

Association Mapping

Lon Cardon

SEA-TAC Airport (South Satellite Terminal) &

London Heathrow (Terminal 4)

Outline

- Linkage vs association
- HapMap/SNP discovery enable whole genome association
- Challenges facing whole genome association
- Outlook for future

Whole Genome Association

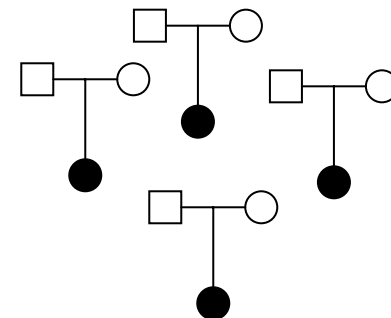
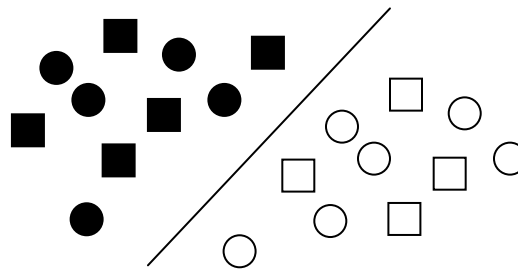
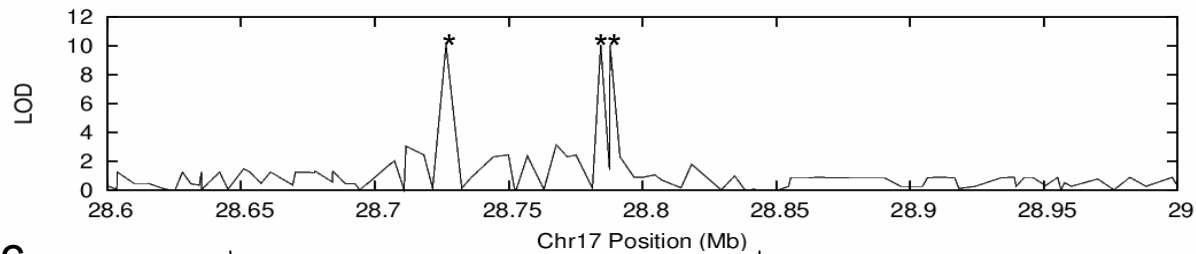
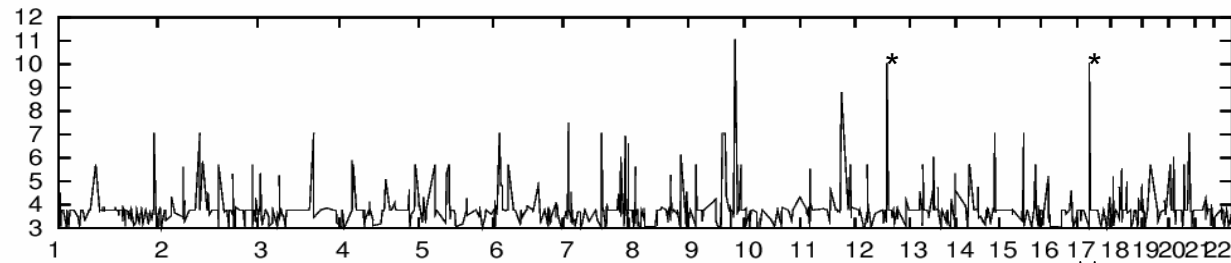
Scan Entire Genome
- 100,000s SNPs



Identify local regions
of interest, examine
genes, SNP density
regulatory regions, etc



Replicate the finding



ARE YOU CAUGHT IN THE NEVER-ENOUGH TRAP? FIND OUT PAGE 20

TRUE STORIES OF
HOPE AND INSPIRATION

Guidenposts

DECEMBER 2006

**SELA
WARD**

**"THE CHRISTMAS
THAT CHANGED
MY LIFE"**

**A BOY'S GIFT,
A TROUBLED
FATHER & THE
LOVE THAT
HEALED
THEM BOTH**

PAGE 50

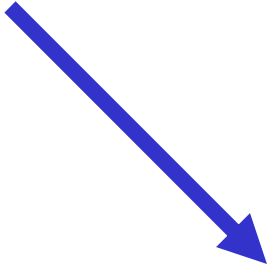
QVC QUACKER QUEEN

**JEANNE BICE
TELLS YOU THE
SECRET OF HER
AMAZING SUCCESS**

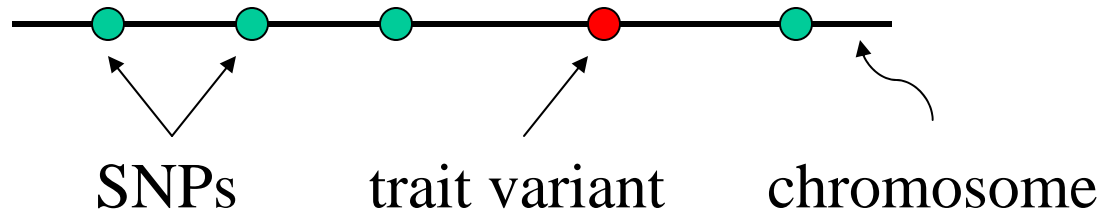
**GENOME
MAPPER
DR. FRANCIS COLLINS
"GOD IS IN
OUR DNA"**

**MYSTERIOUS
HEALINGS
THE STRANGER
AT HIS BEDSIDE**

**The Story of a
PERFECT CHRISTMAS COOKIE**
(recipe included) **PAGE 74**



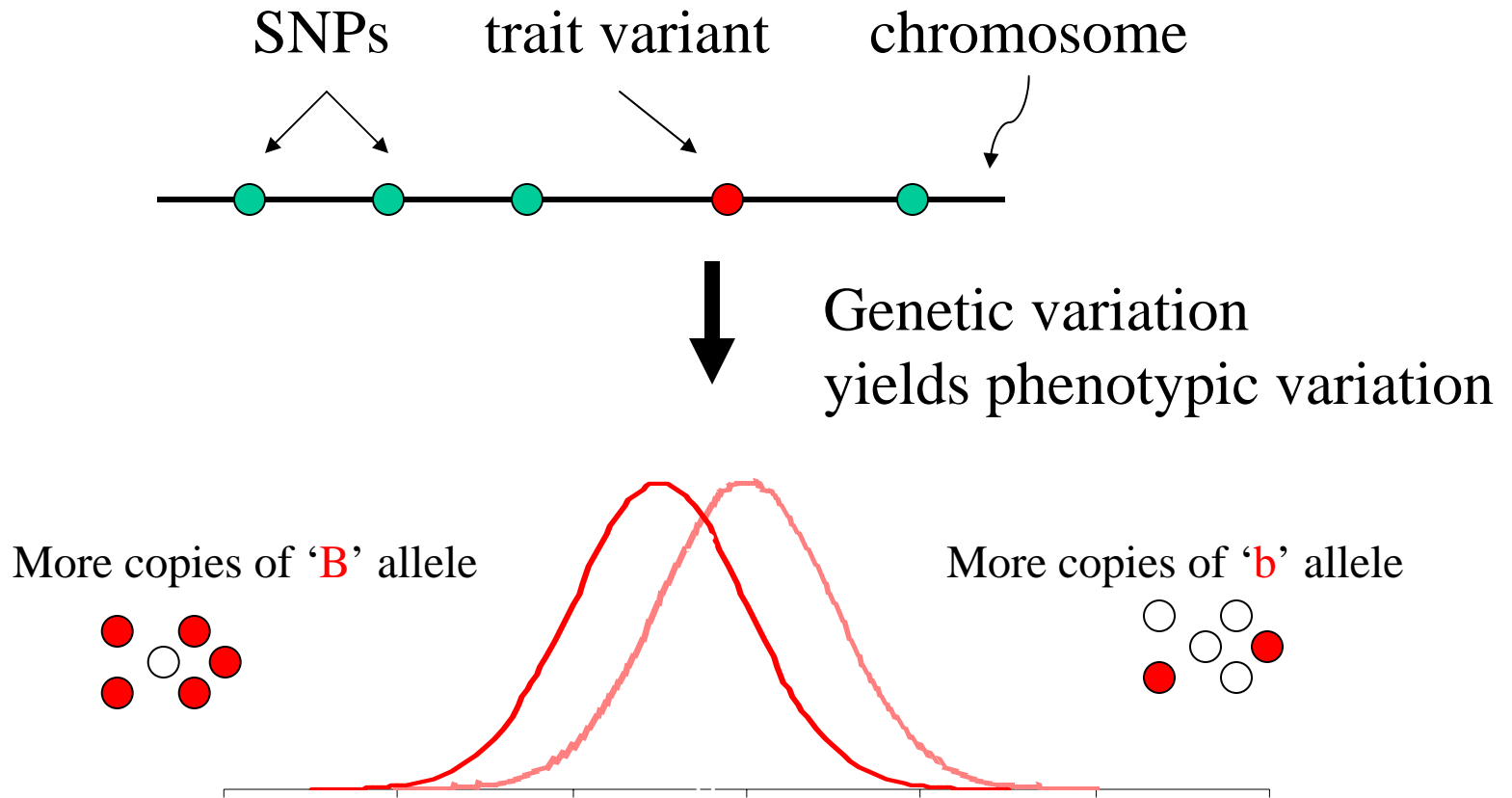
Definitions



Population Data

	Affection	Trait ₁ ...	Trait _n
<p>haplotypes</p>	A	10.3	75.66
<p>genotypes</p>	A	9.9	-99
<p>alleles</p>	U	15.8	101.22

Allelic Association



Simplest Regression Model of Association

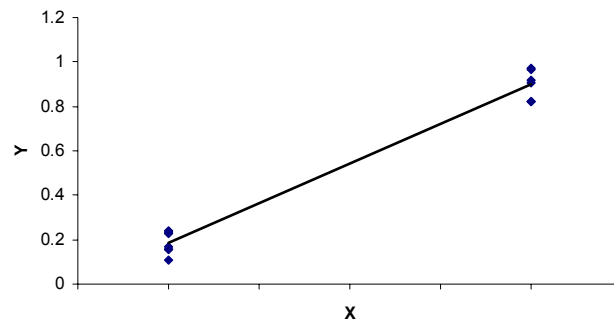
$$Y_i = \alpha + \beta X_i + e_i$$

where

$Y_i =$ trait value for individual i

$X_i =$ 1 if allele individual i has allele 'A'
0 otherwise

i.e., test of mean differences between 'A' and 'not-A' individuals



Association Study Designs and Statistical Methods

- Designs

- Family-based

- Trio (TDT), twins/sib-pairs/extended families (QTDT)

- Case-control

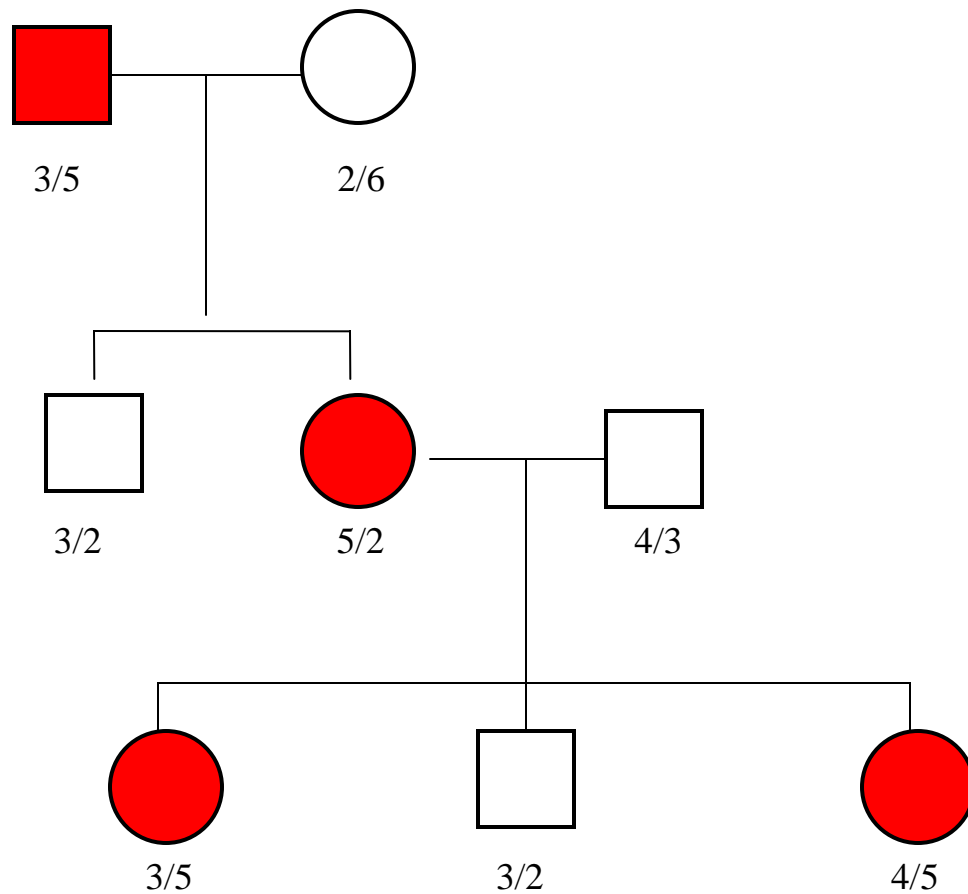
- Collections of individuals with disease, matched with sample w/o disease
 - Some ‘case only’ designs

- Statistical Methods

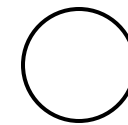
- Wide range: from t-test to evolutionary model-based MCMC

- Principle always same: correlate phenotypic and genotypic variability

Linkage: Allelic association WITHIN FAMILIES



 affected



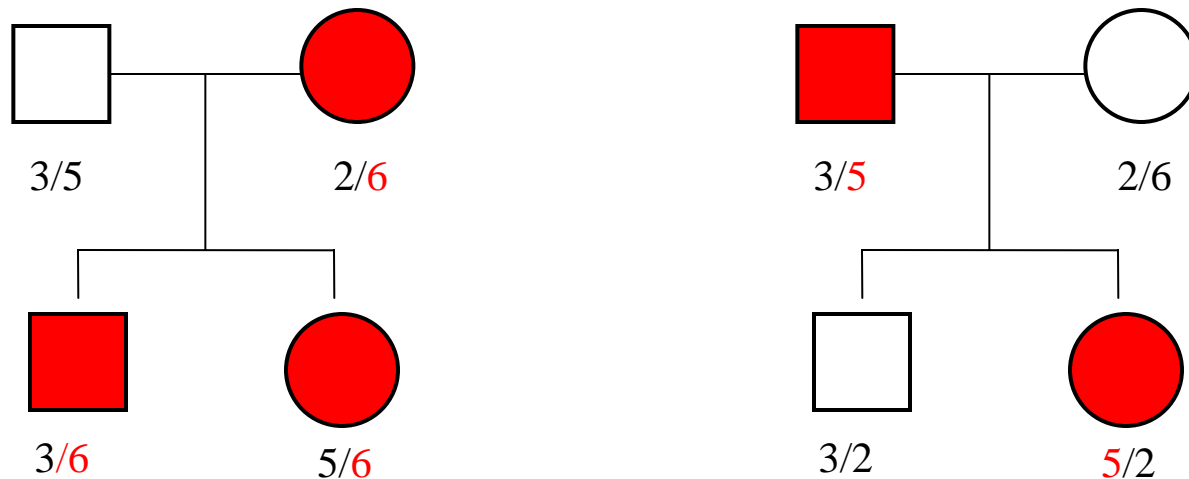
Allele coded by CA copies

2 = CACA

6 = CACACACACACA

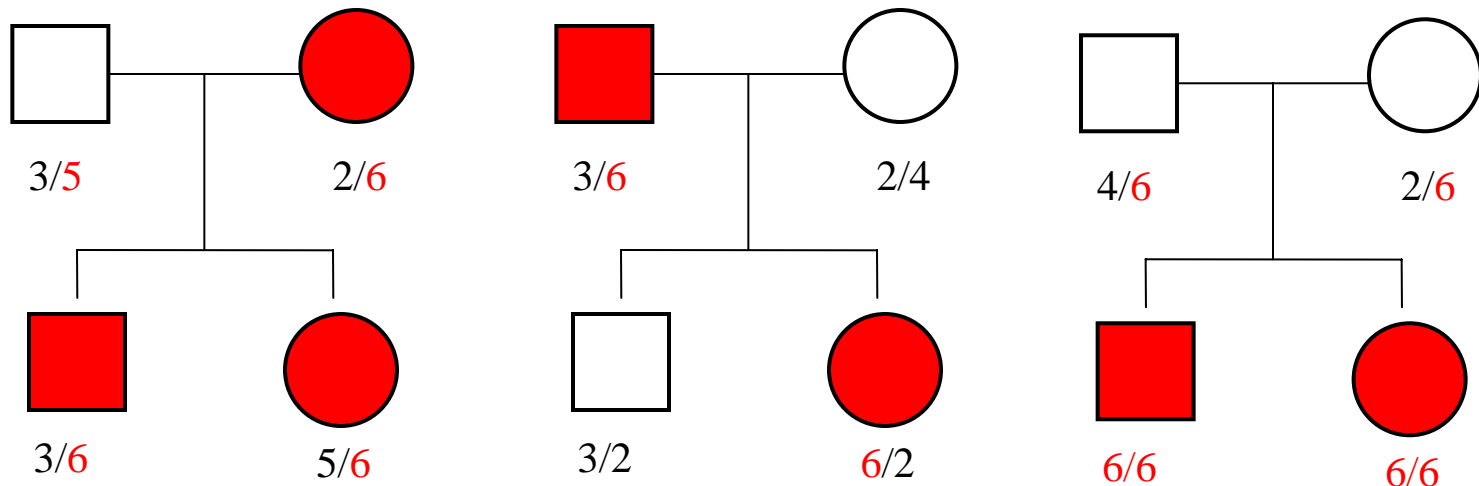
Disease linked to '5'
allele in dominant
inheritance

Allelic Association: Extension of linkage to the population



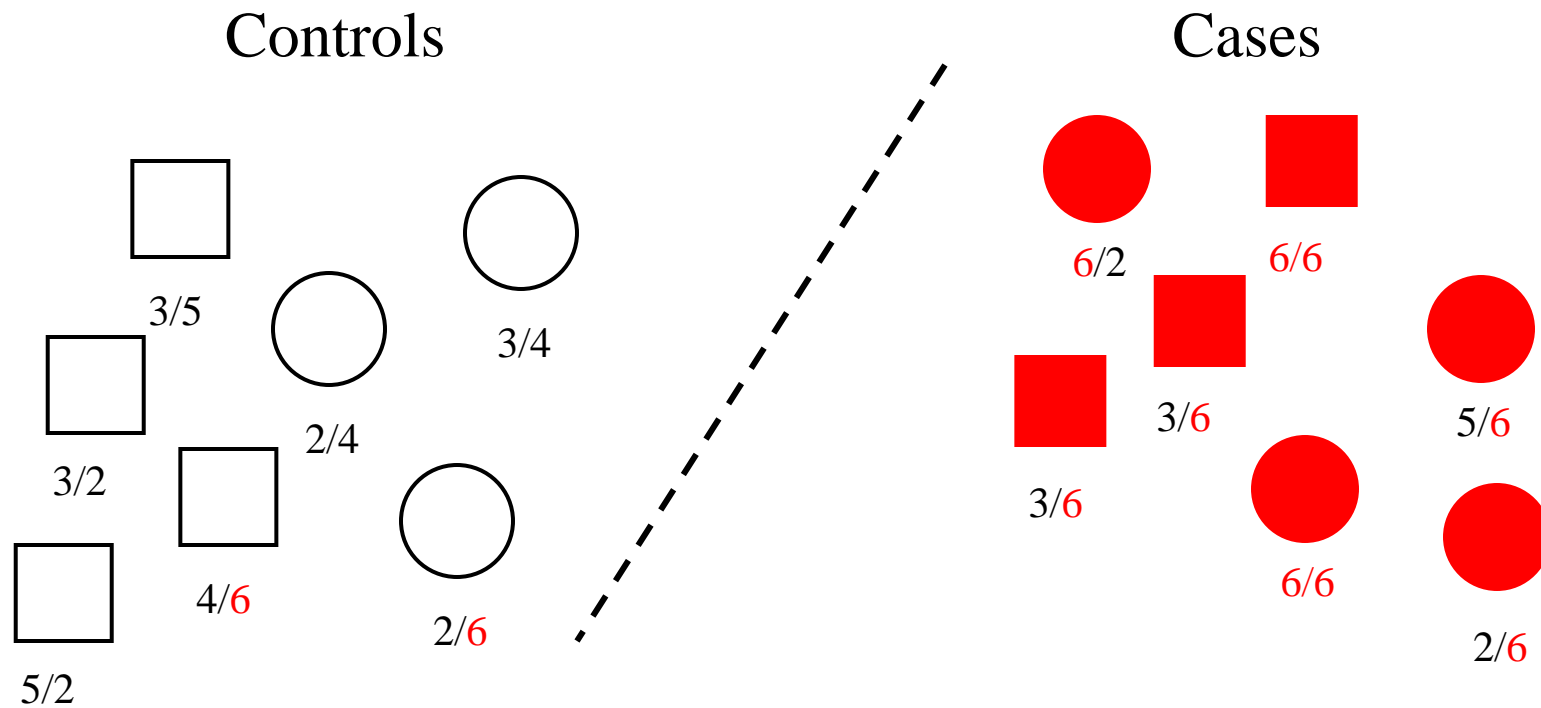
Both families are 'linked' with the marker, but a different allele is involved

Association AND Linkage



All families are 'linked' with the marker
Allele 6 is 'associated' with disease

Allelic Association



Allele 6 is 'associated' with disease

Power of Linkage vs Association

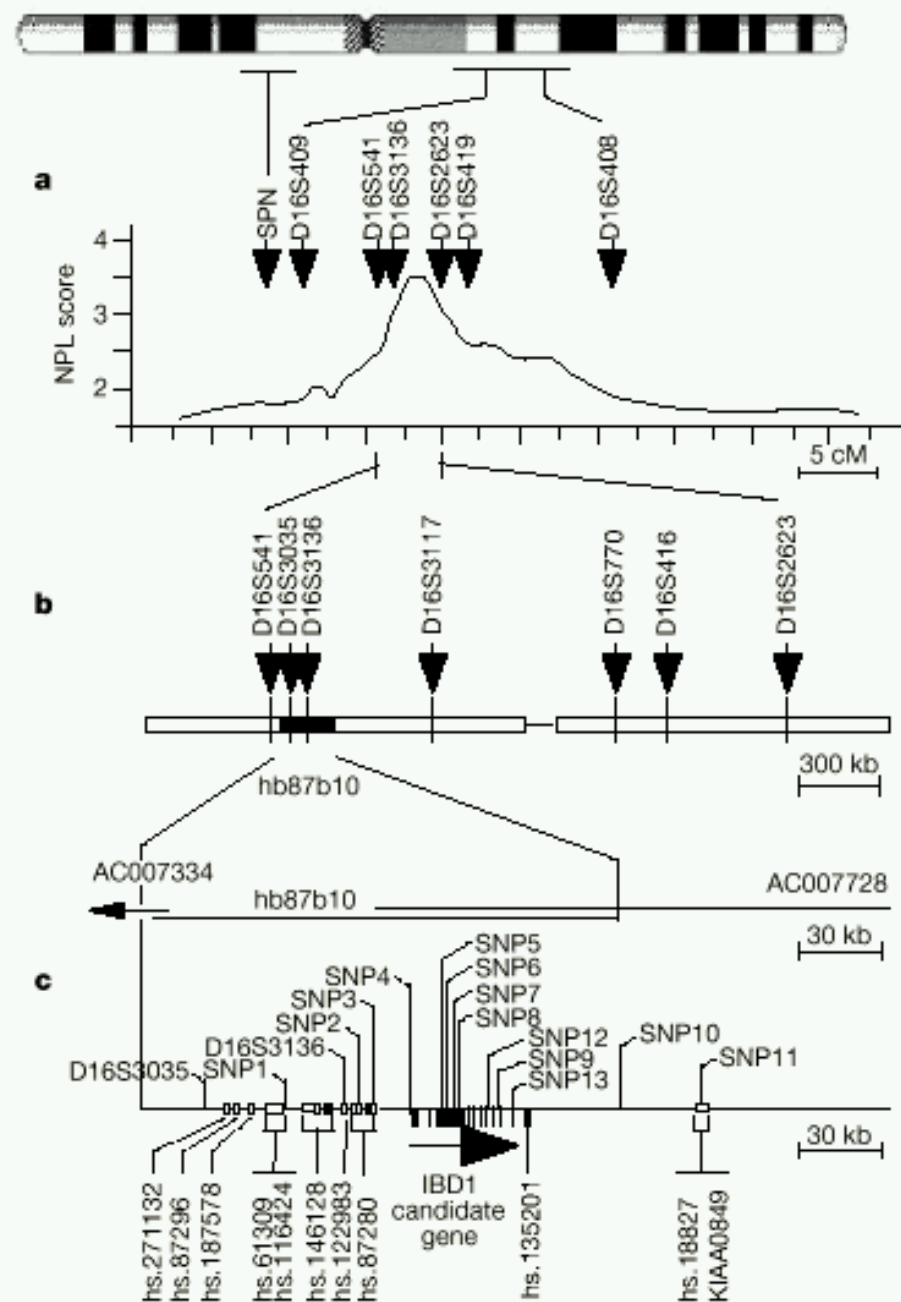
- Association generally has greater power than linkage
 - Linkage based on variances/covariances
 - Association based on means
 - See power lectures in this course

Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease

Jean-Pierre Hugot^{*†‡}, Mathias Chamailard^{*†}, Habib Zouali^{*}, Suzanne Lesage^{*}, Jean-Pierre Cézard[‡], Jacques Belaiche[§], Sven Almer^{||}, Curt Tysk[¶], Colm A. O'Morain[#], Miquel Gassull[☆], Vibeke Binder^{**}, Yigael Finkel^{††}, Antoine Cortot^{‡‡}, Robert Modigliani^{§§}, Pierre Laurent-Puig[†], Corine Gower-Rousseau^{‡‡}, Jeanne Macry^{|||}, Jean-Frédéric Colombel^{‡‡}, Mourad Sahbatou^{*} & Gilles Thomas^{*†§§}

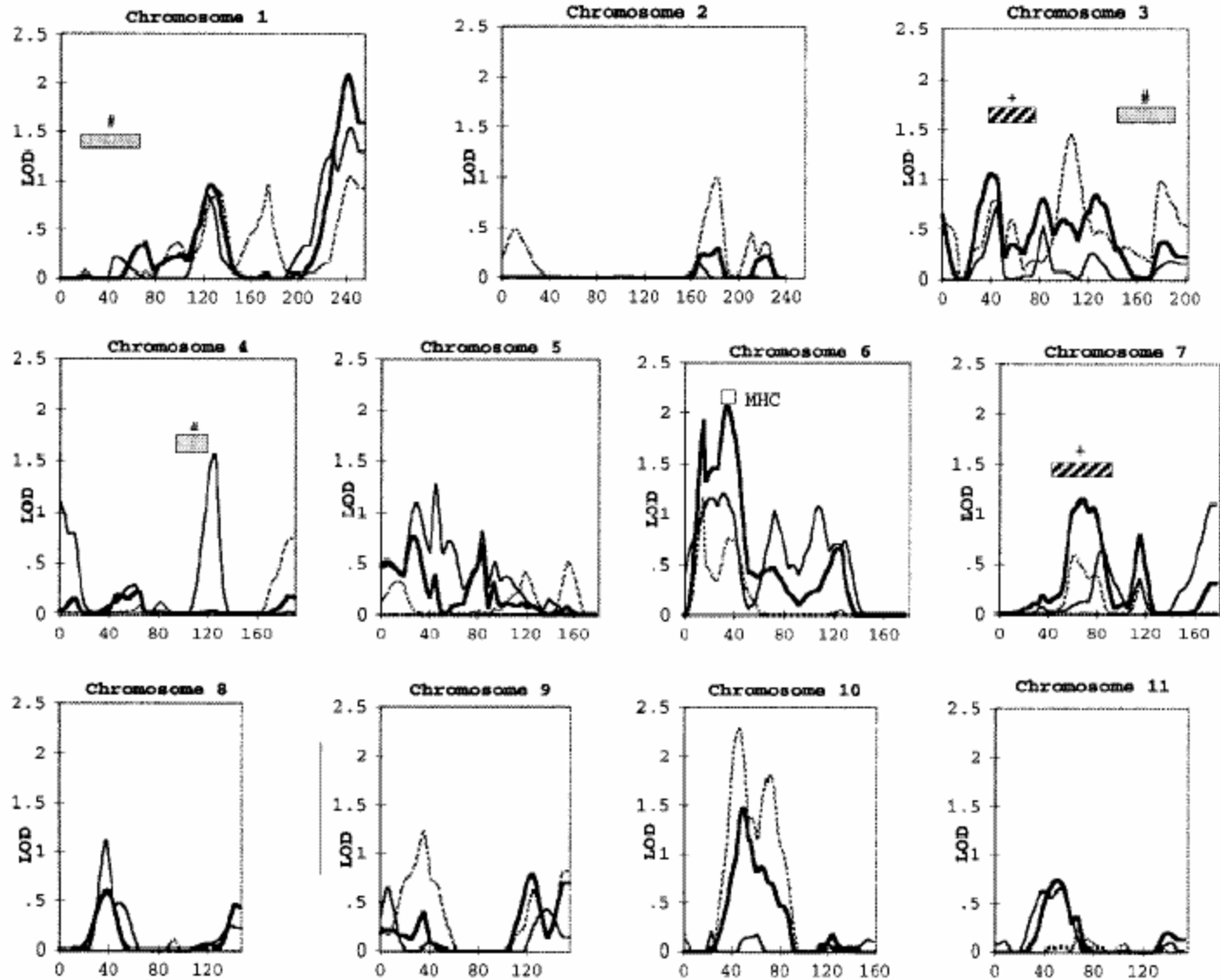
NATURE | VOL 411 | 31 MAY 2001

First (unequivocal) positional cloning of a complex disease gene

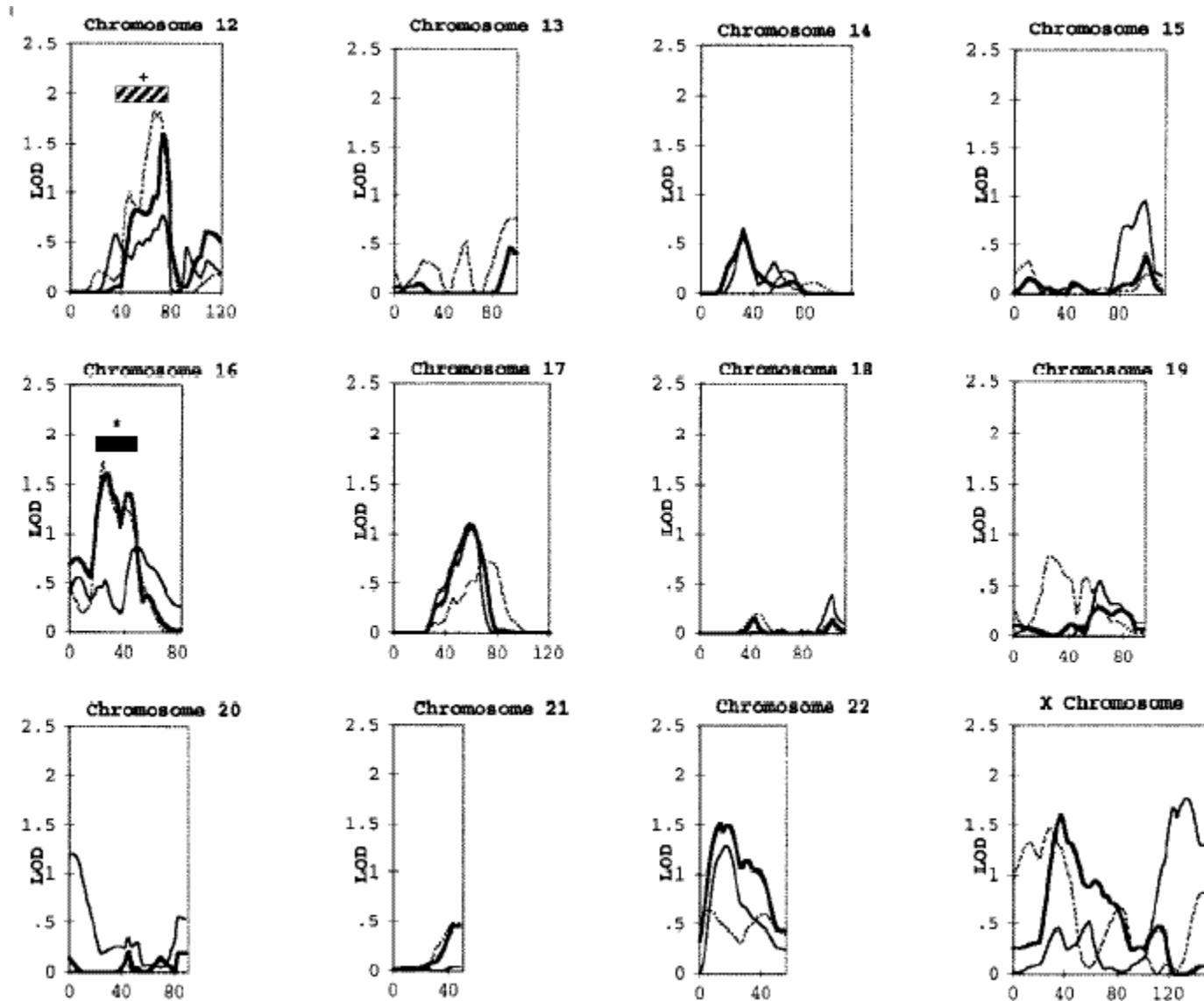


Inflammatory Bowel Disease Genome Screen

Satsangi et al, Nat Genet 1996



Inflammatory Bowel Disease Genome Screen



NOD2 Association Results Stronger than Linkage Evidence

- **Analysis strategy:** same families, same individuals as linkage, but now know mutations. Were the effects there all along?

- **TDT**

TABLE VI. *NOD2* Haplotype Associations With CD

Haplotype ^a	All families n = 294 CD trios ^b	
Pro268/Arg702/Gly908/Leu1007	71 TR : 109 NT	
Pro268Ser /Arg702/Gly908/Leu1007	50 TR : 70 NT	
Pro268Ser/Arg702Trp /Gly908/Leu1007	43 TR : 21 NT	<i>P</i> = 0.01
Pro268Ser /Arg702/Gly908/ Leu1007fsinsC	41 TR : 6 NT	<i>P</i> = 0.000007
Pro268Ser /Arg702/ Gly908Arg /Leu1007	9 TR : 9 NT	<i>P</i> = ns

^aMinor variants shown in bold type.

^b*P*-values shown for over-transmitted haplotypes. Transmission numbers are not identical to Table V, due to haplotype ambiguity and occasional PCR failure.

- **Case-control**

Genotype Rel Risk = 58.9, $p < 10^{-8}$

Same CD cases vs 229 controls

Localization

- **Linkage analysis** yields broad chromosome regions harbouring many genes
 - Resolution comes from recombination events (meioses) in families assessed
 - ‘Good’ in terms of needing few markers, ‘poor’ in terms of finding specific variants involved
- **Association analysis** yields fine-scale resolution of genetic variants
 - Resolution comes from ancestral recombination events
 - ‘Good’ in terms of finding specific variants, ‘poor’ in terms of needing many markers

Linkage vs Association

Linkage

1. Family-based
2. Matching/ethnicity generally unimportant
3. Few markers for genome coverage (300-400 STRs)
4. Can be weak design
5. Good for initial detection; poor for fine-mapping
6. Powerful for rare variants

Association

1. Families or unrelateds
2. Matching/ethnicity crucial
3. Many markers req for genome coverage ($10^5 - 10^6$ SNPs)
4. Powerful design
5. Ok for initial detection; good for fine-mapping
6. Powerful for common variants; rare variants generally impossible

Allelic Association

Three Common Forms

- **Direct Association**
 - Mutant or ‘susceptible’ polymorphism
 - Allele of interest is itself involved in phenotype
- **Indirect Association**
 - Allele itself is not involved, but a nearby correlated marker changes phenotype
- **Spurious association**
 - Apparent association not related to genetic aetiology (most common outcome...)

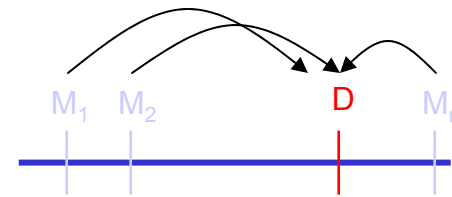
Indirect and Direct Allelic Association

Direct Association



Measure disease relevance (*)
directly, ignoring correlated
markers nearby

Indirect Association & LD



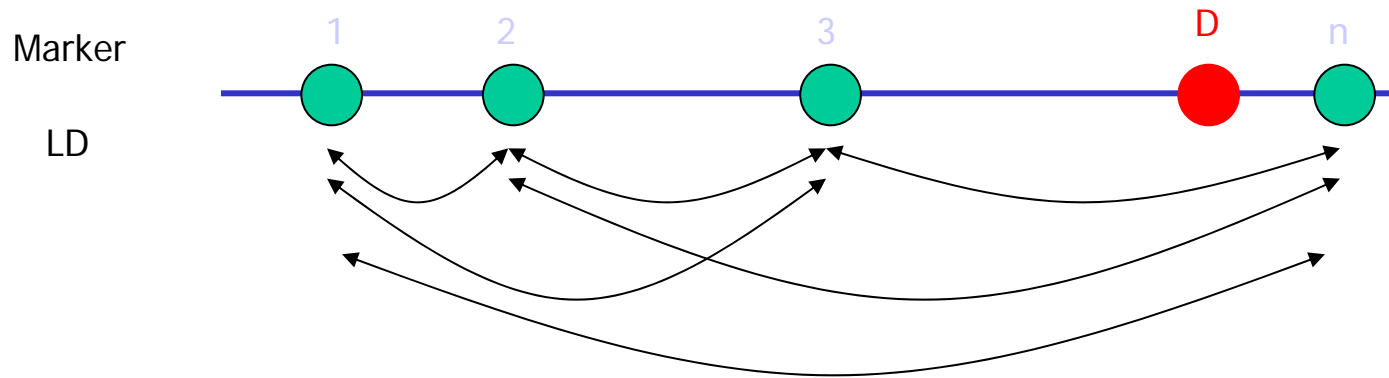
Assess trait effects on **D** via
correlated markers (M_i) rather
than susceptibility/etiologic
variants.

Semantic distinction between

Linkage Disequilibrium: correlation between (any) markers in population

Allelic Association: correlation between marker allele and trait

Linkage Disequilibrium & Allelic Association



Markers close together on chromosomes are often transmitted together, yielding a non-zero correlation between the alleles. This is *linkage disequilibrium*

It is important for allelic association because it means we don't need to assess the exact aetiological variant, but we see trait-SNP association with a neighbouring variant

Building Haplotype Maps for Gene-finding

1. Human Genome Project

→ Good for consensus,
not good for individual
differences



Sept 01



Feb 02



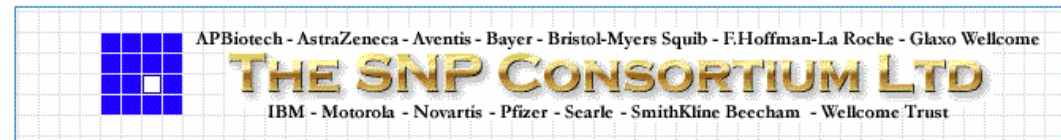
April 04



Oct 04

2. Identify genetic variants

→ Anonymous with respect to
traits.



April 1999 – Dec 01

3. Assay genetic variants

→ Verify polymorphisms,
catalogue correlations
amongst sites

→ Anonymous with respect to
traits



Oct 2002 – 2007...

HapMap Strategy

- Rationale: there are ~10 million common SNPs in human genome
 - We can't afford to genotype them all in each association study
 - But maybe we can genotype them once to catalogue the redundancies and use a smaller set of 'tag' SNPs in each association study
- Samples
 - Four populations, 270 indivs total
- Genotyping
 - 5 kb initial density across genome (600K SNPs)
 - Then second phase to ~ 1 kb across genome (4 million)
 - All data in public domain

Commercial SNP Panels

- Comprise $\approx 100,000 - 550,000$ genetic variants
 - Soon, 1 million
- Cover up to $\sim 85\%$ of common genetic variants

Table 1 Genomic coverage of commercial GWAS products for common SNPs at $r^2 \geq 0.8$, evaluated in Phase II HapMap

	Type	CEU		JPT+CHB		YRI	
		Coverage (%)	Mean r^2	Coverage (%)	Mean r^2	Coverage (%)	Mean r^2
Illumina HumanHap300	Tag	75	0.961	63	0.964	28	0.961
Affymetrix 500K	Random	65	0.975	66	0.974	41	0.971
Affymetrix 111K	Random	31	0.960	31	0.957	15	0.957
Affymetrix 500k + 175K tag	Combination	86	0.975	79	0.978	49	0.973
Illumina Human-1	Gene	26 ^a	0.957	28 ^a	0.955	12 ^a	0.956

Despite the r^2 cutoff of 0.8, the mean r^2 for tagged SNPs is very high; also, 'untagged' SNPs are covered with intermediate values of r^2 , providing modest power to detect such alleles (Supplementary Fig. 1).

^aCoverage estimates for the Human-1 product are underestimates because some of its SNPs were not genotyped in the HapMap project. As these SNPs are largely rare, genic SNPs, it is not expected that they would substantially raise coverage of common variation.

Does having 4 million markers make it easy to find QTLs and disease genes?

- Having more markers makes it easy to do more studies, yes.
- But does it make it easier to find trait-relevant loci?

Historical Performance of Genetic Association Studies

- Pubmed: 27 Feb 2007. “Genetic association” gives 42,294 hits
- 1635 claims of ‘replicated’ genetic association (4%)
- 436 claims of ‘validated’ genetic association (1%)
- In reality, ~ 30-50 confirmed associations for complex traits

Genetic studies of complex diseases have not met anticipated success

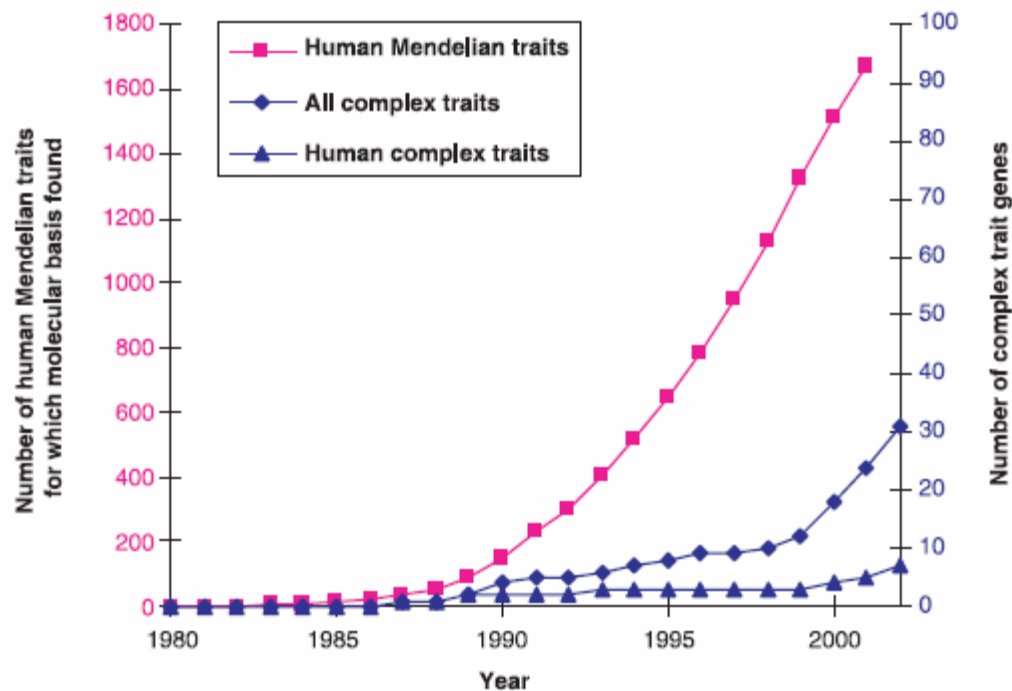
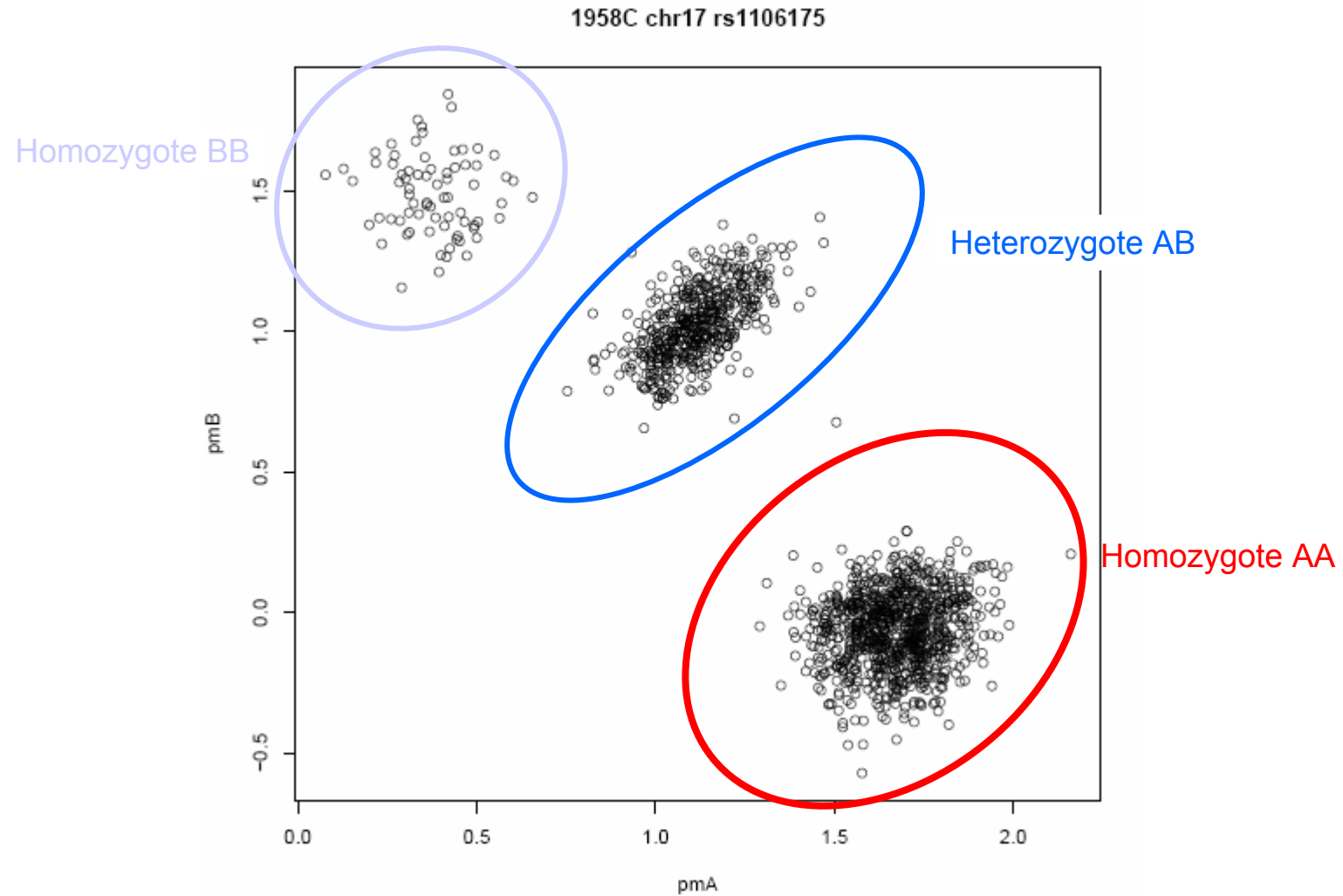


Fig. 1. Identification of genes underlying human Mendelian traits and genetically complex traits in humans and other species. Cumulative data for human Mendelian trait genes (to 2001) include all major genes causing a Mendelian disorder in which causal variants have been identified (58, 59). This reflects mutations in a total of 1336 genes. Complex trait genes were identified by the whole-genome screen approach and denote cumulative year-on-year data described in this review.

Current Association Study Challenges

1) Data Quality

Genotype Calling

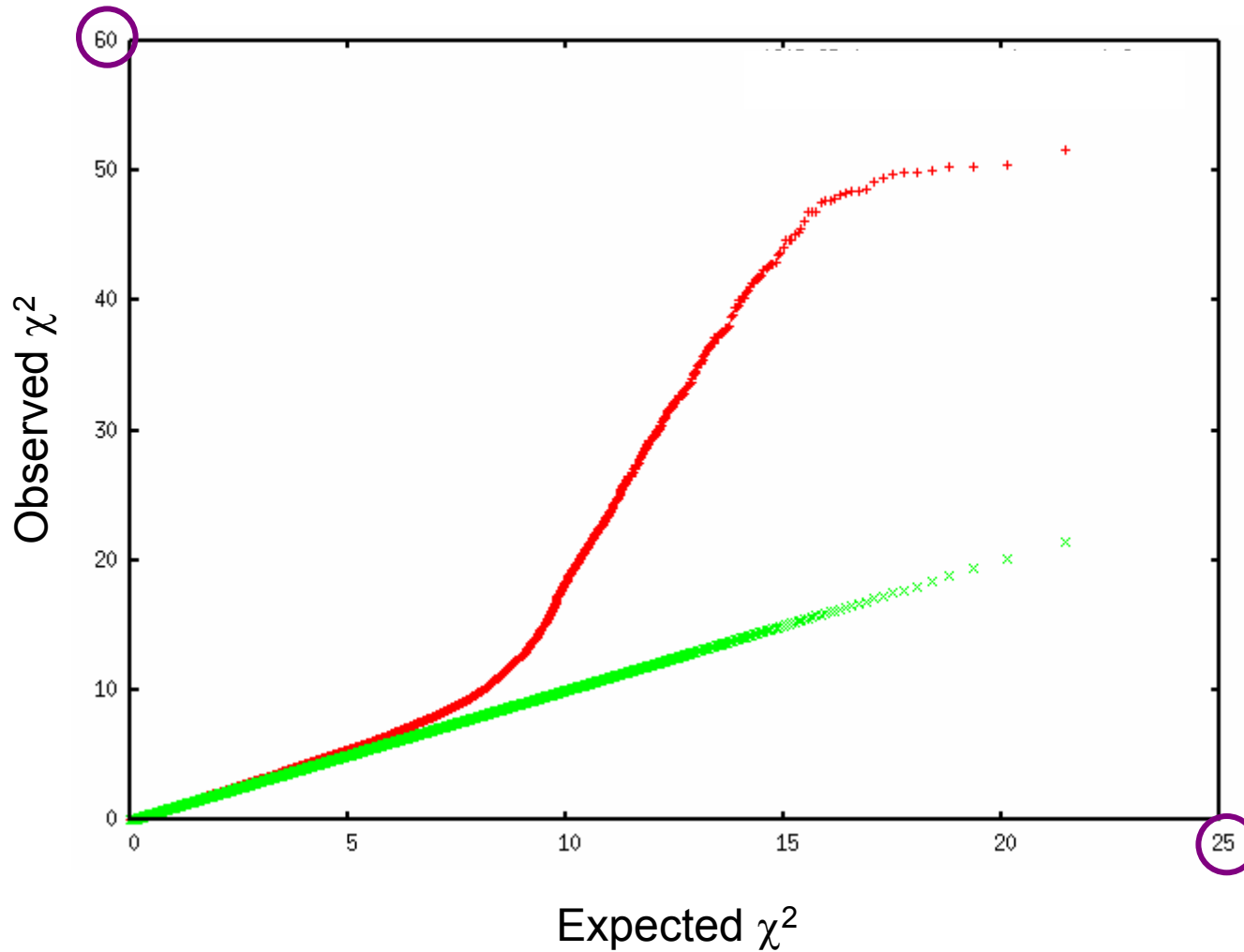


What effect does this have on trait association?

- Following data
 - Affymetrix data
 - Single locus tests
 - ≥ 500 cases/500 controls
 - **Key issue**
 - Genotype calling: batch effects, differential call rates, QC
 - e.g. Clayton et al, Nat Genet 2005

Whole Genome Association

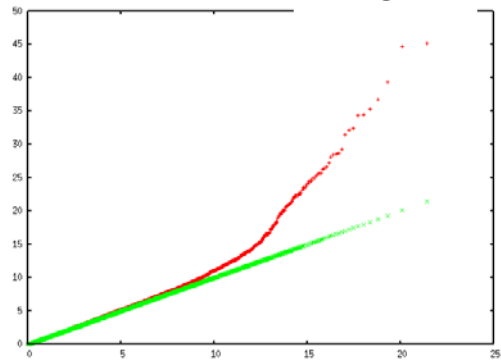
What answer do you want?



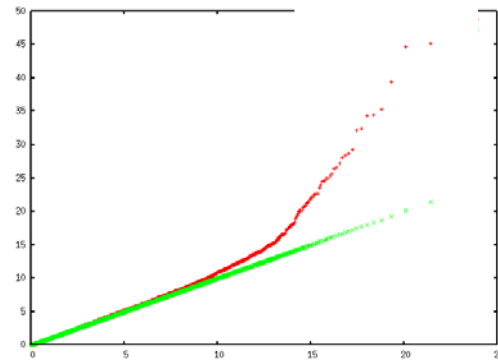
Cleaning Affymetrix Data

Batch Effects and Genotype Calling

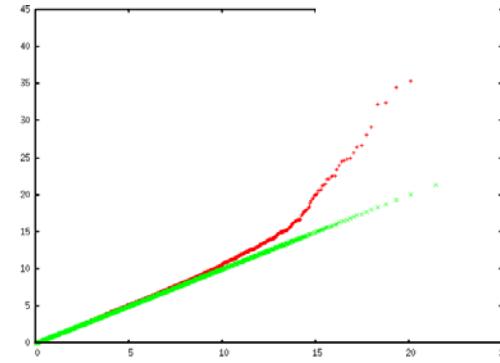
$\leq 10\%$ missing



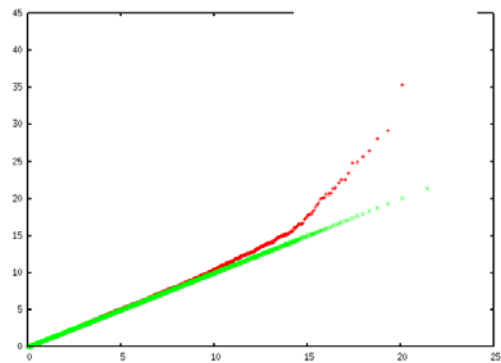
$\leq 9\%$ missing



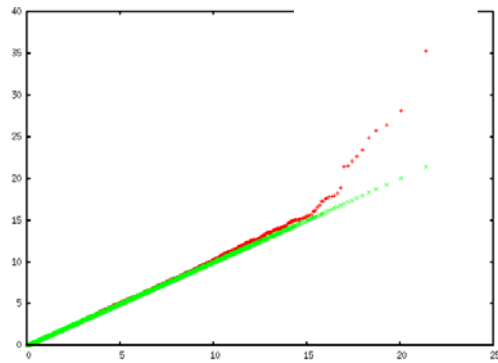
$\leq 8\%$ missing



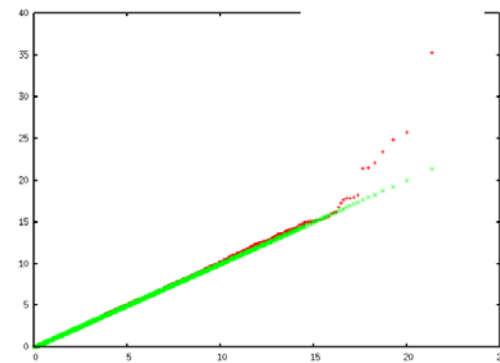
$\leq 7\%$ missing



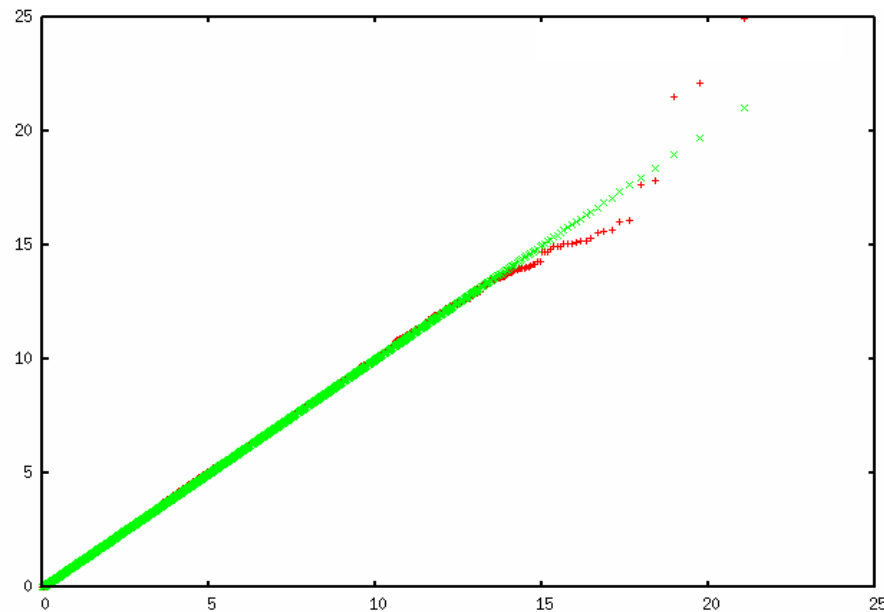
$\leq 6\%$ missing



$\leq 5\%$ missing

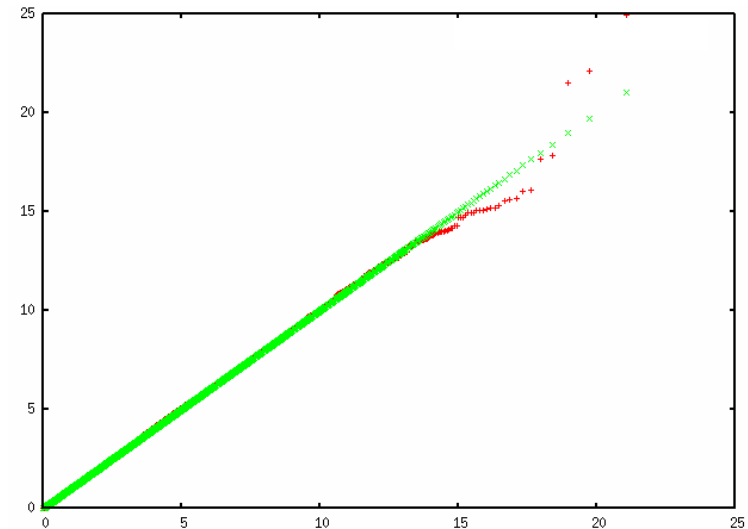
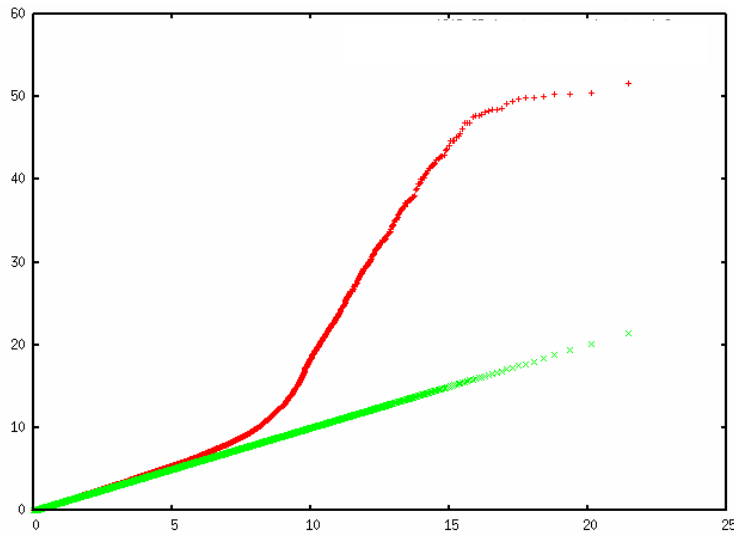


Affymetrix Data – Too Clean?



- As much as 20-30% data eliminated -- including real effects --
- Many 'significant' results can be data errors
 - *'Low Hanging Fruit' sometimes rotten*
- Real effects may not be the most highly significant (power)

Too Many or Too Few?



- Inappropriate genotype calling, study design can mask real effects or make GWA look too good
- How to address this?
 - Multiple controls (e.g., WTCCC)
 - Multiple/better calling algorithms (e.g. Affymetrix)
 - Examination of individual genotypes (manual)

Current Association Study Challenges

2) Do we have the best set of genetic markers

Table 1 | **Priorities for single-nucleotide-polymorphism selection**

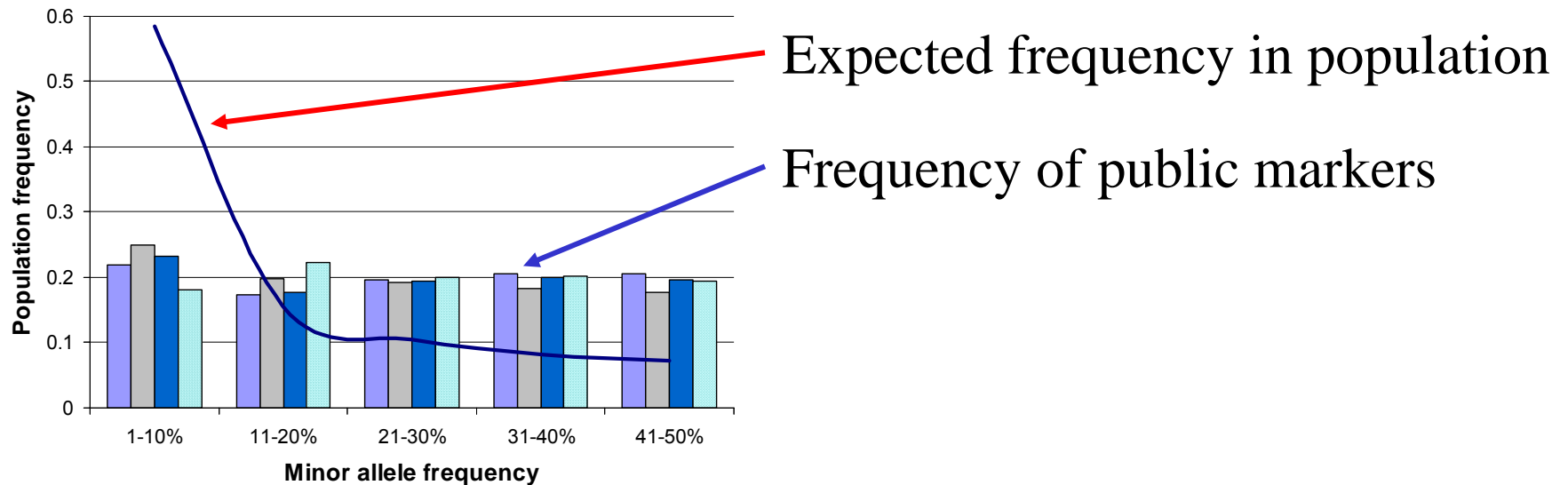
Type of variant	Location	Functional effect	Frequency in genome
Nonsense	Coding sequence	Premature termination of amino-acid sequence	Very low
Missense/ non-synonymous (non-conservative)	Coding sequence	Changes an amino acid in protein to one with different properties	Low
Missense/ non-synonymous (conservative)	Coding sequence	Changes an amino acid in protein to one with similar properties	Low
Insertions/deletions (frameshift)	Coding sequence	Changes the frame of the protein-coding region, usually with very negative consequences for the protein	Low
Insertions/deletions (in frame)	Coding or non-coding	Changes amino-acid sequence	Low
Sense/synonymous	Coding sequence	Does not change the amino acid in the protein – but can alter splicing	Medium
Promoter/regulatory region	Promoter, 5' UTR, 3' UTR	Does not change the amino acid, but can affect the level, location or timing of gene expression	Low to medium
Splice site/intron–exon boundary	Within 10 bp of the exon	Might change the splicing pattern or efficiency of introns	Low
Intronic	Deep within introns	No known function, but might affect expression or mRNA stability	Medium
Intergenic	Non-coding regions between genes	No known function, but might affect expression through enhancer or other mechanisms	High

Current Association Study Challenges

2) Do we have the best set of genetic markers

There exist 6 million putative SNPs in the public domain. Are they the right markers?

Allele frequency distribution is biased toward common alleles



Current Association Study Challenges

3) How to analyse the data

- **Allele based test?**
 - 2 alleles \rightarrow 1 df
 - $E(Y) = a + bX$ $X = 0/1$ for presence/absence
- **Genotype-based test?**
 - 3 genotypes \rightarrow 2 df
 - $E(Y) = a + b_1A + b_2D$ $A = 0/1$ additive (hom); $W = 0/1$ dom (het)
- **Haplotype-based test?**
 - For M markers, 2^M possible haplotypes $\rightarrow 2^M - 1$ df
 - $E(Y) = a + \sum bH$ H coded for haplotype effects
- **Multilocus test?**
 - Epistasis, $G \times E$ interactions, many possibilities

Current Association Study Challenges

4) Multiple Testing

- **Candidate genes:** a few tests (probably correlated)
- **Linkage regions:** 100's – 1000's tests (some correlated)
- **Whole genome association:** 100,000s – 1,000,000s tests (many correlated)
- **What to do?**
 - Bonferroni (conservative)
 - False discovery rate?
 - Permutations?
 -Area of active research

Current Association Study Challenges

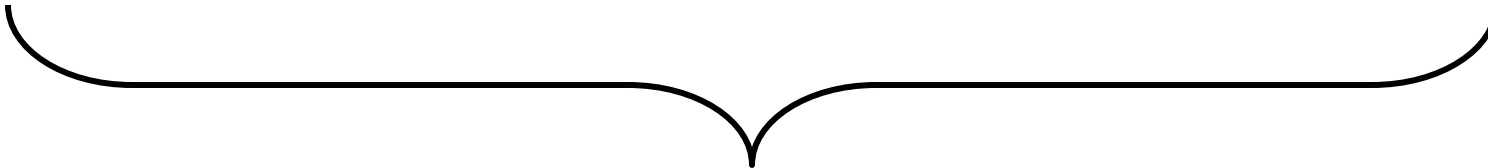
5) Population Stratification

Analysis of mixed samples having different allele frequencies is a primary concern in human genetics, as it leads to false evidence for allelic association.

This is the main blame for past failures of association studies

Population Stratification

Sample 'A'				+	Sample 'B'			
	M	m	Freq.		M	m	Freq.	
Affected	50	50	.10		1	9	.01	
Unaffected	450	450	.90		99	891	.99	
	.50	.50			.10	.90		
χ^2_1 is n.s.				χ^2_1 is n.s.				



	M	m	Freq.
Affected	51	59	.055
Unaffected	549	1341	.945
	.30	.70	

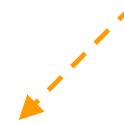
$$\chi^2_1 = 14.84, p < 0.001$$

Spurious Association

Population Stratification: Real Example

Full heritage American Indian Population		
	+	-
Gm ^{3;5,13,14}	~1%	~99%
(NIDDM Prevalence ≈ 40%)		

Caucasian Population		
	+	-
Gm ^{3;5,13,14}	~66%	~34%
(NIDDM Prevalence ≈ 15%)		



Study without knowledge of genetic background:

Gm ^{3;5,13,14} haplotype	Cases	Controls
+	7.8%	29.0%
-	92.2%	71.0%

OR=0.27
95%CI=0.18 to 0.40



Proportion with NIDDM by heritage and marker status

<i>Index of Indian Heritage</i>	Gm ^{3;5,13,14} haplotype	
	+	-
0	17.8%	19.9%
4	28.3%	28.8%
8	35.9%	39.3%

Current Association Study Challenges

6) What constitutes a replication?

GOLD Standard for association studies

Replicating association results in different laboratories is often seen as most compelling piece of evidence for 'true' finding

But.... in any sample, we measure

Multiple traits

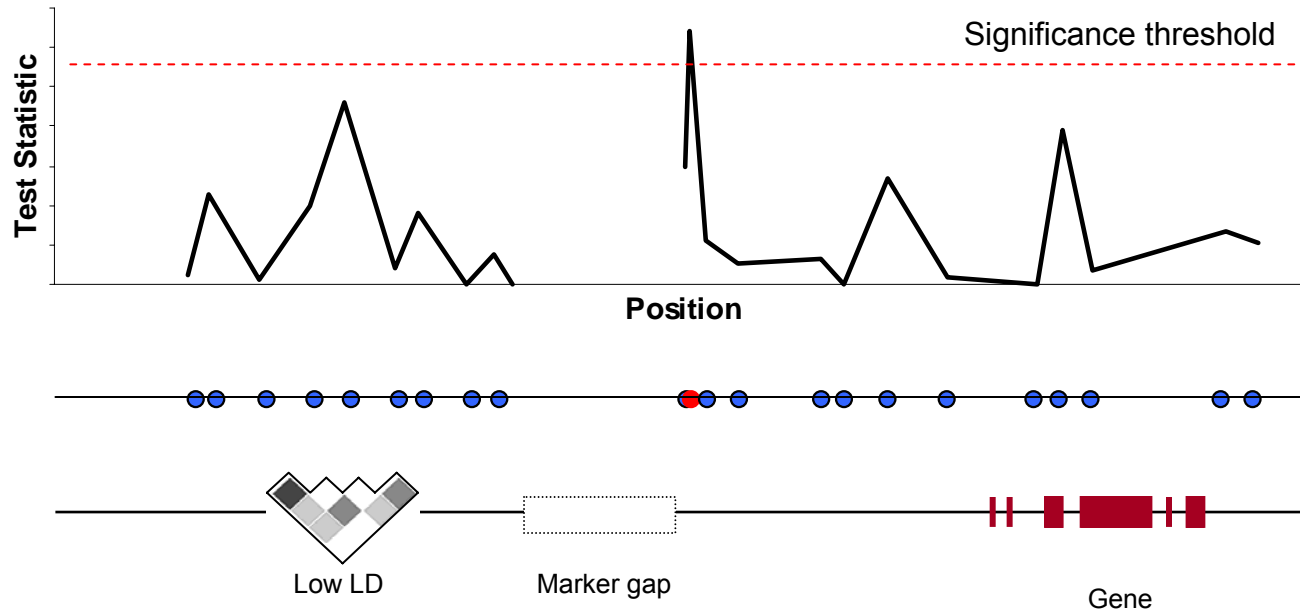
Multiple genes

Multiple markers in genes

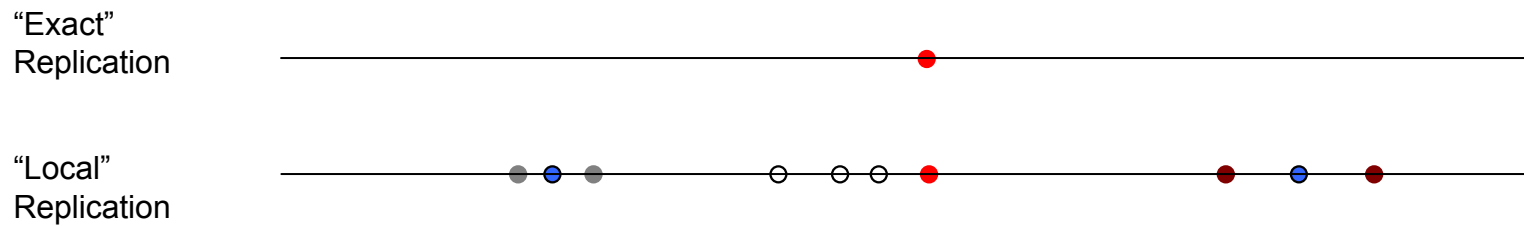
and we analyse all this using multiple statistical tests

What is a true replication?

Initial Study



Replication Strategy



What is a true replication?

Replication Outcome

- Association to same trait, but different gene
- Association to same trait, same gene, different SNPs (or haplotypes)
- Association to same trait, same gene, same SNP – but in opposite direction (protective $\leftarrow\rightarrow$ disease)
- Association to different, but correlated phenotype(s)
- No association at all

Explanation

- Genetic heterogeneity
- Allelic heterogeneity
- Allelic heterogeneity/popln differences
- Phenotypic heterogeneity
- Sample size too small

Measuring Success by Replication

- Define objective criteria for what is/is not a replication *in advance*
- Design initial and replication study to have enough power
 - ‘Lumper’: use most samples to obtain robust results in first place
 - Great initial detection, may be weak in replication
 - ‘Splitter’: Take otherwise large sample, split into initial and replication groups
 - One good study → two bad studies.
 - Poor initial detection, poor replication

Despite challenges: upcoming association studies hold promise

- Large, epidemiological-sized samples emerging
- Availability of millions of genetic markers
 - Genotyping costs decreasing rapidly
- Background LD patterns characterized
 - International HapMap and other projects

GWA: Recent Success

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 Hoh⁷ †

www.sciencemag.org SCIENCE VOL 308 15 APRIL 2005

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,^{1,10} 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,^{4,2,4} Lisa Wu Datta,⁵ Emily O. Kistner,¹⁵ L. Philip Schumm,¹⁵ Annette T. Lee,¹⁶ Peter K. Gregersen,¹⁶ M. Michael Bamada,² Jerome I. Rotter,^{3,4} Dan L. Nicolae,^{13,2,7} Judy H. Cho^{12*}

www.sciencemag.org SCIENCE VOL 314 1 DECEMBER 2006

HTRA1 Promoter Polymorphism in V Age-Related Macular Degeneration

Andrew DeWan,¹ Mugen Liu,^{2*} Stephen Hartman,^{3*} Samuel Shao-Min Zhang,^{2*} David T. Connie Zhao,⁵ Pancy O. S. Tam,⁴ Wai Man Chan,⁴ Dennis S. C. Lam,⁴ Michael Snyder,¹ Colin Barnstable,² Chi Pui Pang,⁴ Josephine Hoh^{1,2} †

www.sciencemag.org SCIENCE VOL 314 10 NOVEMBER 2006

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 Shen¹, David Serre¹, Philippe Boutin⁵, Daniel Vincent⁴, Alexandre Belisle⁴, Samy Hadjadj⁶, Beverley Balkau⁷, Barbara Heude⁷, Guillaume Charpentier⁸, Thomas J. Hudson^{4,9}, Alexandre Montpetit⁴, Alexey V. Pshezhetsky¹⁰, Marc Prentki^{10,11}, Barry I. Posner^{2,12}, David J. Balding¹³, David Meyre⁵, Constantin Polychronakos^{1,3} & Philippe Froguel^{5,14}

doi:10.1038/nature05616

nature

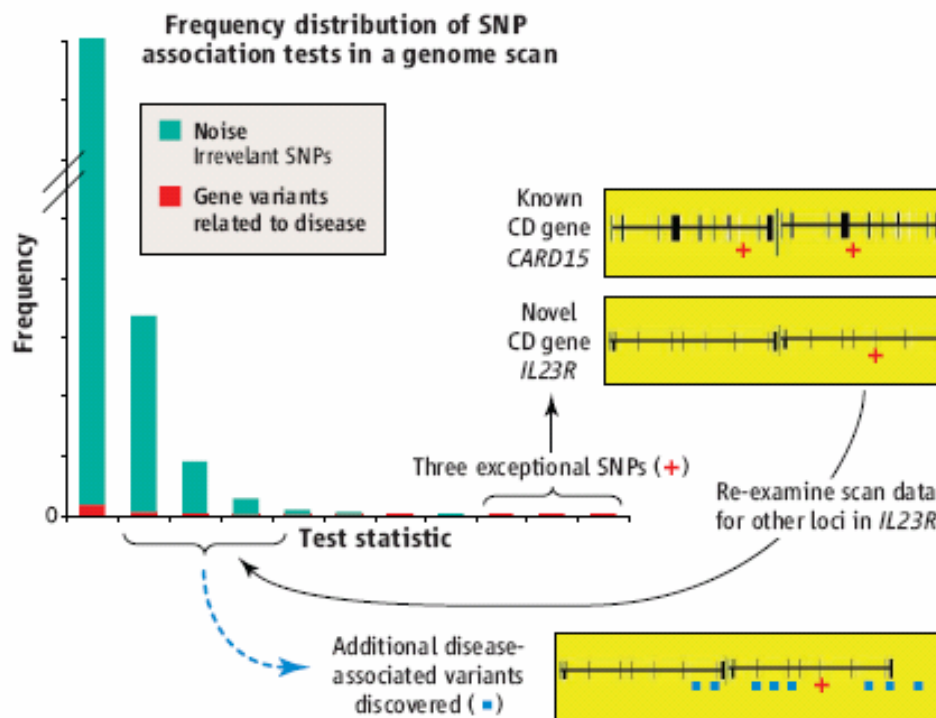
A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*

Jochen Hampe^{1,2,10}, Andre Franke^{1,10}, Philip Rosenstiel^{1,9}, Andreas Till¹, Markus Teuber¹, Klaus Huse³, Mario Albrecht⁴, Gabriele Mayr⁴, Francisco M De La Vega⁵, Jason Briggs⁵, Simone Günther⁵, Natalie J Prescott⁶, Clive M Onnie⁶, Robert Häslér¹, Bence Sipos⁷, Ulrich R Fölsch², Thomas Lengauer⁴, Matthias Platzer³, Christopher G Mathew⁶, Michael Krawczak⁸ & Stefan Schreiber^{1,2}

NATURE GENETICS VOLUME 39 | NUMBER 2 | FEBRUARY 2007

IL23R-Crohn's Disease Finding

- 500 cases/controls
 - Illumina 317k
 - 3 highly significant SNPs
 - 2 in CARD15 (known)
 - 1 novel (IL23R)
 - 2 independent replications
-
- Highly significant SNPs led them to look at less significant SNPs
- Multiple independent associations



Cardon, Science, 2006

IL23R is real: GWA can work

Replication in Oxford Samples

(subset of WTCCC)

- 604 cases/1149 controls
- Genotyped same markers
- Used same statistical procedures

Results

- Convincing replication of main findings
- No clinical specificity
- Same direction of effect
- Accurate effect sizes (smaller)
- Epistasis?

SNP	CD		P-value	OR
	Cases	Controls		
rs1004819	0.371	0.3002	7.03E-05	1.37 (1.17-1.60)
rs7517847	0.344	0.4472	2.07E-08	0.65 (0.55-0.75)
rs10489629	0.386	0.455	1.60E-04	0.75 (0.65-0.87)
rs2201841	0.369	0.3057	3.20E-04	1.33 (1.14-1.55)
rs11209026	0.028	0.06011	8.20E-05	0.46(0.31-0.68)
rs1343151	0.278	0.3393	4.00E-04	0.75 (0.63-0.88)
rs11209032	0.389	0.3404	0.006604	1.23 (1.06-1.43)
rs1495965	0.505	0.4738	0.08752	1.14 (0.98-1.31)

- All carriers of rare *protective* allele carry at least 1 *IBD5 risk* haplotype

	A	G
IBD5-ve	0	294
IBD5+ve	30	814

2007: The Year of Whole Genome Association

- There are ~ 20 studies nearing completion now
- Many of them have new findings
 - Not 100s of new genes, but not 0 either
- They are being replicated and validated externally
- All data will go into public domain

- Association studies do work, but they don't find everything