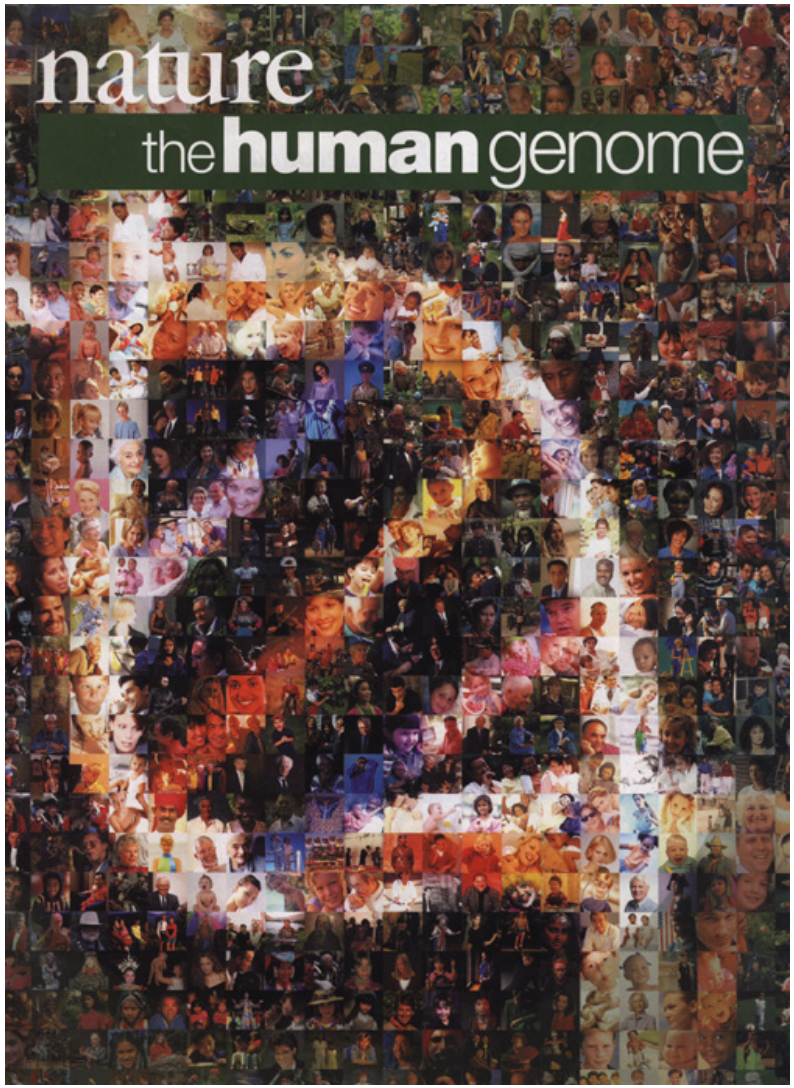
Human Genome Project: **Sequence** of the genome (base sequence)

Variation in the genome (e.g. microsatellites, SNPs, duplicons, copy number variation) related to variation in phenotype?

DNA methylation

Expression of the genome (RNA)

Metabolomics

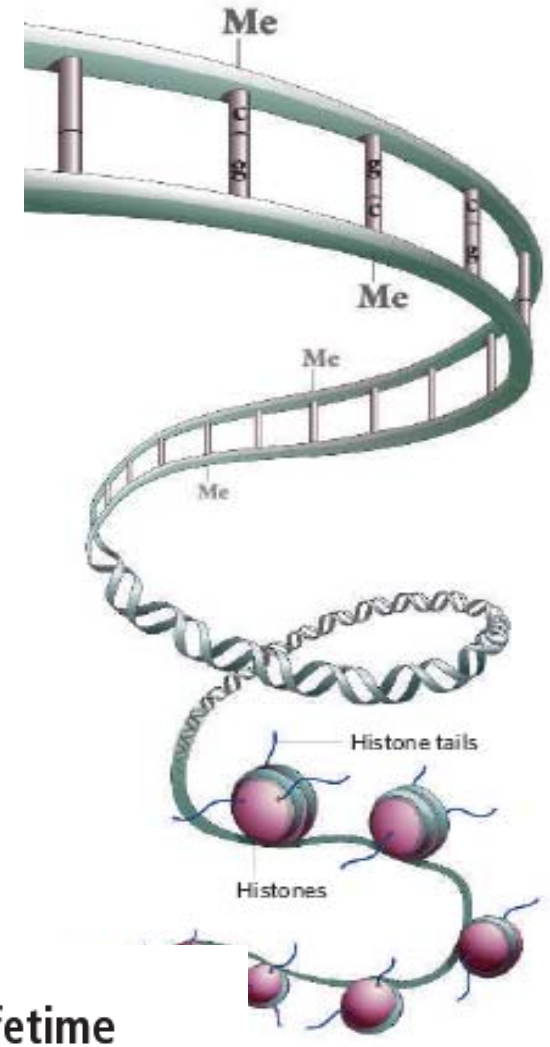


Co-twin control design

DISCORDANCE IN IDENTICAL TWINS

A role for Epigenetics?

Does epigenetics depend on age?



The two main components of the epigenetic code

DNA methylation
Methyl marks added to certain DNA bases repress gene activity.

Histone modification
A combination of different molecules can attach to the 'tails' of proteins called histones. These alter the activity of the DNA wrapped around them.

PNAS

Epigenetic differences arise during the lifetime of monozygotic twins

Mario F. Fraga*, Esteban Ballestar*, Maria F. Paz*, Santiago Ropero*, Fernando Setien*, Maria L. Ballestar†, Damia Heine-Suñer‡, Juan C. Cigudosa§, Miguel Urioste¶, Javier Benitez¶, Manuel Boix-Chornet†, Abel Sanchez-Aguilera†, Charlotte Lingl, Emma Carlssonl, Pernille Poulsen**, Allan Vaag**, Zarko Stephan††, Tim D. Spector††, Yue-Zhong Wu‡‡, Christoph Plass‡‡, and Manel Esteller*§§

*Epigenetics, §Cytogenetics, and ¶Genetic Laboratories, Spanish National Cancer Centre (CNIO), Melchor Fernandez Almagro 3, 28029 Madrid, Spain; †Department of Behavioral Science, University of Valencia, 46010 Valencia, Spain; ‡Molecular Genetics Laboratory, Genetics Department, Son Dureta Hospital, 07014 Palma de Mallorca, Spain; lDepartment of Clinical Sciences, University Hospital Malmö, Lund University, S-205 02 Malmö, Sweden; **Steno

Please cite this article in press as: Bruder et al., Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles, The American Journal of Human Genetics (2008), doi:10.1016/j.ajhg.2007.12.011

REPORT

Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles

Carl E.G. Bruder,[1,*] Arkadiusz Piotrowski,[1] Antoinet A.C.J. Gijsbers,[2,3] Robin Andersson,[4] Stephen Erickson,[5] Teresita Diaz de Ståhl,[6] Uwe Menzel,[6] Johanna Sandgren,[7] Desiree von Tell,[1] Andrzej Poplawski,[1] Michael Crowley,[1] Chiquito Crasto,[1] E. Christopher Partridge,[1] Hemant Tiwari,[5] David B. Allison,[1,5] Jan Komorowski,[4] Gert-Jan B. van Ommen,[2,3] Dorret I. Boomsma,[8] Nancy L. Pedersen,[9] Johan T. den Dunnen,[2,3] Karin Wirdefeldt,[9] and Jan P. Dumanski[1,6]

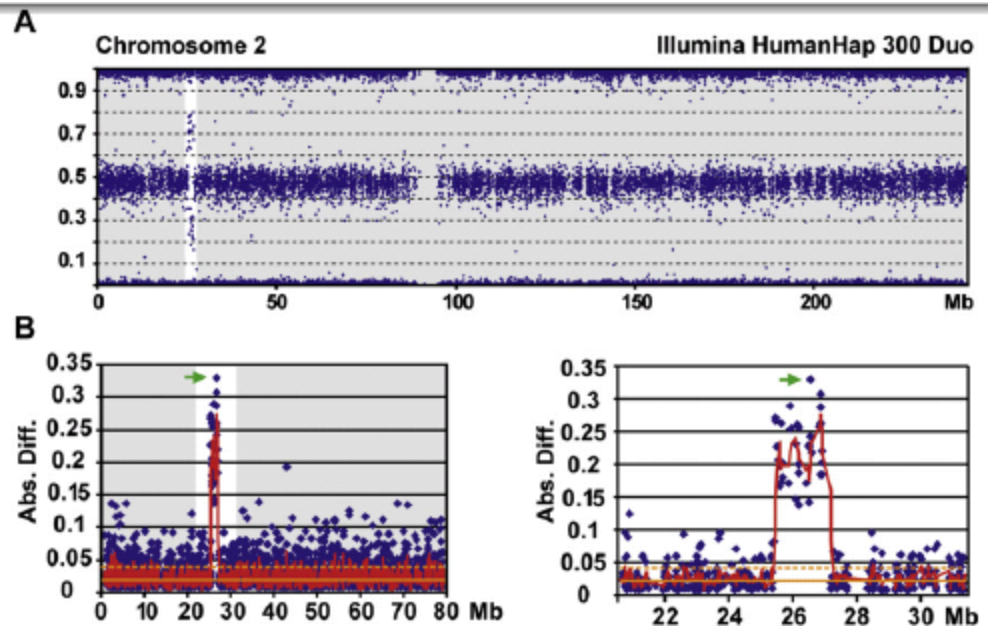


Figure 3. CNV Analysis of Twin D8 Showing the 1.6 Mb Deletion on Chromosome 2

(A) Profile of the entire chromosome 2 from Illumina HumanHap 300 Duo beadchip showing the values of SNP allele ratios. True heterozygous SNPs are expected to be distributed around a value of 0.5. In the highlighted region (white box), the allele ratios differ significantly from 0.5, indicating an imbalance in the allele signals caused by a 1.6 Mb deletion.

(B) Two enlarged views of the deleted region, plotted as values of absolute difference between the heterozygous SNP allele frequencies in twin D8 versus twin D7, calculated in a similar way as shown in Figures 1C and 1E. The red line in both graphs displays the moving average, with a period of

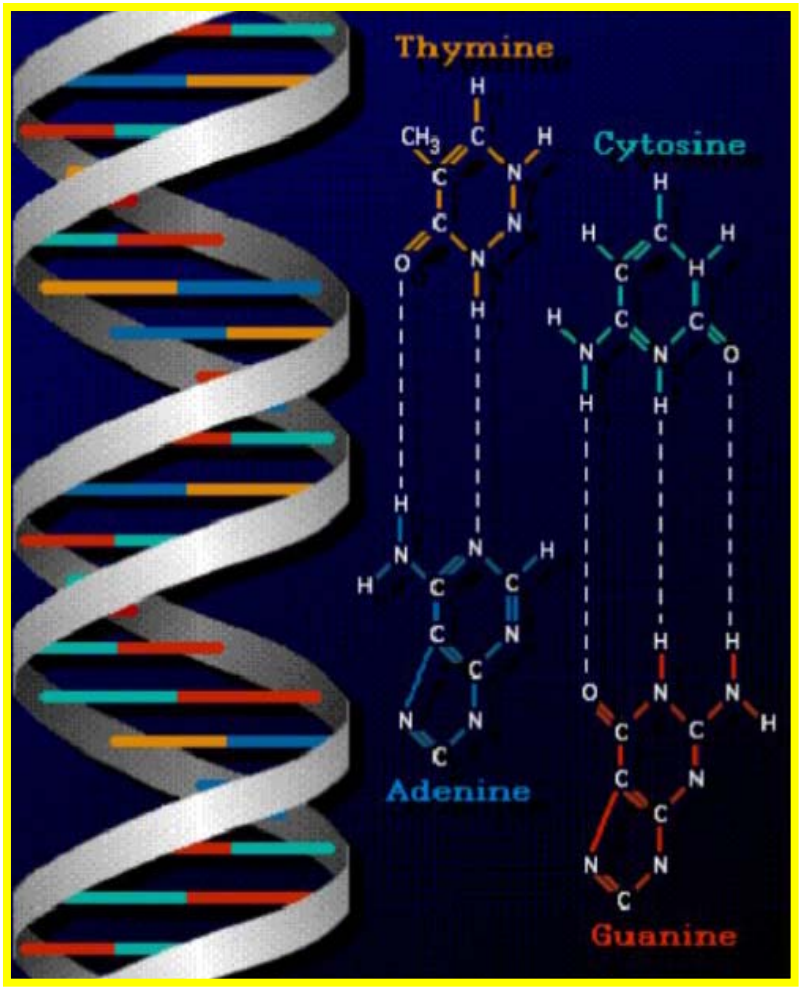
Unselected NTR twins (10 MZ pairs)

- CNV: gains and losses of large chunks of DNA sequence consisting of between ten thousand and five million letters (known as Copy Number Variation).

- Based on shared CNVs patterns twin pairs were easily recognized.
- However, we also detected an unexpected number of unique differences within the monozygotic twin pairs.
- The number of CNVs identified depends mainly on the settings of the scoring algorithms; in the size range of 0.3-1.2 Mb we detect 1-2 per pair.
- CNVs are not present in 100% of the cells. This suggests somatic mosaicism, i.e. a post-meiotic emergence.

Genetic differences
= differences in DNA sequence

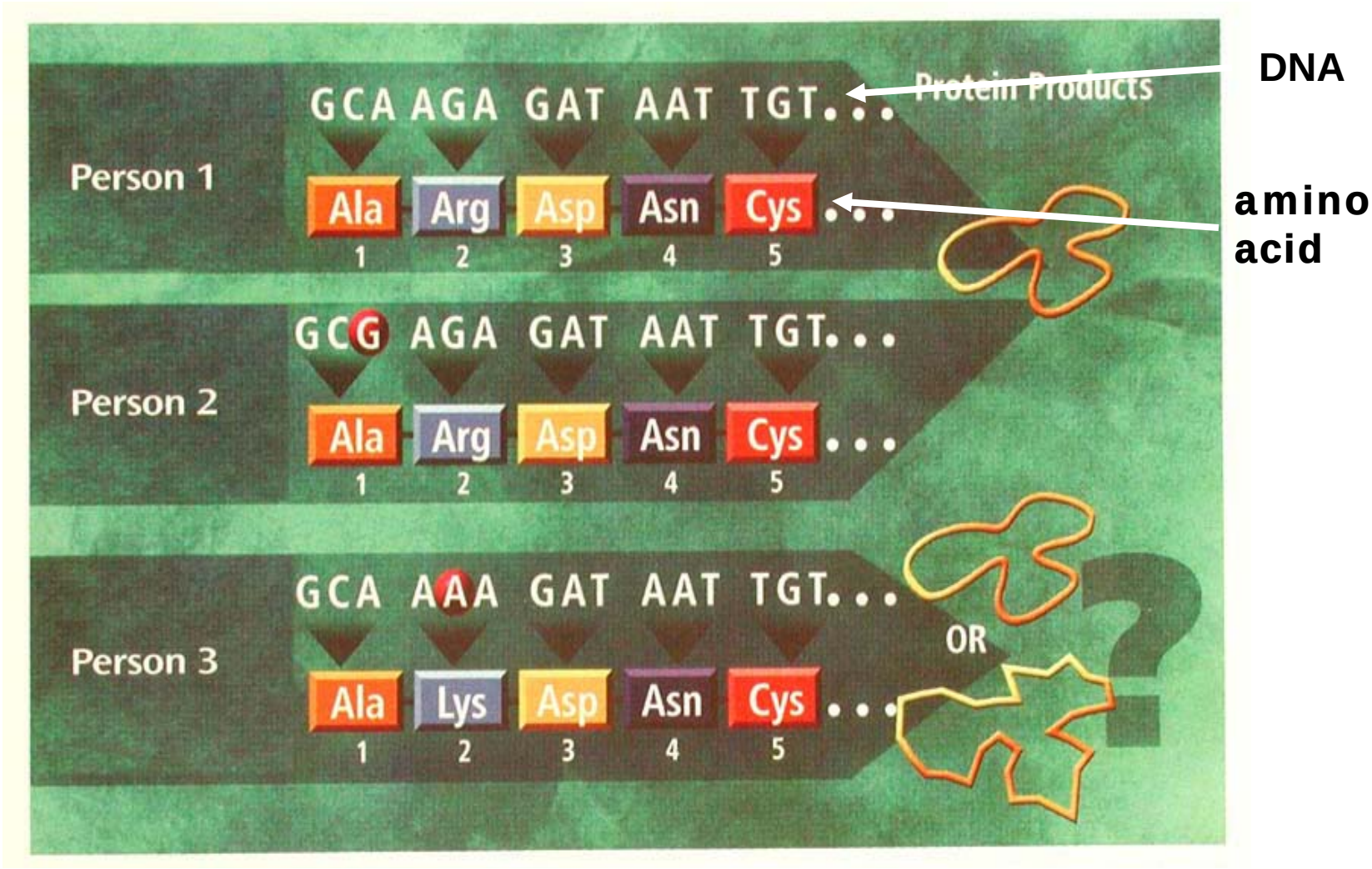
Human-Human 1:1000 = 0.1%



Human-Chimp 1:100 = 1%

Human-Mouse 1:8 = 15%

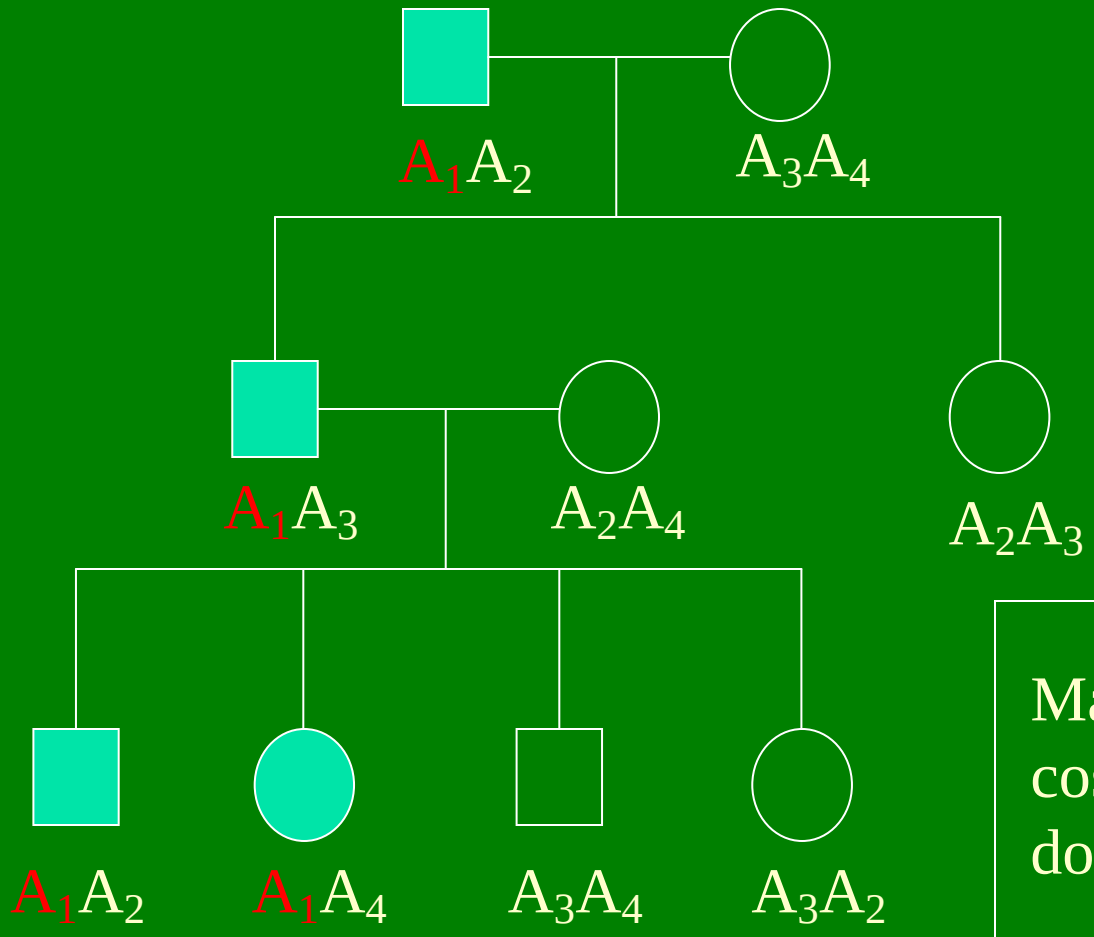
Sequence differences between individuals



3 Stages of Genetic Mapping

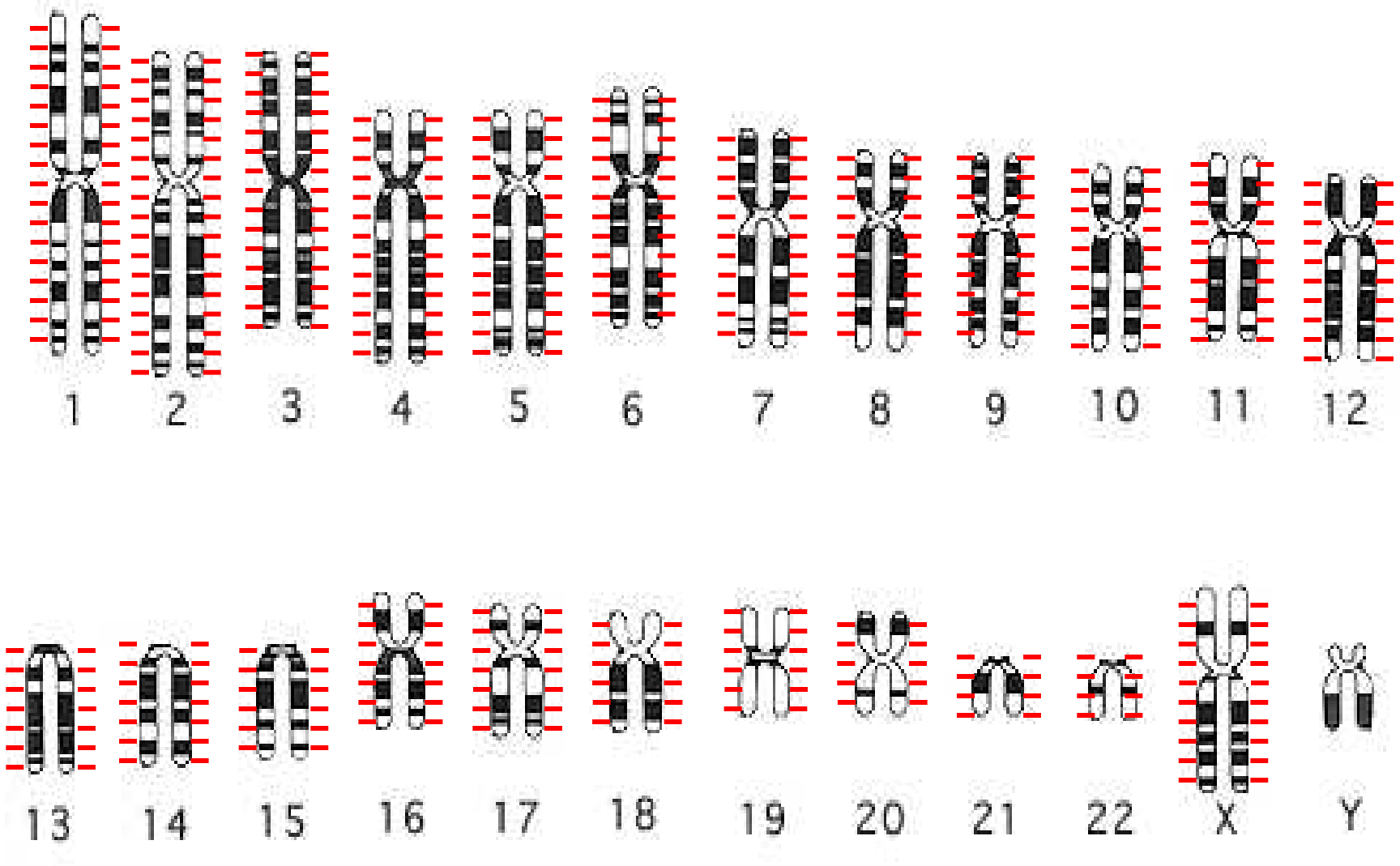
- Are there genes influencing this trait?
  - Genetic epidemiological studies
- Where are those genes?
  - **Linkage analysis**
  - (look for quantitative trait loci: QTL)
- What are those genes?
  - Association analysis

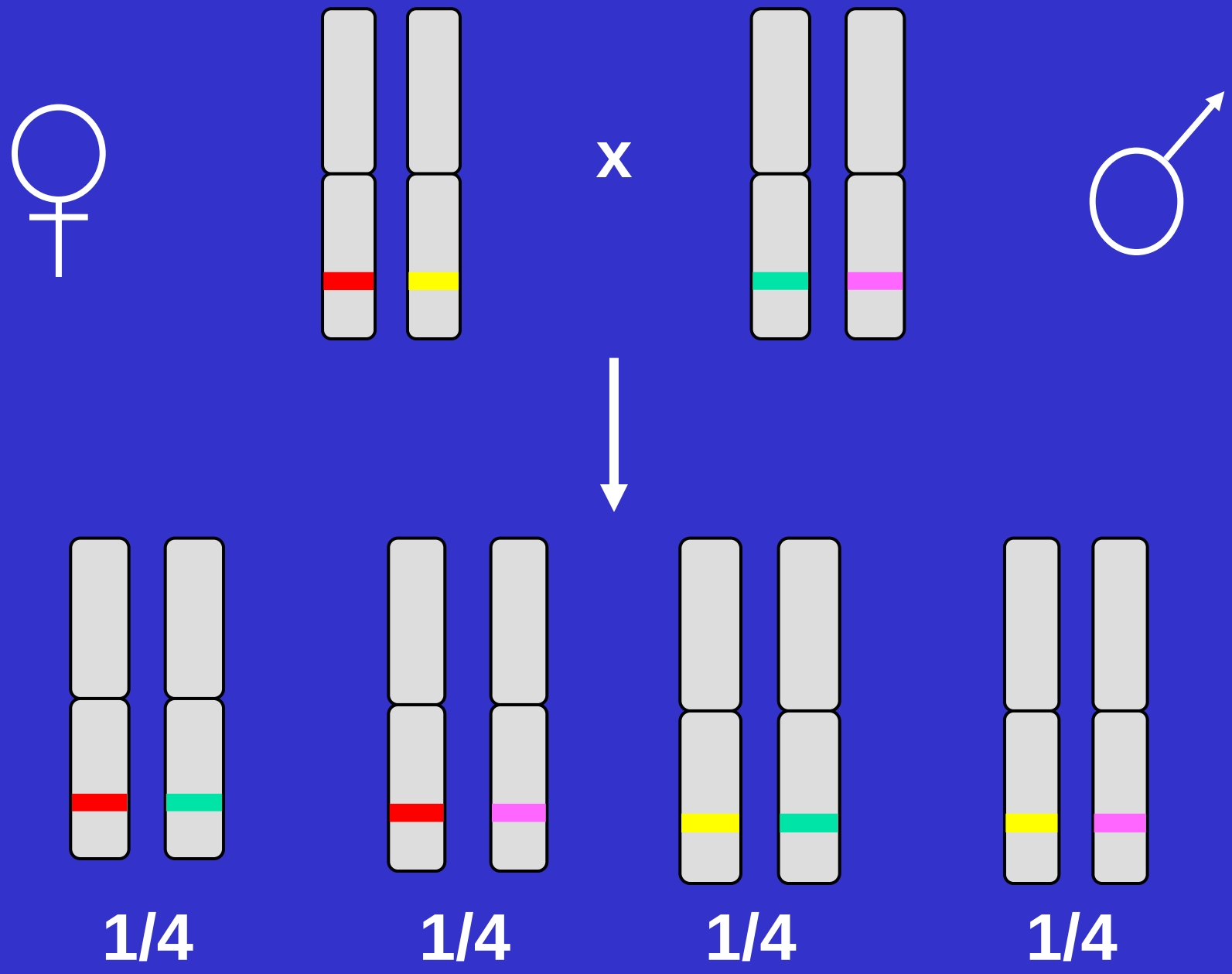
Linkage = Co-segregation



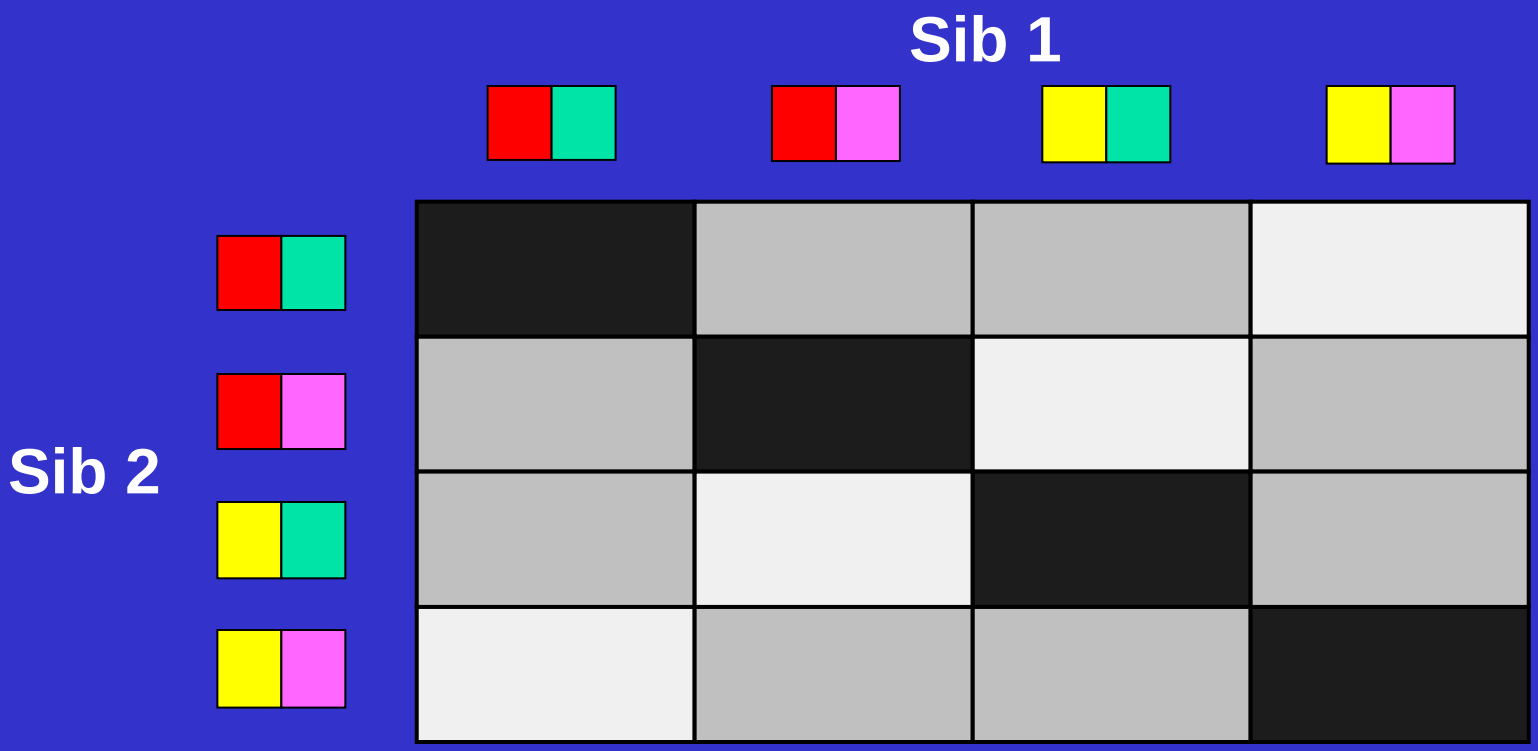
Marker allele A_1 cosegregates with dominant disease

Linkage Markers...




x
1/4
1/4
1/4
1/4

IDENTITY BY DESCENT

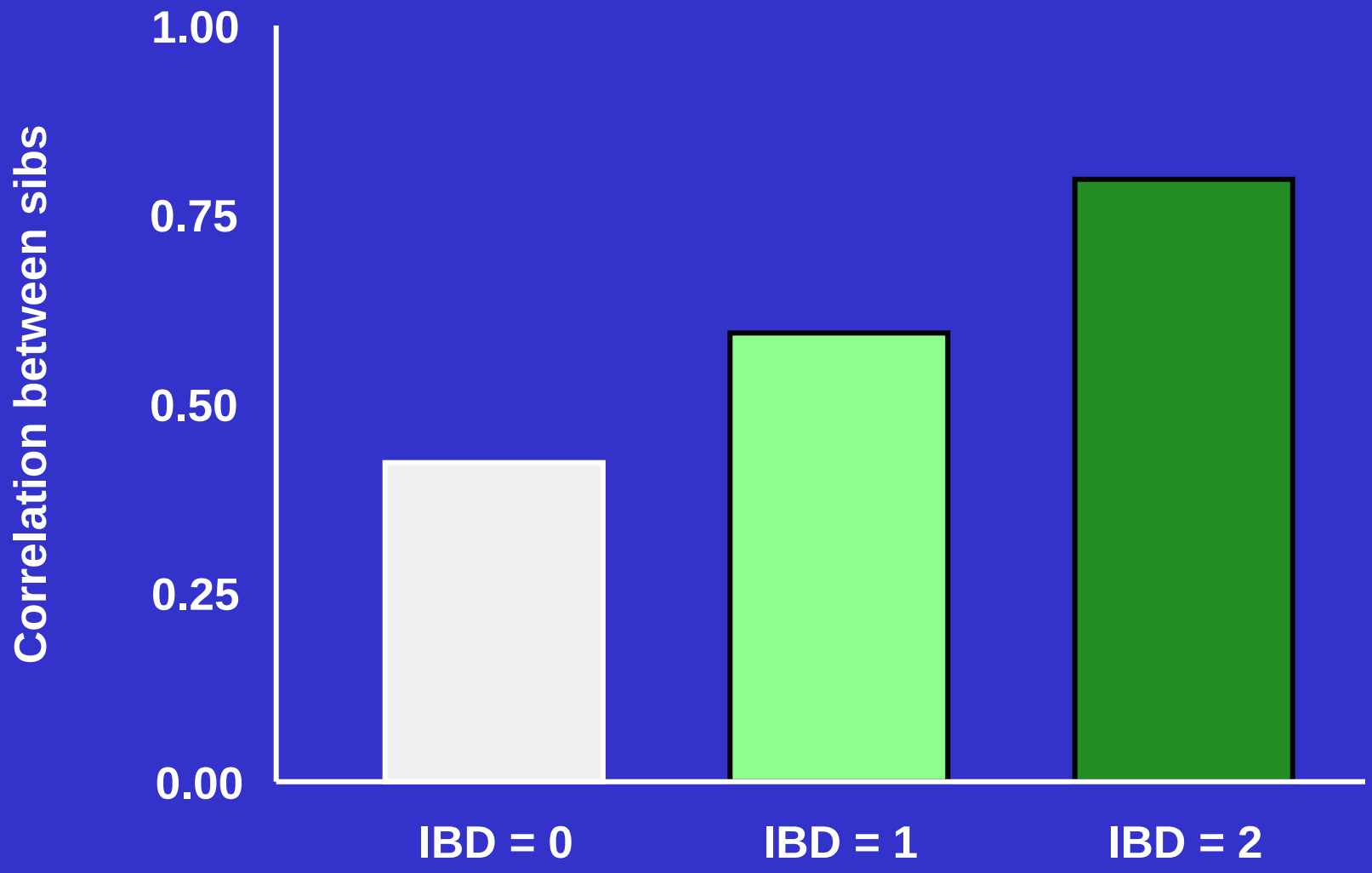


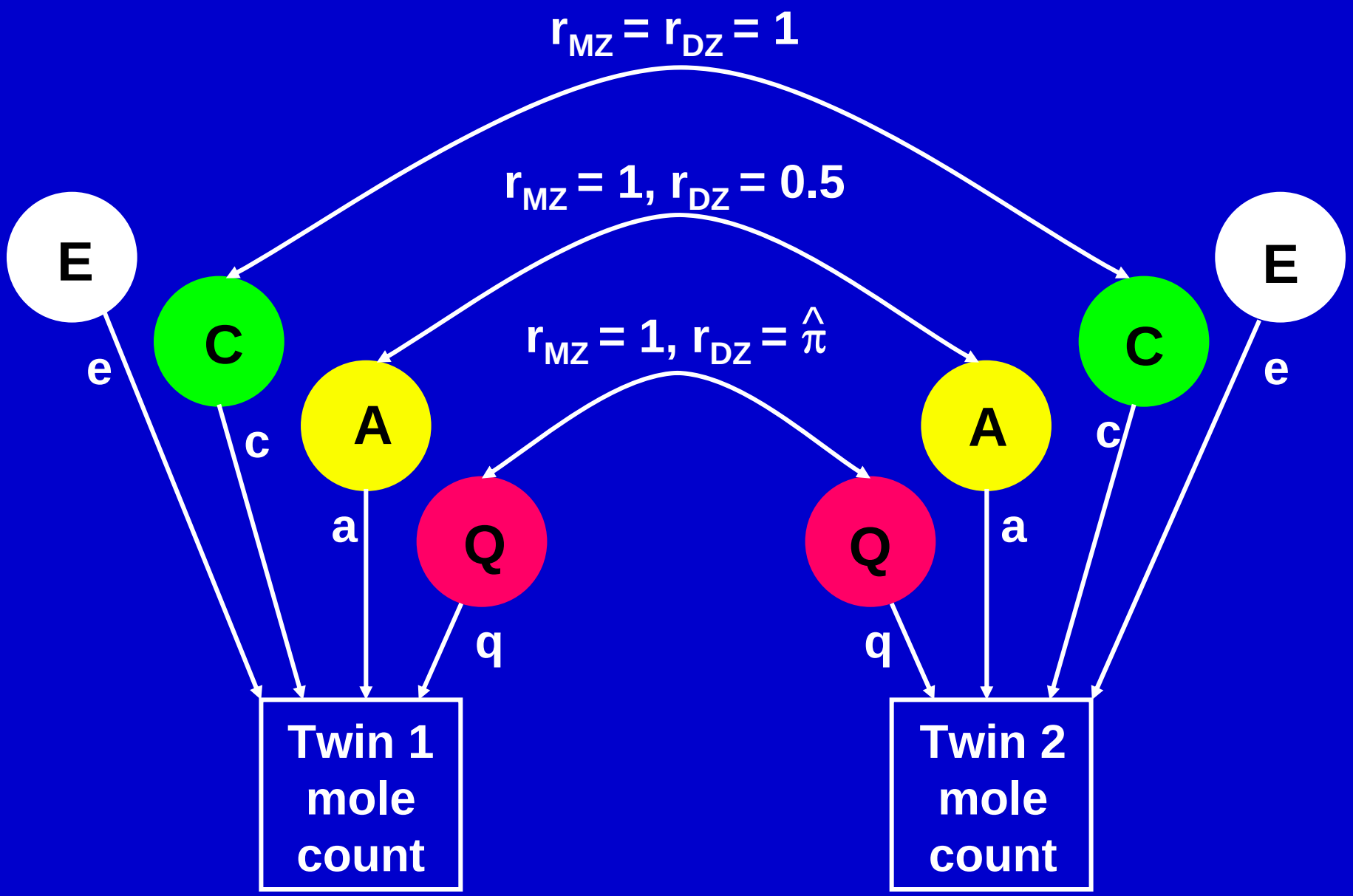
4/16 = 1/4 sibs share BOTH parental alleles IBD = 2

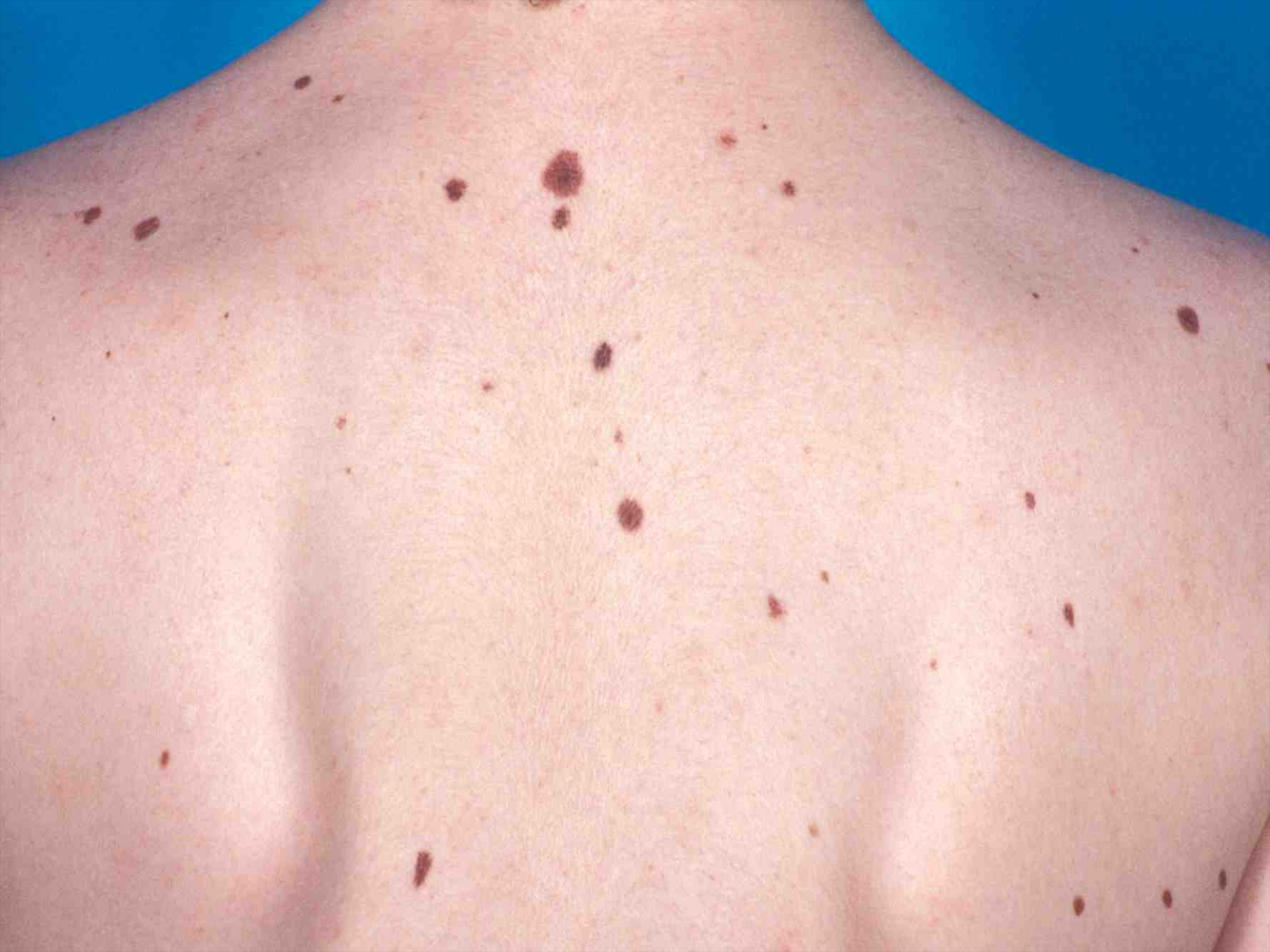
8/16 = 1/2 sibs share ONE parental allele IBD = 1

4/16 = 1/4 sibs share NO parental alleles IBD = 0

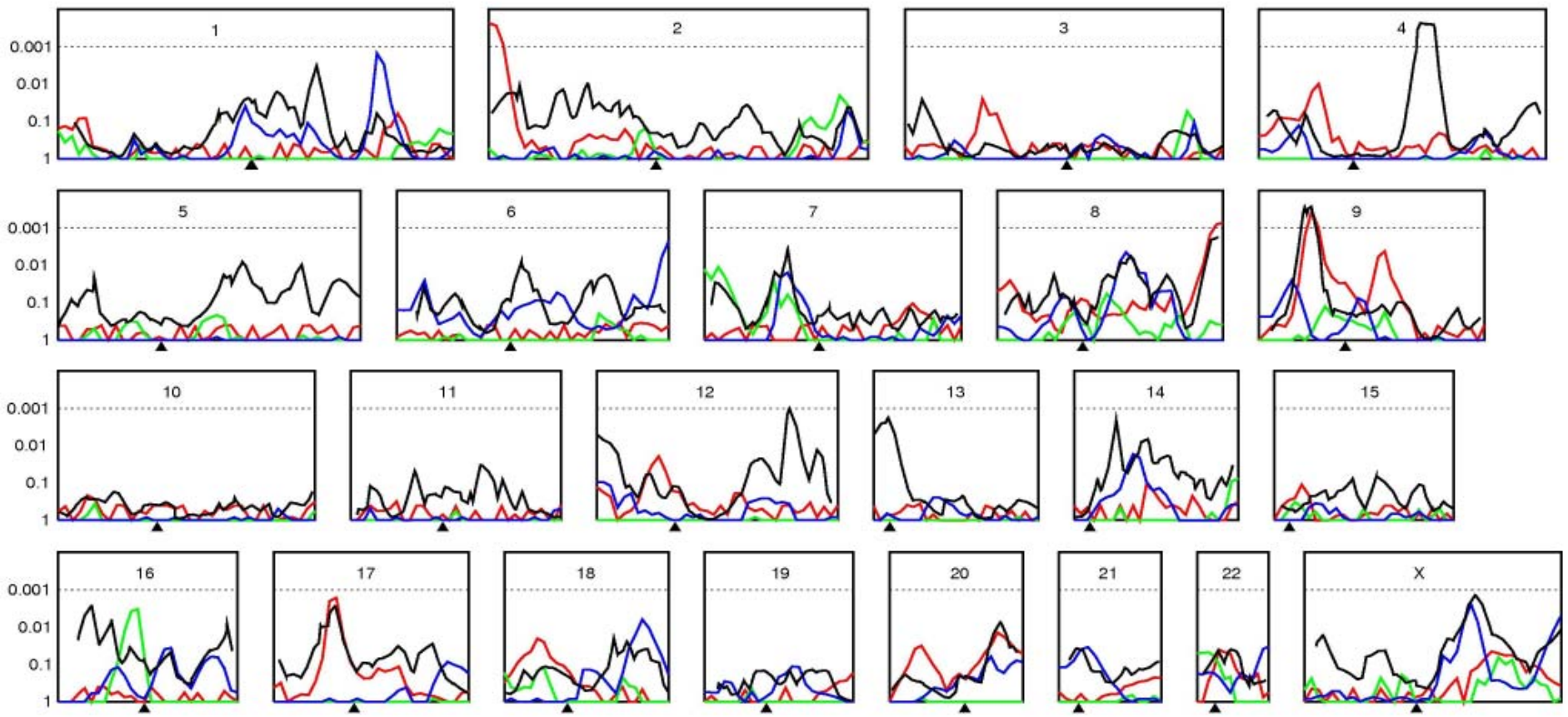
For continuous measures
Unselected sib pairs




$r_{MZ} = r_{DZ} = 1$
$r_{MZ} = 1, r_{DZ} = 0.5$
$r_{MZ} = 1, r_{DZ} = \hat{\pi}$
E
C
A
Q
E
C
A
Q
e
c
a
q
q
a
c
e
Twin 1 mole count
Twin 2 mole count



Linkage for mole counts in Australian twin families



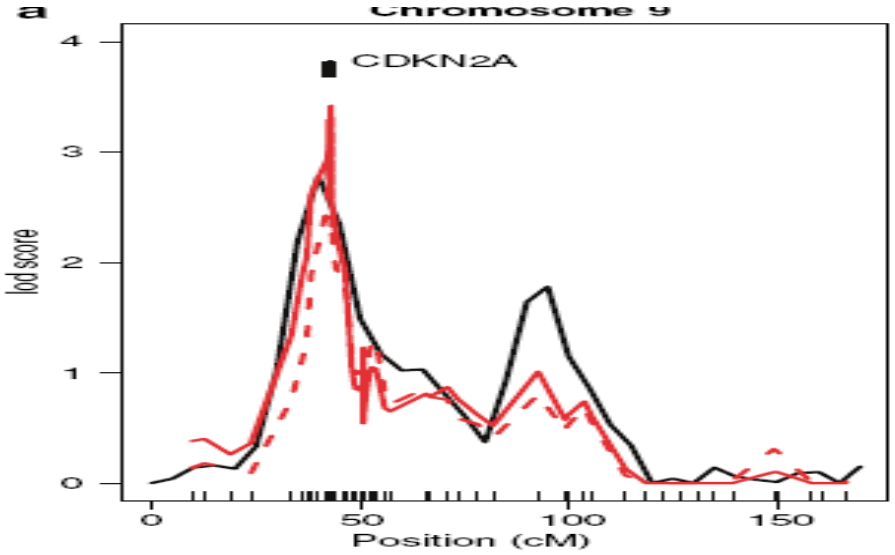
A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions

Gu Zhu[1], Grant W Montgomery[1], Michael R James[1], Jeff M Trent[2], Nicholas K Hayward[1], Nicholas G Martin[1] and David L Duffy*[,1]

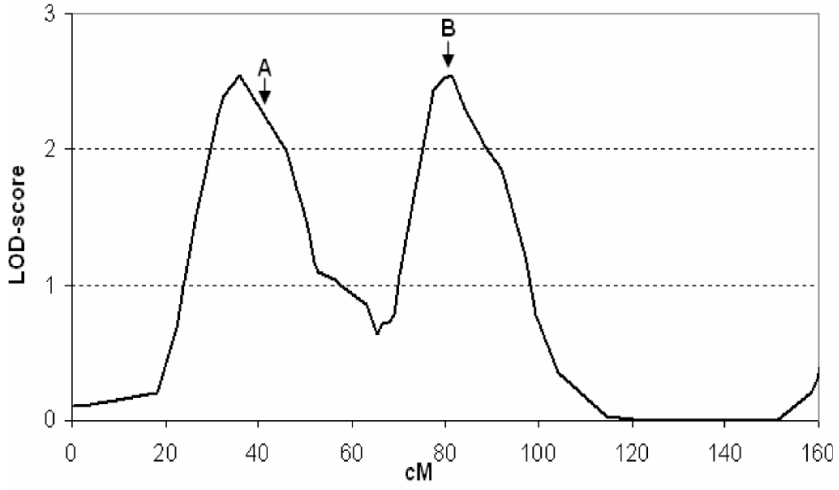
European Journal of Human Genetics (2007) 15, 94–102

Flat mole count: chromosome 9 linkage in Australian and UK twins

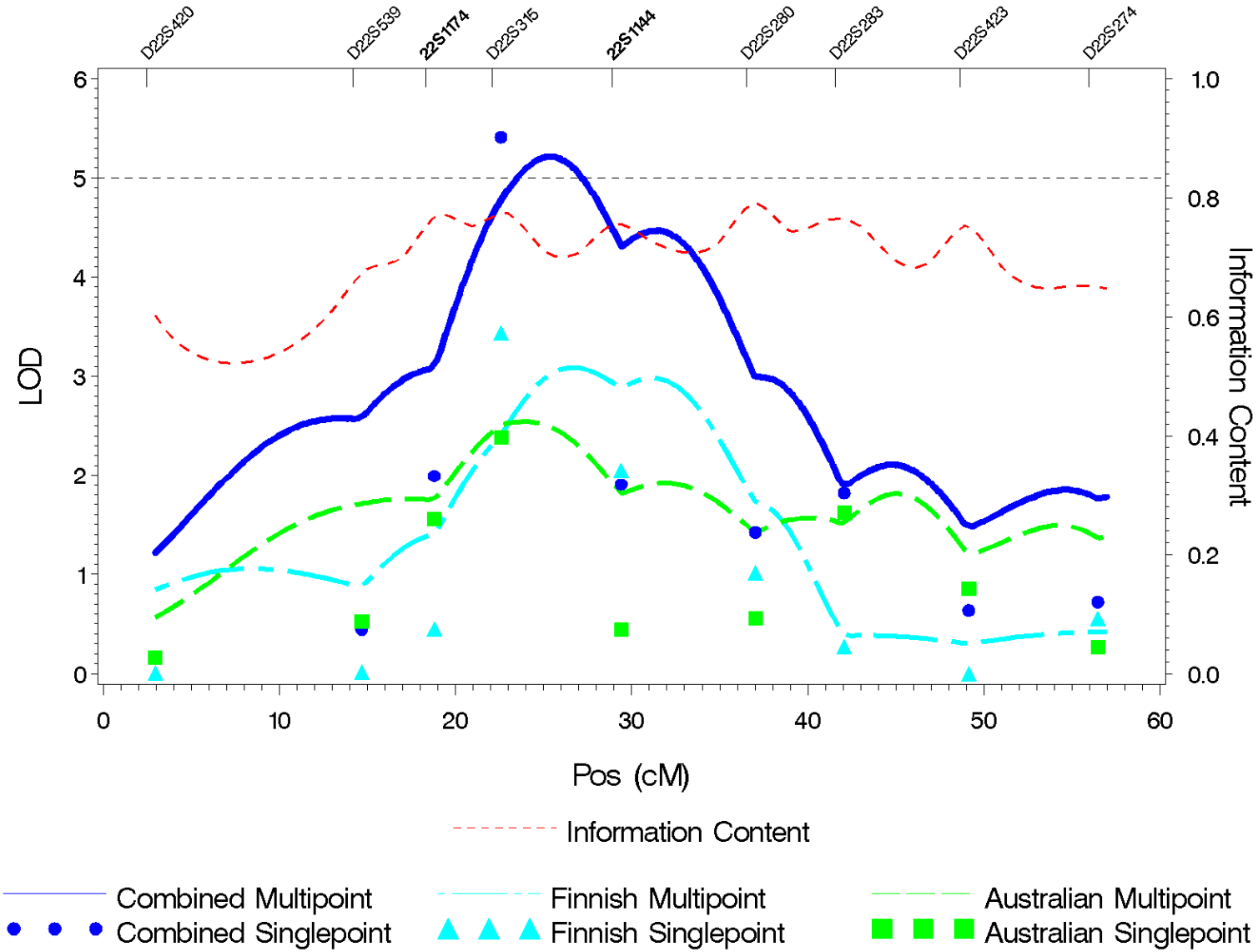
Australia



UK



Linkage for MaxCigs24 in Australia and Finland



AJHG, in press

Linkage Analysis

- Models the **covariance structure** among family members

- Marker sharing between relatives
  - Identifies large regions
    - Include several candidates
- Complex disease
  - Scans on sets of small families popular
  - No strong assumptions about disease alleles
  - Low power
  - Limited resolution

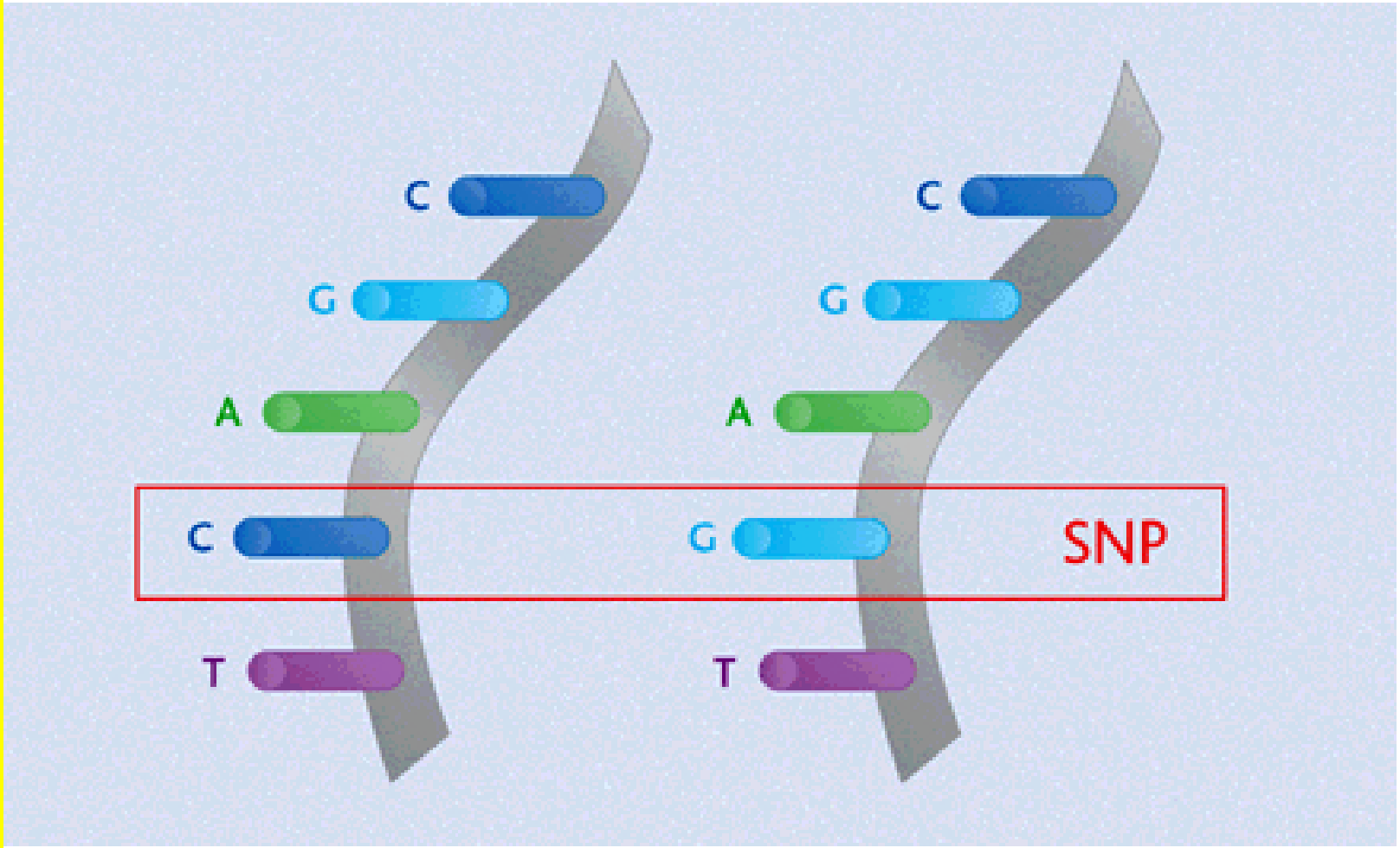
Association

- Models **“mean”** values
- Looks for correlation between *specific alleles* and a phenotype (quantitative trait value, disease risk)
- E.g. cases and controls (affected / unaffected)
- Or high and low scoring Ss

Association

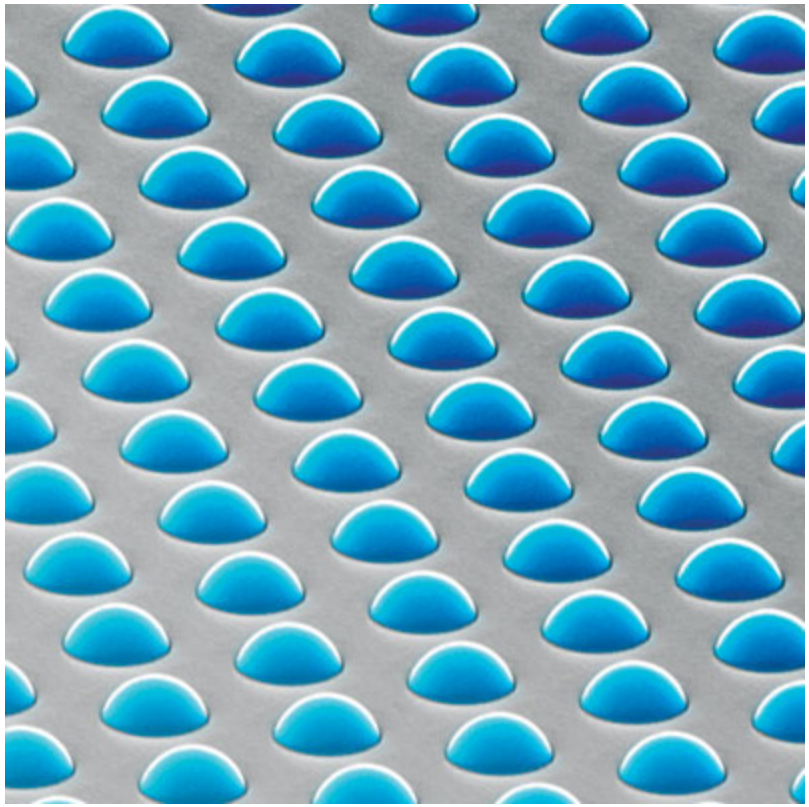
- More sensitive to small effects
- Need to “guess” gene/alleles (“candidate gene”) or be close enough for linkage disequilibrium with nearby loci (GWA: Genome Wide Association)
- May get spurious association (“stratification”) – need to have genetic controls to be convinced
- May get too many “positive” results (if the number of tests is large)

Variation: Single Nucleotide Polymorphisms

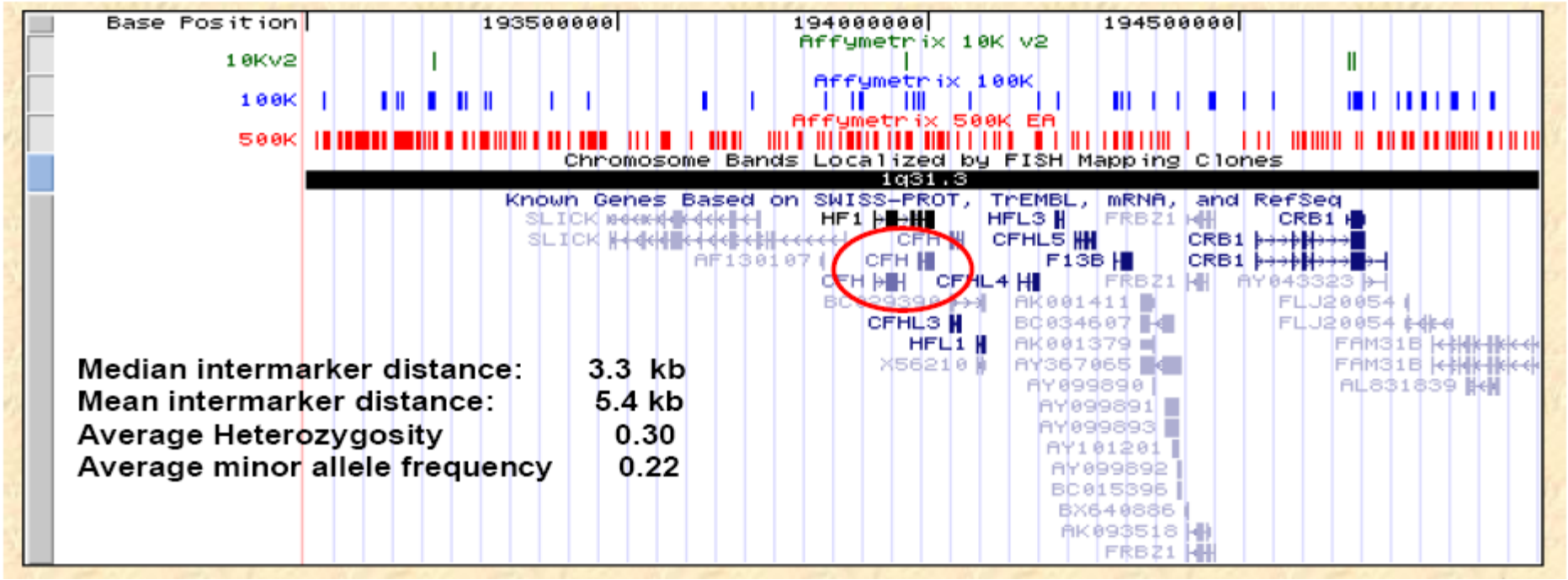


Complex disease marker? SNPs are single-base differences in DNA

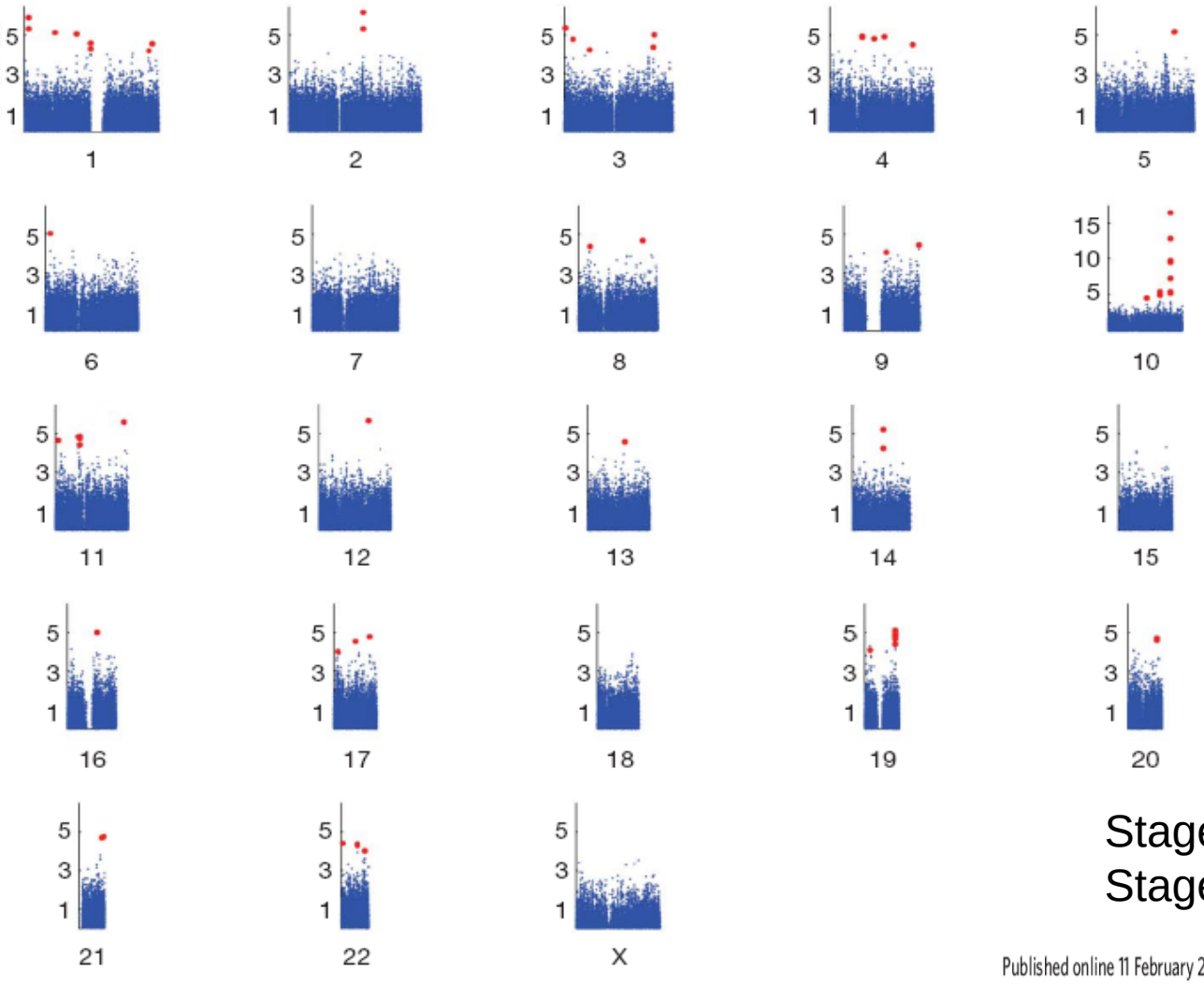
High density SNP arrays – up to 1 million SNPs



Comparison of Affymetrix 10k, 100k, 500k SNP chips



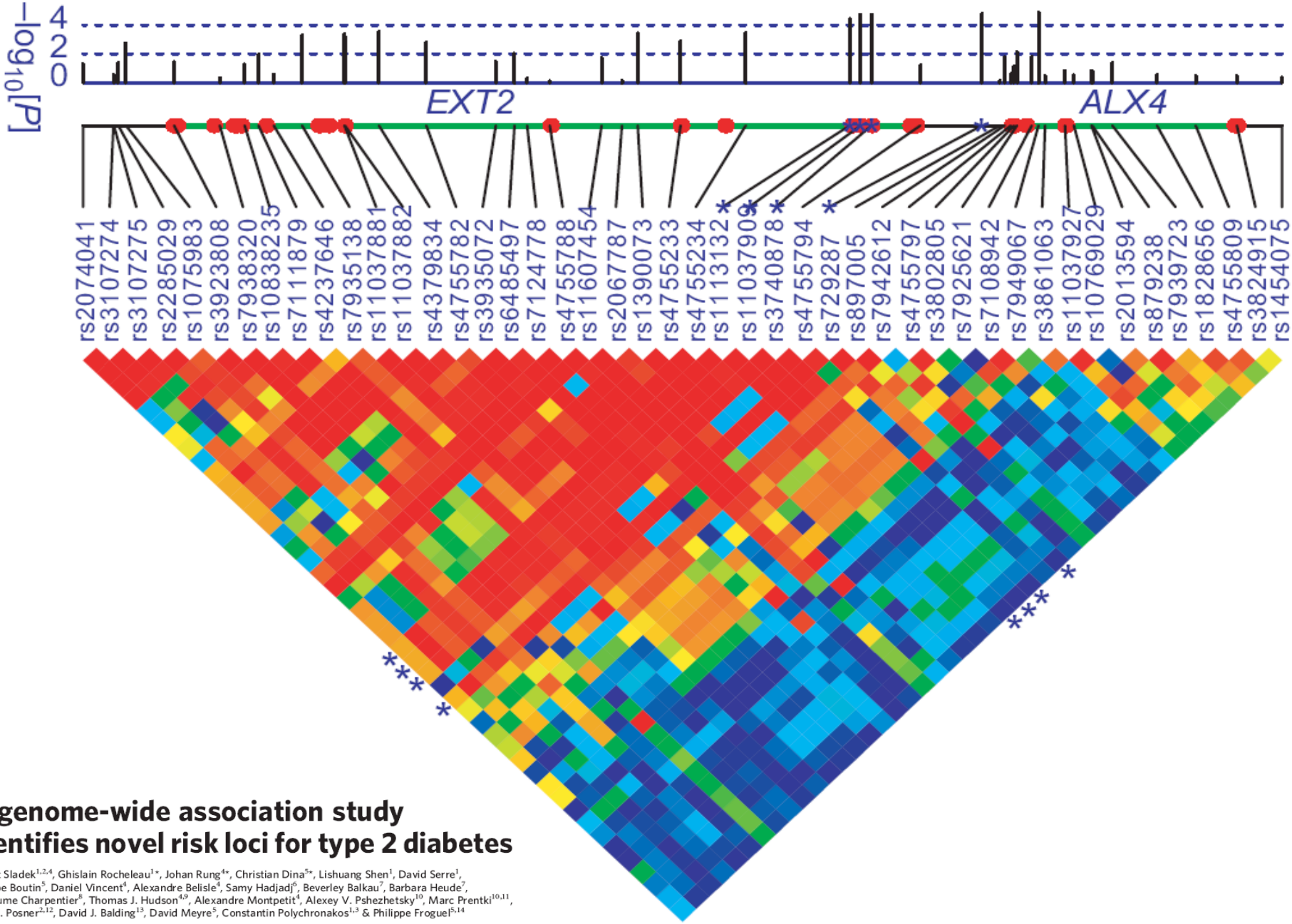
A genome-wide association study identifies novel risk loci for type 2 diabetes



Stage 1: Illumina 100k+300k
Stage 2: Sequenom Iplex

Published online 11 February 2007

nature

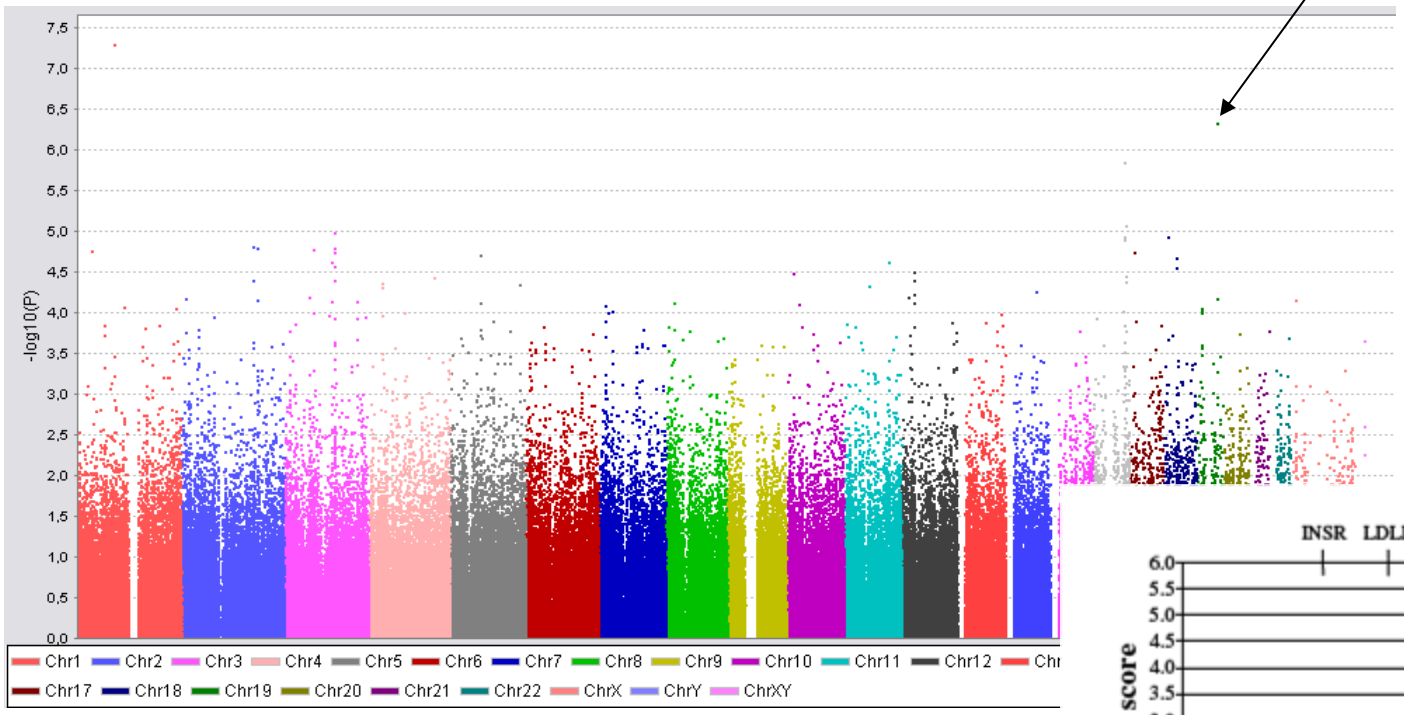


A genome-wide association study identifies novel risk loci for type 2 diabetes

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]

LDL cholesterol (levels)

19!!



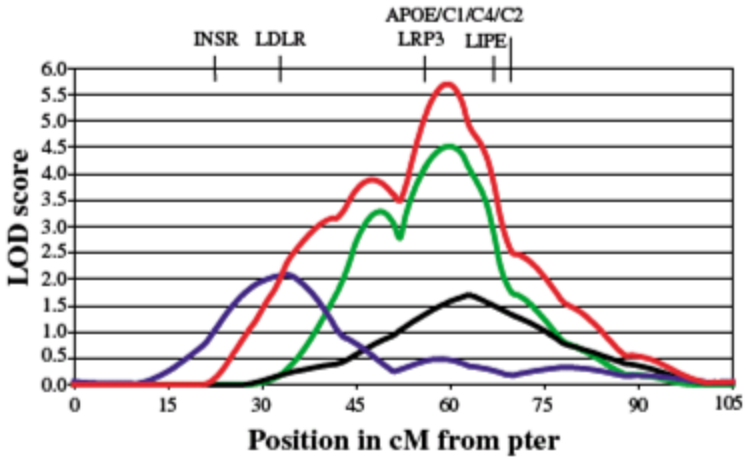


Figure 1 Linkage of LDL cholesterol levels with chromosome 19 in adult Dutch (green line), Swedish (black line) and Australian (blue line) twins in separate analyses and a combined analysis (red line).

STATISTICS AND MEDICINE

Drinking from the Fire Hose — Statistical Issues in Genomewide Association Studies

David J. Hunter, M.B., B.S., and Peter Kraft, Ph.D.

Related article, page 443

"There have been few, if any, similar bursts of discovery in the history of medical research…"

The past 3 months have seen the publication of a series of

ating the need for guessing which genes are likely to harbor variants

The main problem with this strategy is that because of the

ost studained in samples ower to

and in this issue of the *Journal*, coronary artery disease (reported by Samani et al., pages 443–453). These genomewide association studies have been able to examine interpatient differences in inherited genetic variability at an unprecedented level of resolution, thanks to the development of microarrays, or chips, capable of as-

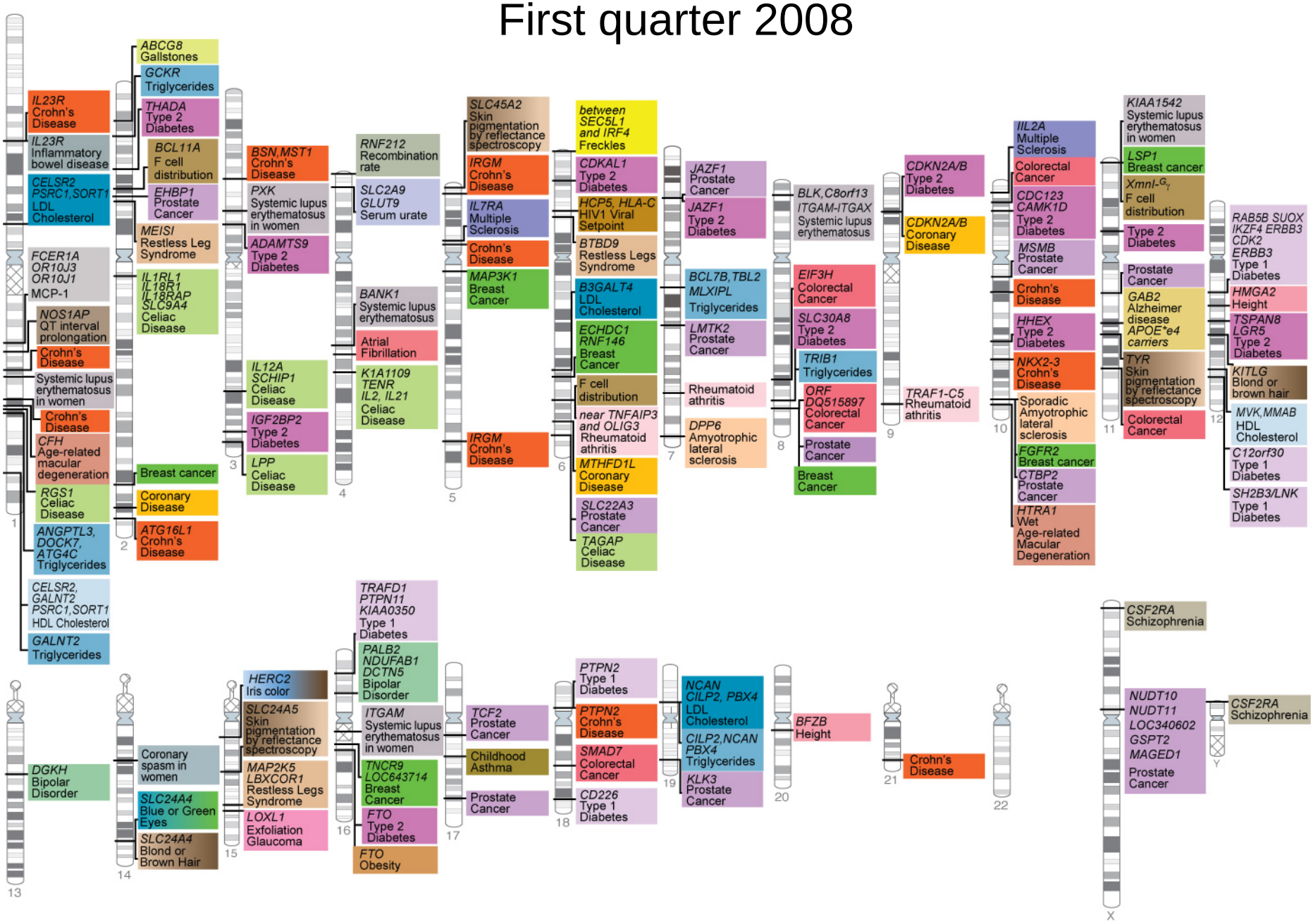
lated to the disease. Some of these associations have been found in regions not even known to harbor genes, such as the 8q24 region, in which multiple variants have been found to be associated with prostate cancer.[2] Such findings promise to open up new avenues of research, through both the discovery of new genes rele-

generate P values as small as 10^{-7}. In addition, most variants identified recently have been associated with modest relative risks (e.g., 1.3 for heterozygotes and 1.6 for homozygotes), and many true associations are not likely to exceed P values as extreme as 10^{-7} in an initial study. On the other hand, a "statistically significant" finding

Hunter DJ and Kraft P, *N Engl J Med* 2007; 357:436-439.

Stephen Channock

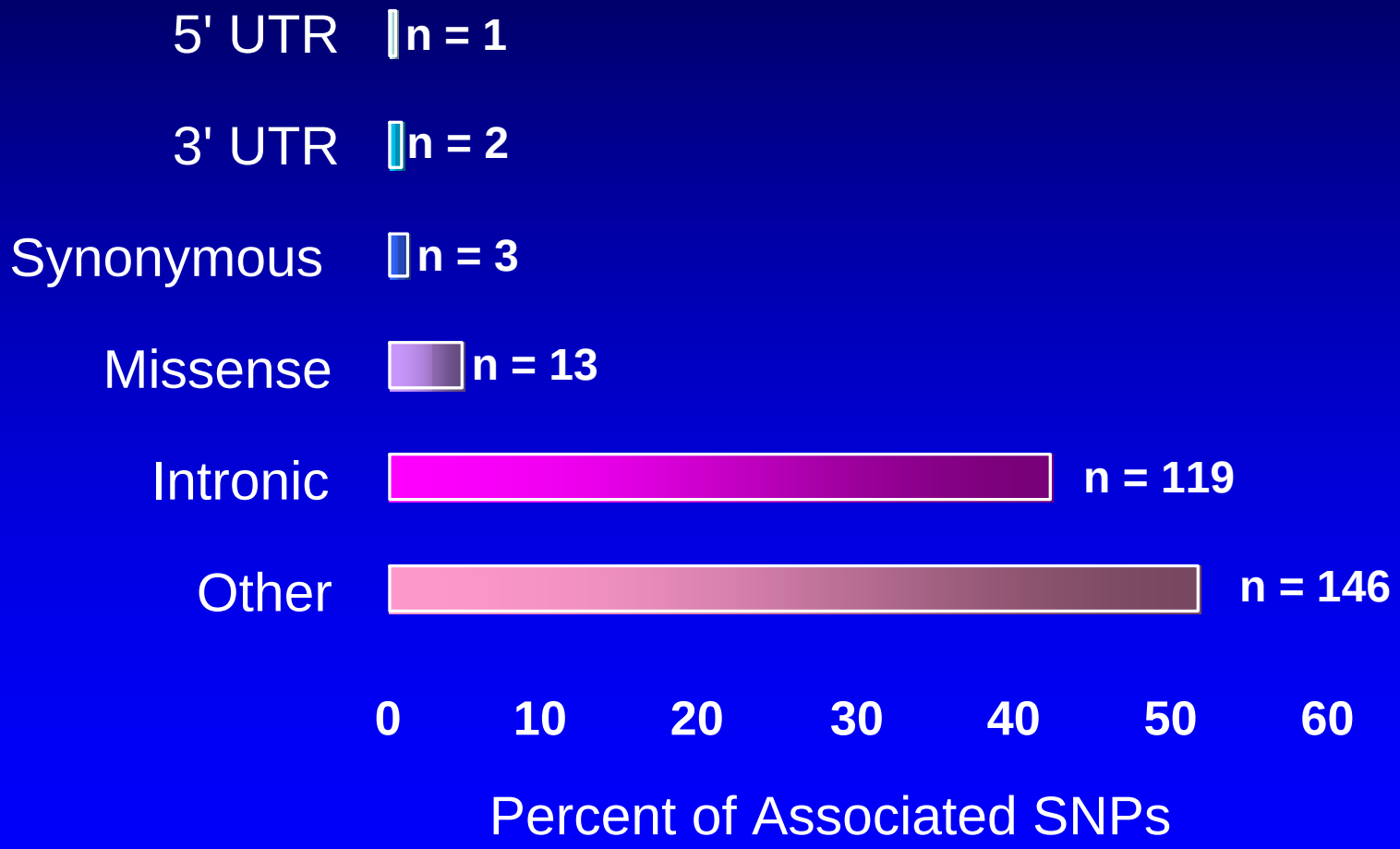
First quarter 2008



Manolio, Brooks, Collins, J. Clin. Invest., May 2008

Stephen Channock

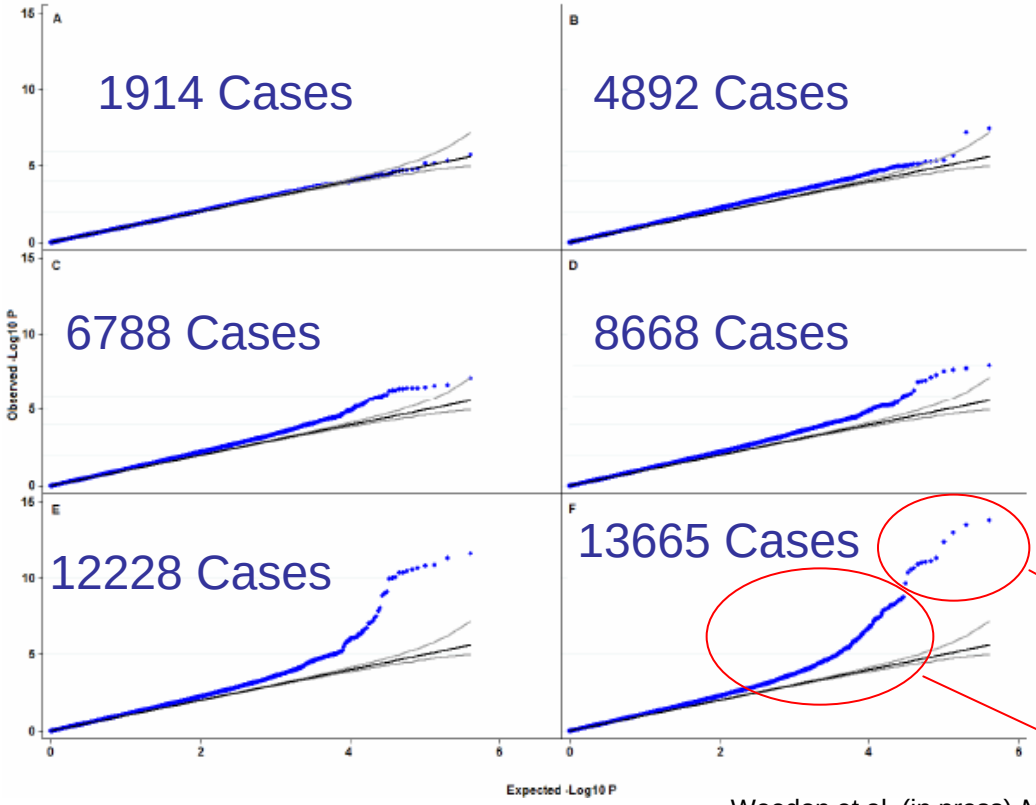
Functional Classification of 284 SNPs Associated with Complex Traits



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

Stephen Channock

GWA of Height



Weedon et al. (in press) *Nat Genet*

Significant results

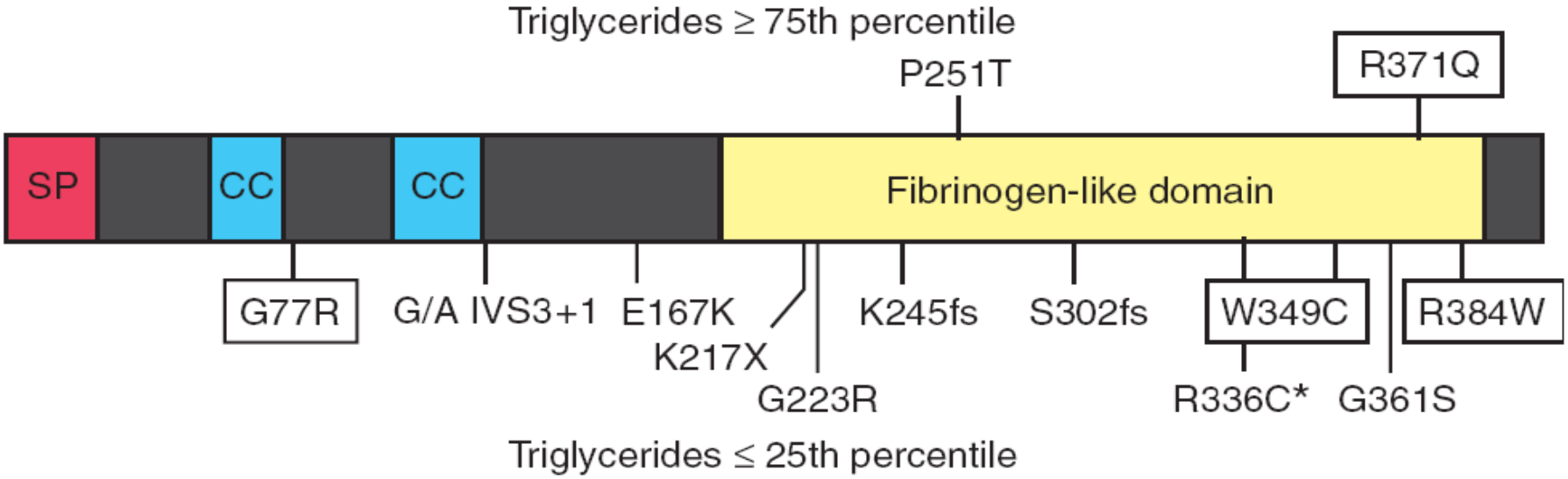
Other loci?

- Large numbers are needed to detect QTLs !!!
- Collaboration is the name of the game !!!

Even for “simple” diseases the number of alleles is large

- Ischaemic heart disease (LDR) >190
- Breast cancer (BRAC1) >300
- Colorectal cancer (MLN1) >140

Large numbers of rare variants affect quantitative traits



Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL

Stefano Romeo[1,7], Len A Pennacchio[2,7], Yunxin Fu[3], Eric Boerwinkle[3], Anne Tybjaerg-Hansen[4], Helen H Hobbs[1,6] & Jonathan C Cohen[1,5]

NATURE GENETICS VOLUME 39 | NUMBER 4 | APRIL 2007

The next stage – large scale resequencing to detect new/rare variants

illumina®
making sense out of life

LOG IN TO ICOM | ORDER | SEARCH

ABOUT ILLUMINA | PRODUCTS & SERVICES | SUPPORT | COMMUNITY | CORPORATE

products & services

- overview
- systems & software
- dna analysis solutions
- rna analysis solutions
- solexa applications
- services
- product literature

- print this page

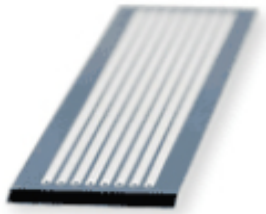
solexa sequencing applications

Illumina's Solexa Sequencing technology offers a powerful new approach to some of today's most important applications for genetic analysis and functional genomics, including:

sequencing and resequencing
Whether you need to sequence an entire genome or a large candidate region, the Illumina Genome Analyzer System is today's most productive and economical sequencing tool. Solexa sequencing technology and reversable terminator chemistry deliver unprecedented volumes of high quality data, rapidly and economically.

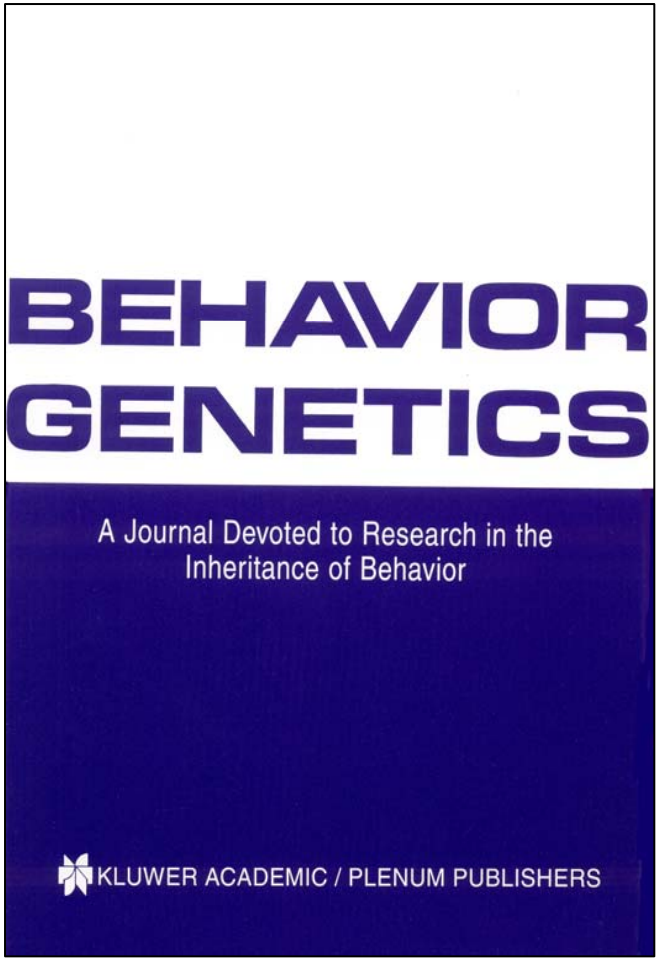
expression profiling
Sequencing millions of short cDNA tags per sample, the Genome Analyzer allows you to generate digital expression profiles at costs comparable to current analog methods. Because our protocol does not require any transcript-specific probes, you can apply the technology to discover and quantitate transcripts in any organisms, irrespective of the annotation available on the organism.

small rna identification and quantification
Solexa sequencing technology also offers a unique and powerful solution for the comprehensive discovery and characterization of small RNAs in a wide range of species. The massively parallel sequencing protocol allows researchers to discover and analyze genome-wide profiles of small RNA in any species. With the potential to generate several million sequence tags economically, the Illumina Genome Analyzer offers investigators the opportunity to uncover global profiles of small RNA at an unprecedented scale.



important information
- product literature
- publications
- faqs
- have a rep contact me

We also run two journals (1)



- Editor: John Hewitt
- Editorial assistant Christina Hewitt
- Publisher: Kluwer /Plenum
- Fully online
- http://www.bga.org

We also run two journals (2)

- Editor: Nick Martin
- Editorial assistant + subscriptions: Lorin Grey
- Publisher: Australian Academic Press
- Fully online
- http://www.ists.qimr.edu.au/journal.html