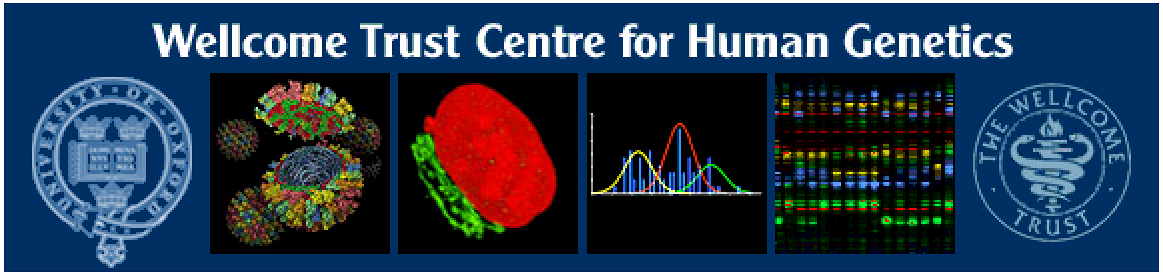
# Association Mapping and the Human Genome

Lon Cardon

Wellcome Trust Centre for Human Genetics



# Association Study Applications

Candidate genes for specific diseases

common practice in medicine/genetics

Pharmacogenetics

genotyping clinically relevant samples (toxicity vs efficacy)

Insurance purposes

contentious, but likely at some point

Positional cloning

the most frequent source of new loci at present

Genome-wide association

with millions available SNPs, can search whole genome exhaustively

# Association Studies and the Human Genome

1. Mendelian disorders and positional cloning
2. Complex trait association models
3. Current status
4. Near-term challenges

# Mendelian Disorders

- Measured phenotype caused by single gene
  - May have multiple mutations in gene
  - May be additional (presumably environmental) causes
- Follow clear segregation patterns in families
- Typically rare in population
- Examples
  - Duchenne Muscular Dystropy
  - Cystic Fibrosis (1989)
  - Huntington's Disease (1993)
  - ~ 1200 have been mapped

# Positional Cloning

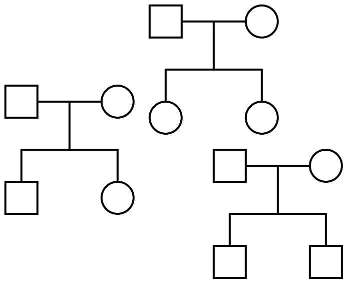
*The identification of a gene based solely on its position in the genome*

- Most widespread strategy in human genetics in past 15 years
- Most ongoing association studies initiated on basis of this model
- Strengths
  - No knowledge of function of gene product required
  - Very strong track record in single gene disorders
- Weaknesses
  - Understanding of function not a certain outcome
  - Poor track record with multifactorial conditions

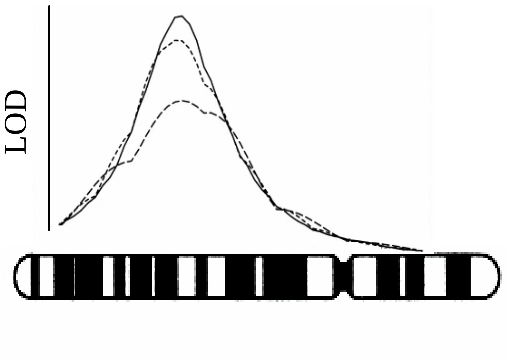
# Positional Cloning

## Genetics

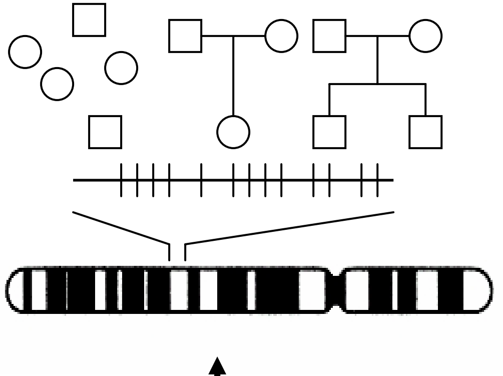
Families



Chromosome Region



Association Study

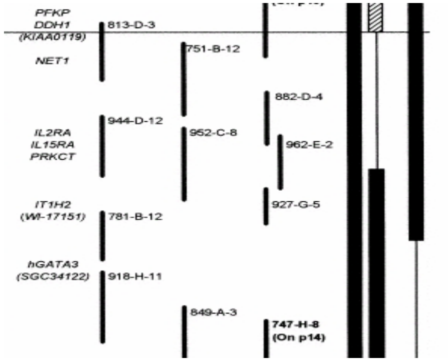


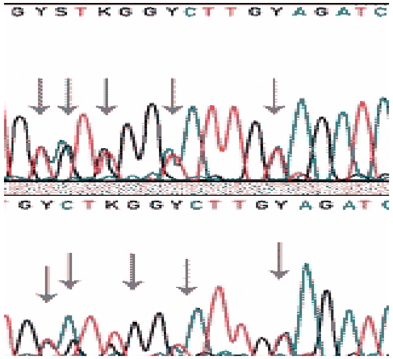
## Genomics

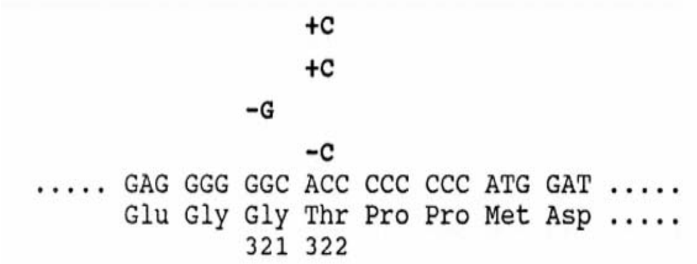
Physical Mapping/
Sequencing

Candidate Gene Selection/
Polymorphism Detection

Mutation Characterization/
Functional Annotation







```
                    +C
                    +C
                -G
                    -C
..... GAG GGG GGC ACC CCC CCC ATG GAT .....
      Glu Gly Gly Thr Pro Pro Met Asp .....
                321 322
```

# Genetic Linkage

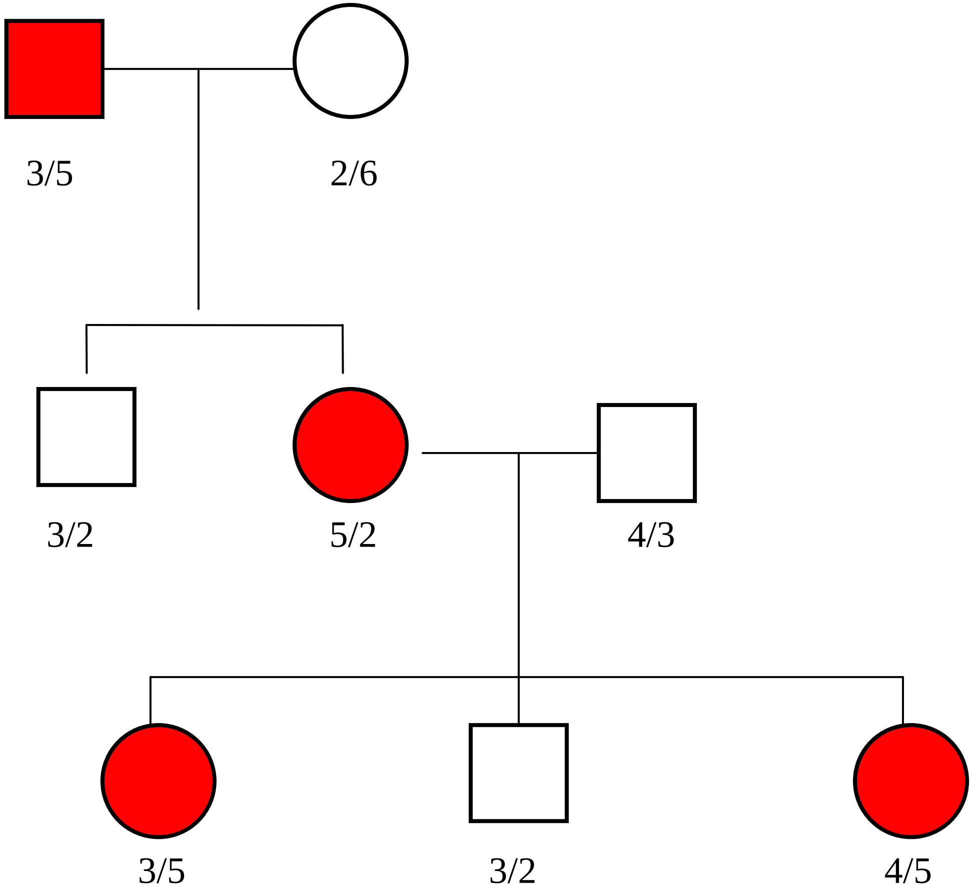
*Co-segregation of marker alleles with disease alleles* *within families*

Aim: Identify broad chromosome regions (20-30 cM)
harbouring etiologic variants
(~200 – 400+ genes)

Requirements:

(i) Many families with trait of interest
(ii) Informative marker panels

# Single Gene Linkage Analysis



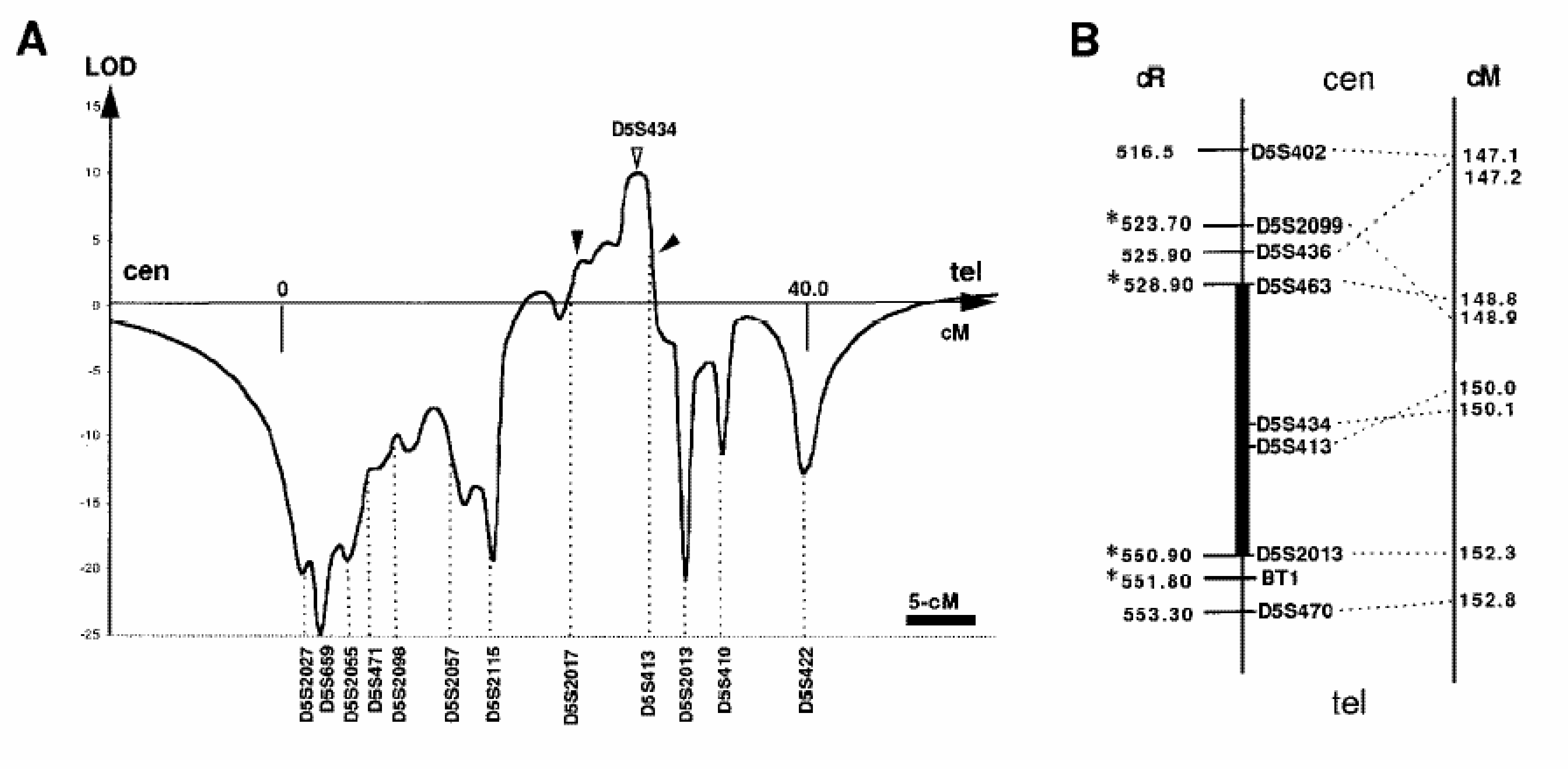
■ affected

○ unaffected

Allele coded by CA copies
2 = CACA
6 = CACACACACACA

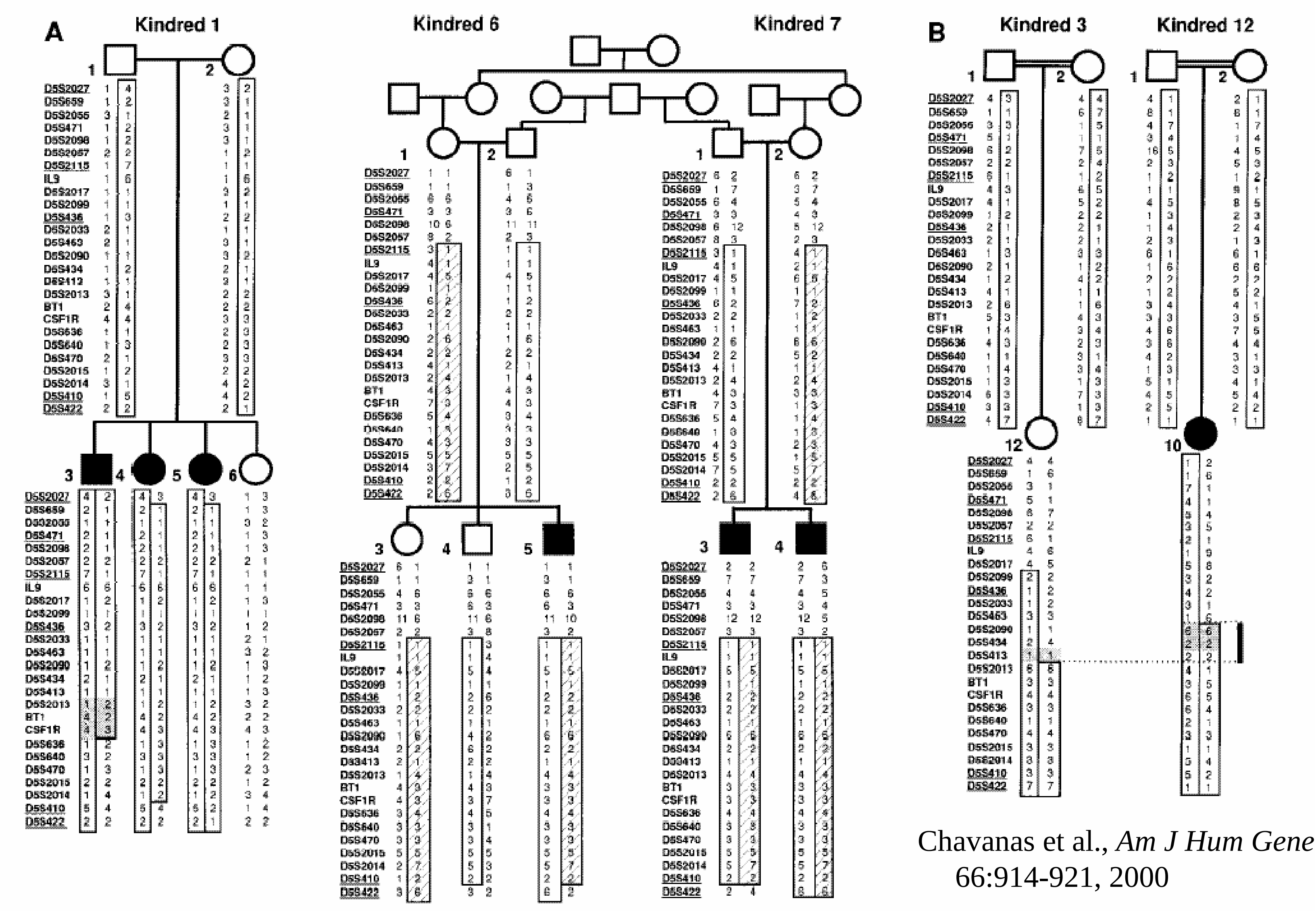
Disease linked to ‘5’
allele in dominant
inheritance pattern

# Netherton Syndrome Linkage

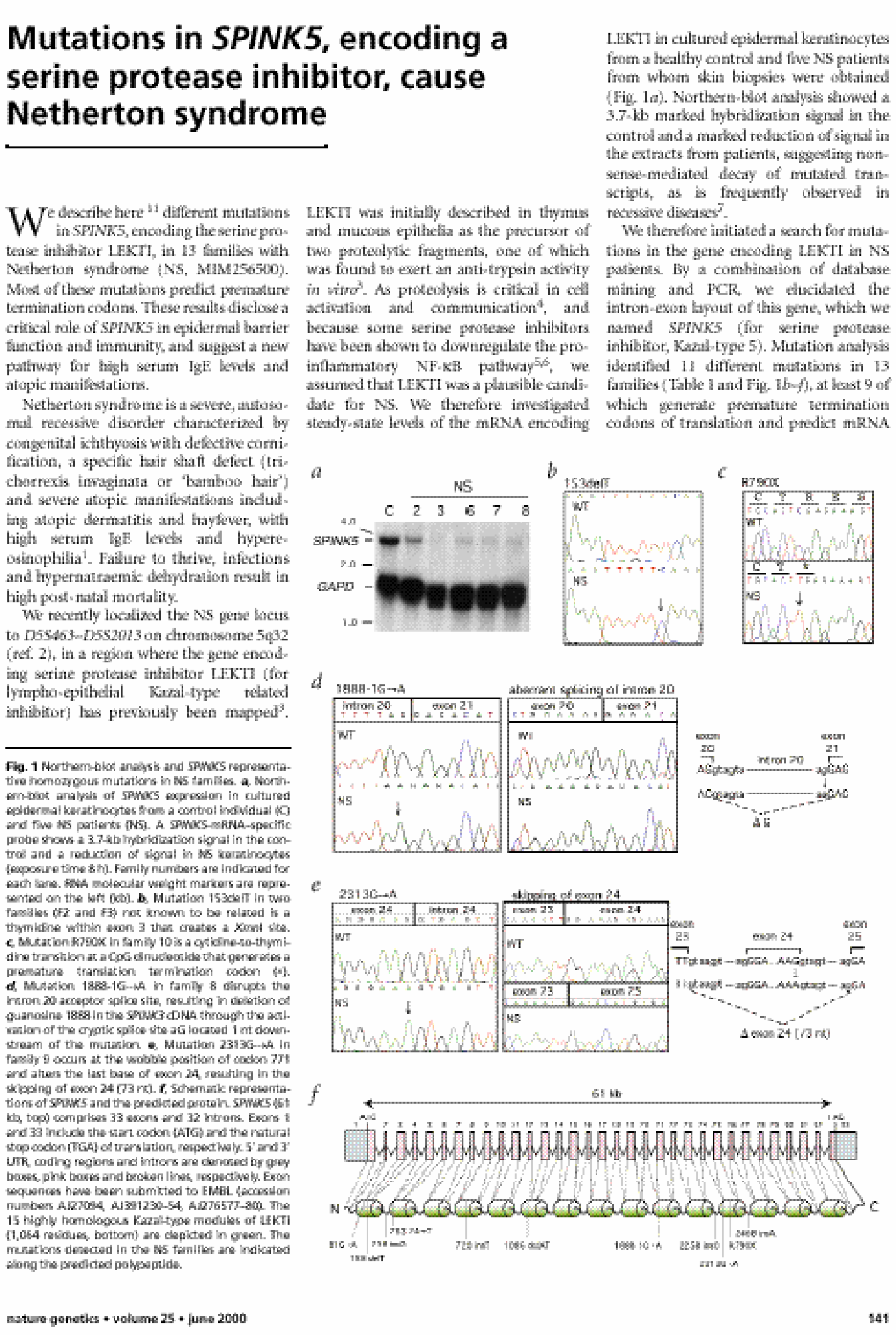


Chavanas et al., *Am J Hum Genet*, 66:914-921, 2000

# Netherton Syndrome Haplotypes



Chavanas et al., *Am J Hum Genet*, 66:914-921, 2000



# Mutations in *SPINK5*, encoding a serine protease inhibitor, cause Netherton syndrome

We describe here 11 different mutations in *SPINK5*, encoding the serine protease inhibitor LEKTI, in 13 families with Netherton syndrome (NS, MIM256500). Most of these mutations predict premature termination codons. These results disclose a critical role of *SPINK5* in epidermal barrier function and immunity, and suggest a new pathway for high serum IgE levels and atopic manifestations.

Netherton syndrome is a severe, autosomal recessive disorder characterized by congenital ichthyosis with defective cornification, a specific hair shaft defect (trichorrexis invaginata or 'bamboo hair') and severe atopic manifestations including atopic dermatitis and hayfever, with high serum IgE levels and hypereosinophilia[1]. Failure to thrive, infections and hypernatraemic dehydration result in high post-natal mortality.

We recently localized the NS gene locus to *D5S463–D5S2013* on chromosome 5q32 (ref. 2), in a region where the gene encoding serine protease inhibitor LEKTI (for lympho-epithelial Kazal-type related inhibitor) has previously been mapped[3]. LEKTI was initially described in thymus and mucous epithelia as the precursor of two proteolytic fragments, one of which was found to exert an anti-trypsin activity *in vitro*[3]. As proteolysis is critical in cell activation and communication[4], and because some serine protease inhibitors have been shown to downregulate the pro-inflammatory NF-κB pathway[5,6], we assumed that LEKTI was a plausible candidate for NS. We therefore investigated steady-state levels of the mRNA encoding LEKTI in cultured epidermal keratinocytes from a healthy control and five NS patients from whom skin biopsies were obtained (Fig. 1*a*). Northern-blot analysis showed a 3.7-kb marked hybridization signal in the control and a marked reduction of signal in the extracts from patients, suggesting nonsense-mediated decay of mutated transcripts, as is frequently observed in recessive diseases[7].

We therefore initiated a search for mutations in the gene encoding LEKTI in NS patients. By a combination of database mining and PCR, we elucidated the intron-exon layout of this gene, which we named *SPINK5* (for serine protease inhibitor, Kazal-type 5). Mutation analysis identified 11 different mutations in 13 families (Table 1 and Fig. 1*b–f*), at least 9 of which generate premature termination codons of translation and predict mRNA

**Fig. 1** Northern-blot analysis and *SPINK5* representative homozygous mutations in NS families. **a**, Northern-blot analysis of *SPINK5* expression in cultured epidermal keratinocytes from a control individual (C) and five NS patients (NS). A *SPINK5*-mRNA–specific probe shows a 3.7-kb hybridization signal in the control and a reduction of signal in NS keratinocytes (exposure time 8 h). Family numbers are indicated for each lane. RNA molecular weight markers are represented on the left (kb). **b**, Mutation 153delT in two families (F2 and F3) not known to be related is a thymidine within exon 3 that creates a *Xmn*I site. **c**, Mutation R790X in family 10 is a cytidine-to-thymidine transition at a CpG dinucleotide that generates a premature translation termination codon (*). **d**, Mutation 1888-1G→A in family 8 disrupts the intron 20 acceptor splice site, resulting in deletion of guanosine 1888 in the *SPINK5* cDNA through the activation of the cryptic splice site aG located 1 nt downstream of the mutation. **e**, Mutation 2313G→A in family 9 occurs at the wobble position of codon 771 and alters the last base of exon 24, resulting in the skipping of exon 24 (73 nt). **f**, Schematic representations of *SPINK5* and the predicted protein. *SPINK5* (61 kb, top) comprises 33 exons and 32 introns. Exons 1 and 33 include the start codon (ATG) and the natural stop codon (TGA) of translation, respectively. 5′ and 3′ UTR, coding regions and introns are denoted by grey boxes, pink boxes and broken lines, respectively. Exon sequences have been submitted to EMBL (accession numbers AJ27094, AJ391230–54, AJ276577–80). The 15 highly homologous Kazal-type modules of LEKTI (1,064 residues, bottom) are depicted in green. The mutations detected in the NS families are indicated along the predicted polypeptide.

Chavanas et al. 2000, *Nature Genetics*

# Multifactorial Traits (aka “Complex Disease”)

- Caused by > 1 gene
- Possibly triggered by environment
- Each gene (env) may have small effect
- No clear segregation pattern in families
- Epistasis or intra-genic interactions likely
- Pleiotropy, environmental influences, G x E interactions likely
- Epigenetic influences possible
- Measurement of disease or phenotype not highly reliable

# Assessing genetic contributions to complex traits

- Continuous characters (wt, blood pressure)
  - Heritability: Proportion of observed variance in phenotype explained by genetic factors

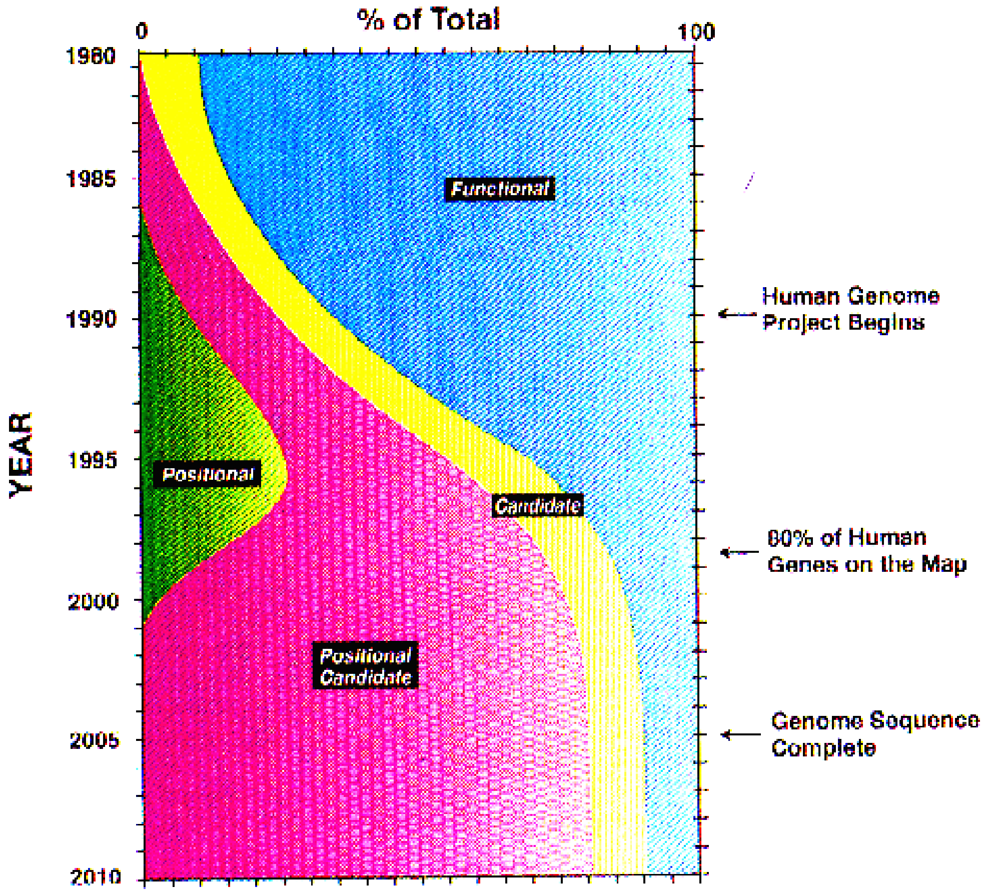
- Discrete characters (disease)
  - Relative risk ratio: $\lambda$ = risk to relative of an affected individual/risk in general population
  - $\lambda$ encompasses all genetic and environmental effects, not just those due to any single locus

# λs examples

| | |
|---|---|
| Huntington’s Disease | >1000 |
| Cystic Fibrosis | 400 |
| Autism | 75 |
| Inflammatory Bowel Disease | 60 |
| Multiple Sclerosis | 20 |
| Juvenile Diabetes | 15 |
| Schizophrenia | 10 |
| Asthma | 6 |
| Prostate Cancer | 5 |
| Late Onset Diabetes | 2-3 |
| Breast Cancer | 2 |

NB: all are crude estimates as different sampling strategies give different values

# Cloning Predictions 1995



Collins, F.S. Positional cloning moves from perditional to traditional, *Nat Genet*, 9:347-350, 1995

# Genome Screens in Complex Traits

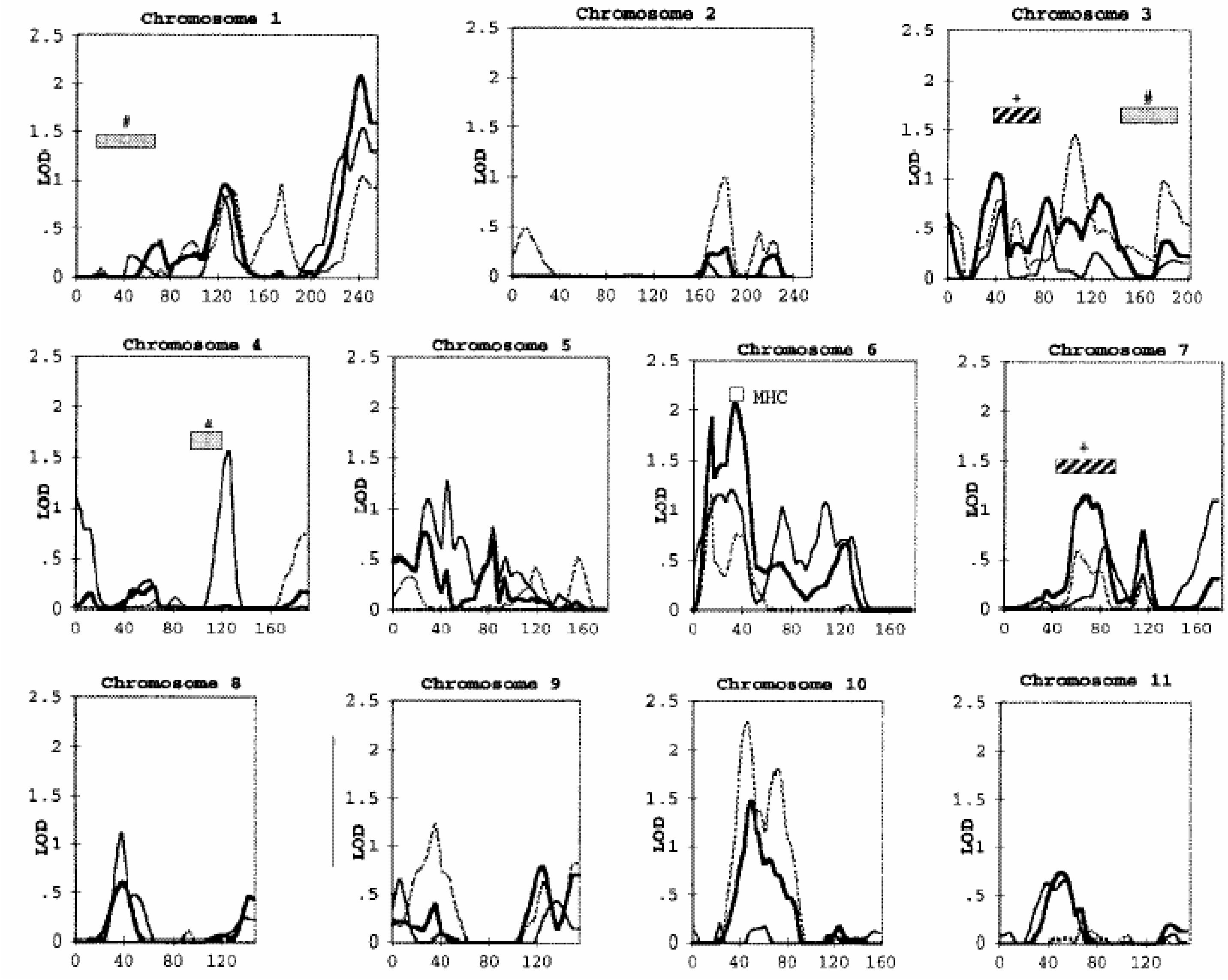
## 1997/98

- Diabetes (IDDM + NIDDM)
- Asthma/atopy
- Osteoporosis
- Obesity
- Multiple Sclerosis
- Rheumatoid arthritis
- Systemic lupus erythematosus
- Ankylosing spondylitis
- Epilepsy
- Inflammatory Bowel Disease
- Celiac Disease
- Psychiatric Disorders (incl. Scz, bipolar)
- Behavioral traits (incl. Personality, panic)
- others missed...

## 1999

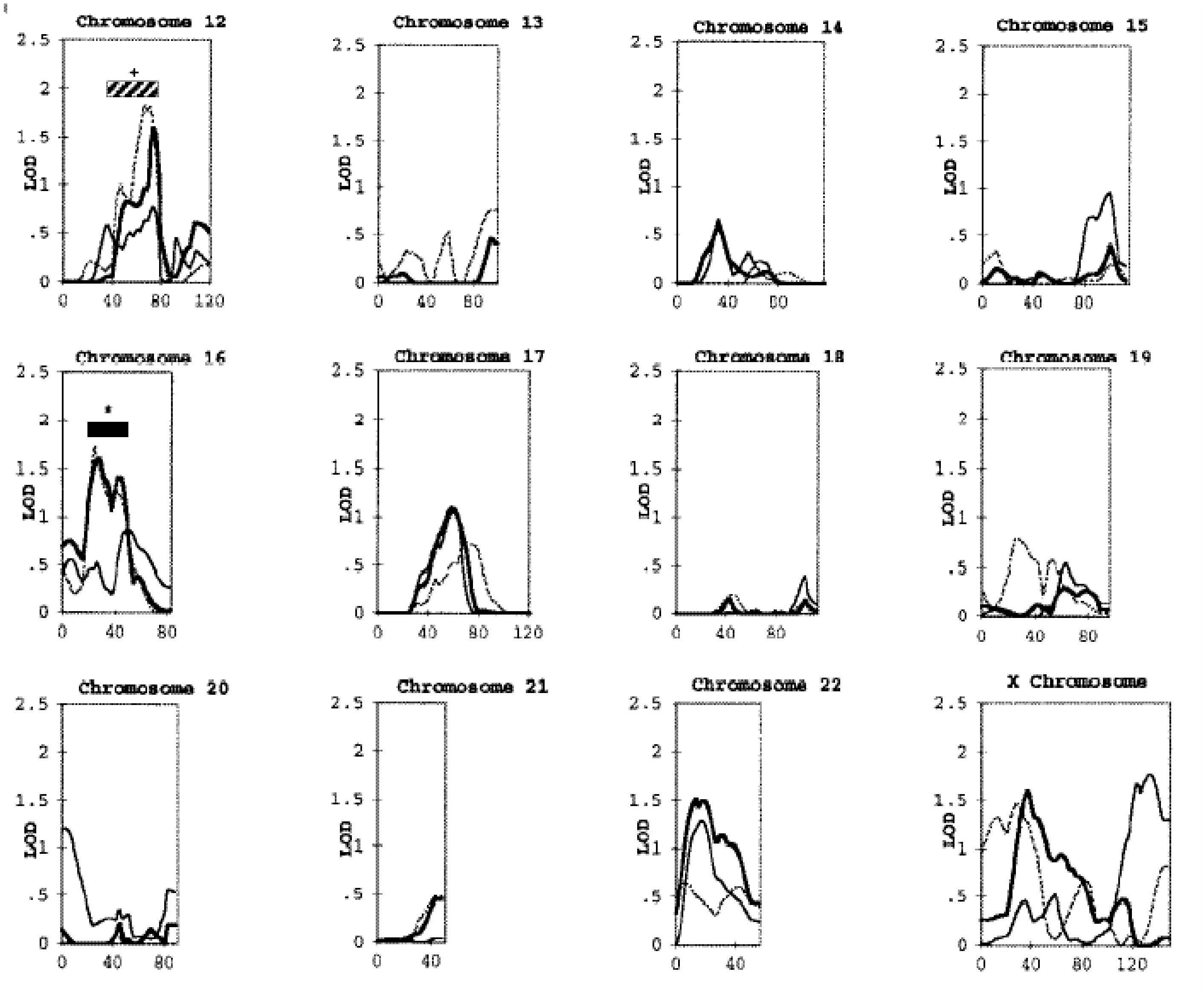
- NIDDM
- Asthma/atopy
- Psoriasis
- Inflammatory Bowel Disease
- Osteoporosis/Bone Mineral Density
- Obesity
- Epilepsy
- Thyroid disease
- Pre-eclampsia
- Blood pressure
- Psychiatric disorders (incl. Scz, bipolar)
- Behavioral traits (incl. smoking, alcoholism, autism)
- Familial combined hyperlipidemia
- Tourette syndrome
- Systemic lupus erythematosus
- others missed…

# Inflammatory Bowel Disease Genome Screen



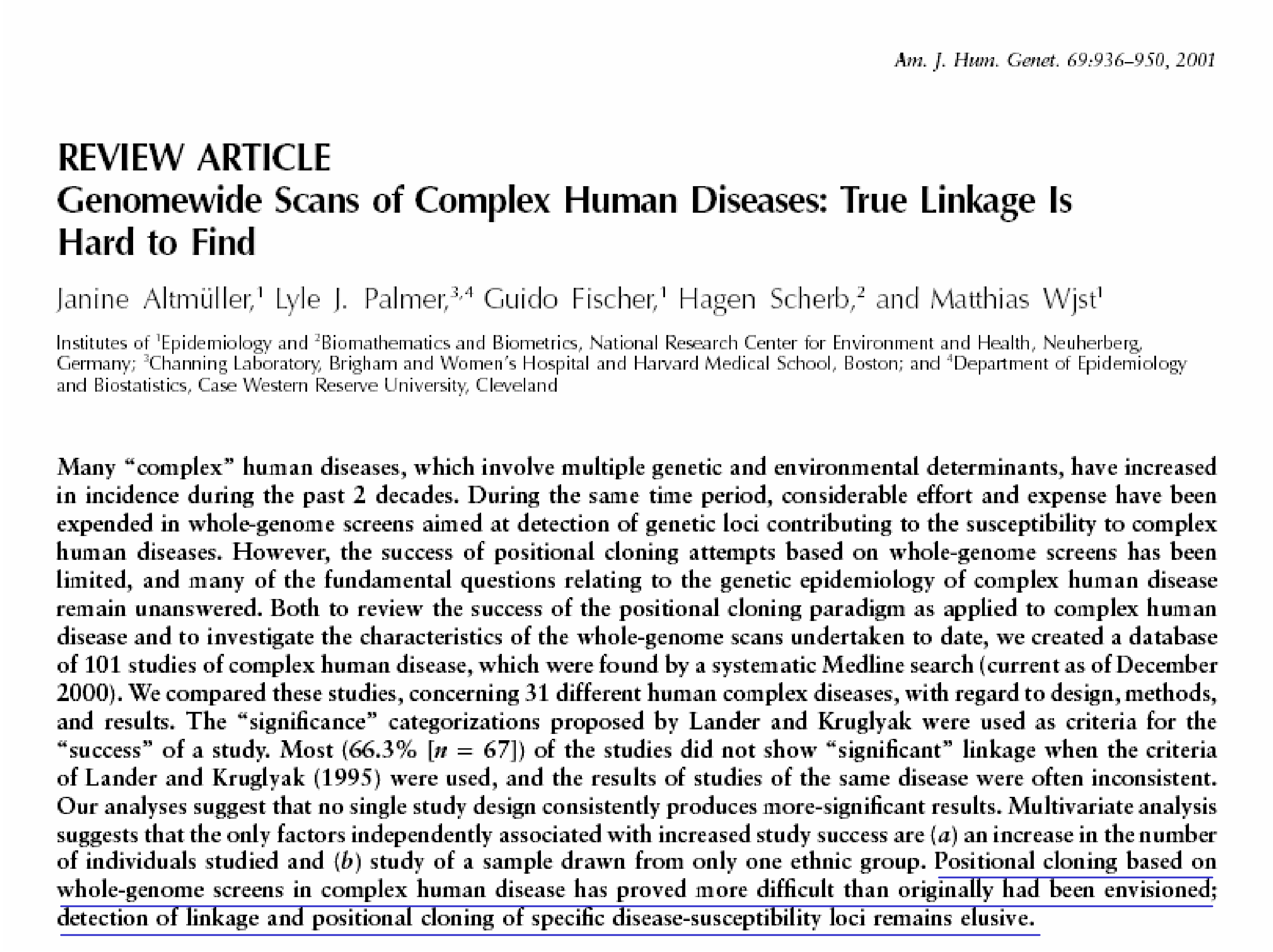
Hampe et al., *Am J Hum Genet*, 64:808-816, 1999

# Inflammatory Bowel Disease Genome Screen



Hampe et al., *Am J Hum Genet*, 64:808-816, 1999

# Linkage Outcomes for Complex Traits



Am. J. Hum. Genet. 69:936–950, 2001

REVIEW ARTICLE

## Genomewide Scans of Complex Human Diseases: True Linkage Is Hard to Find

Janine Altmüller,[1] Lyle J. Palmer,[3,4] Guido Fischer,[1] Hagen Scherb,[2] and Matthias Wjst[1]

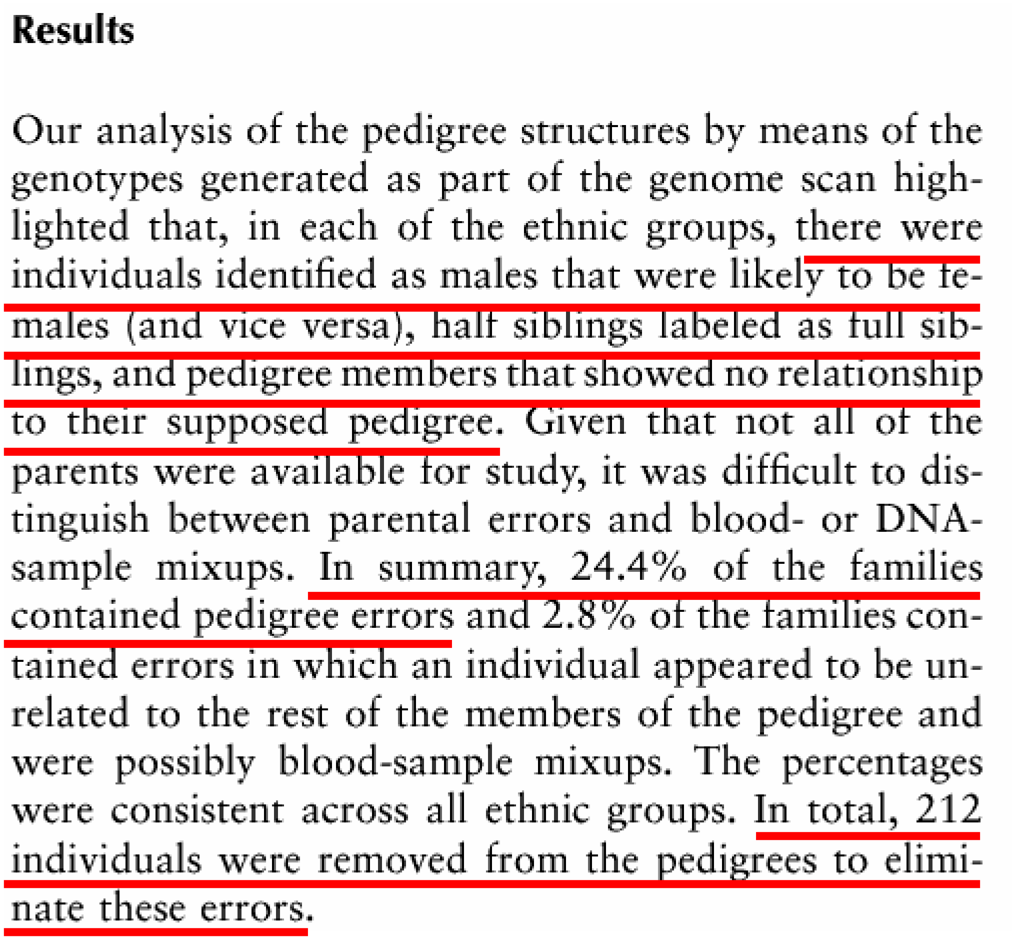
Institutes of [1]Epidemiology and [2]Biomathematics and Biometrics, National Research Center for Environment and Health, Neuherberg, Germany; [3]Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, Boston; and [4]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland

Many "complex" human diseases, which involve multiple genetic and environmental determinants, have increased in incidence during the past 2 decades. During the same time period, considerable effort and expense have been expended in whole-genome screens aimed at detection of genetic loci contributing to the susceptibility to complex human diseases. However, the success of positional cloning attempts based on whole-genome screens has been limited, and many of the fundamental questions relating to the genetic epidemiology of complex human disease remain unanswered. Both to review the success of the positional cloning paradigm as applied to complex human disease and to investigate the characteristics of the whole-genome scans undertaken to date, we created a database of 101 studies of complex human disease, which were found by a systematic Medline search (current as of December 2000). We compared these studies, concerning 31 different human complex diseases, with regard to design, methods, and results. The "significance" categorizations proposed by Lander and Kruglyak were used as criteria for the "success" of a study. Most (66.3% [$n$ = 67]) of the studies did not show "significant" linkage when the criteria of Lander and Kruglyak (1995) were used, and the results of studies of the same disease were often inconsistent. Our analyses suggest that no single study design consistently produces more-significant results. Multivariate analysis suggests that the only factors independently associated with increased study success are (*a*) an increase in the number of individuals studied and (*b*) study of a sample drawn from only one ethnic group. Positional cloning based on whole-genome screens in complex human disease has proved more difficult than originally had been envisioned; detection of linkage and positional cloning of specific disease-susceptibility loci remains elusive.

# Why such limited success with complex trait linkage studies?

- Power
  - Power calculations have always indicated need for many 100’s, probably thousands of families to detect genes of even moderate effect
  - N ~ 200 for most studies conducted to date
    - For QTL, this is about enough to detect a locus explaining 25% of the total variance in the trait
- Hope for ‘low-hanging’ fruit
  - If there are one or a few monogenic-like loci within oligogenic spectrum, could lead to pathway information
  - Not supported by data.
- Practical problems:  errors in data

# Pedigree Errors

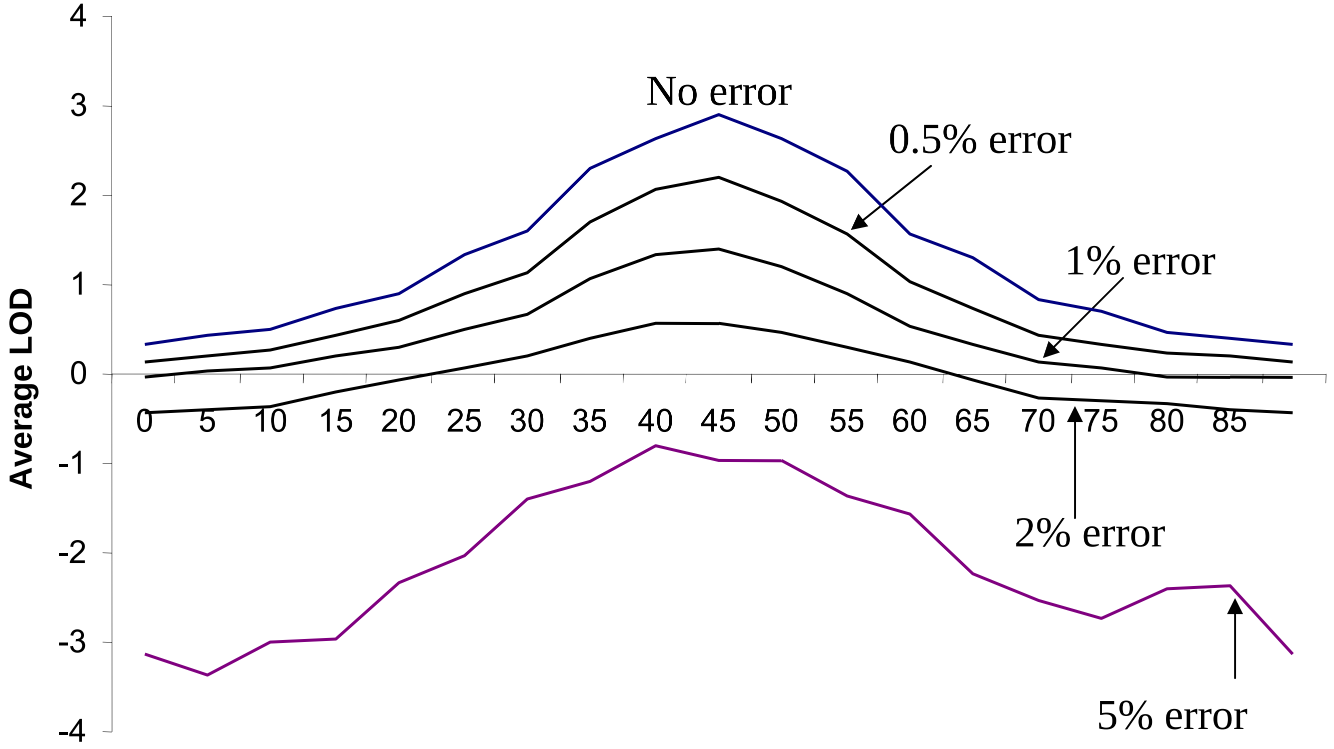


**Results**

Our analysis of the pedigree structures by means of the genotypes generated as part of the genome scan highlighted that, in each of the ethnic groups, there were individuals identified as males that were likely to be females (and vice versa), half siblings labeled as full siblings, and pedigree members that showed no relationship to their supposed pedigree. Given that not all of the parents were available for study, it was difficult to distinguish between parental errors and blood- or DNA-sample mixups. In summary, 24.4% of the families contained pedigree errors and 2.8% of the families contained errors in which an individual appeared to be unrelated to the rest of the members of the pedigree and were possibly blood-sample mixups. The percentages were consistent across all ethnic groups. In total, 212 individuals were removed from the pedigrees to eliminate these errors.

Excerpt from *Am J Hum Genet*, 2000

# Genotyping Error: Affected Sib Pair Sample



$\lambda_s$ = 1.5; Lods calculated using Kong & Cox (signed) procedure

# Genotype Error

- Realistic error rates in past linkage studies probably ~1-3%
- Small error rates can have dramatic consequences
  - 1% error costs 50% of test statistic in ASP linkage
- Detection more important than correction (probably)
- Detection without families hard problem (esp for association)
- These are (partly) avoidable problems by rigorous study design

# Positional Cloning of Complex Traits: Lack of Success

*...Not surprisingly, progress in analyzing complex genetic disorders has been more modest. What success there has been has basically come from one of two approaches:*

*(i) Identification of a sub-phenotype in pedigrees...*
(akin to Mendelian disorder)
*(ii) Genetic studies in isolated human populations*
(reduced genetic variation)

(Collins et al, Science, 278:1580-81, 1997)

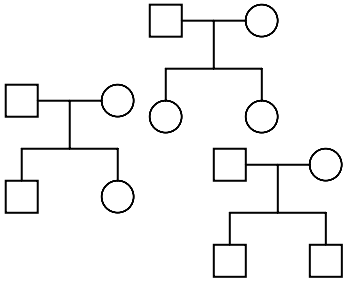
**This has not improved in past 8 years…**

Weiss & Terwilliger (2000), Altschuler et al (2000), others

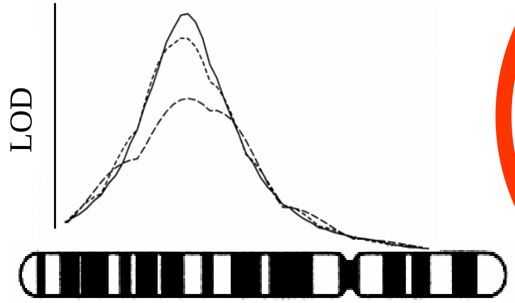
# The weakest link?

## Genetics

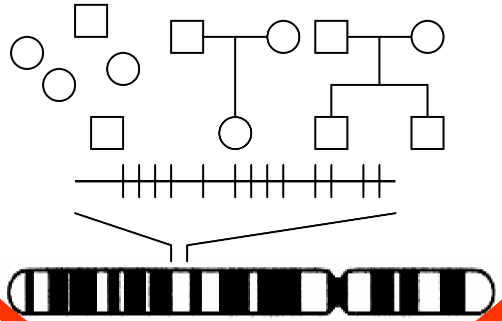
Sib pairs



Chromosome Region

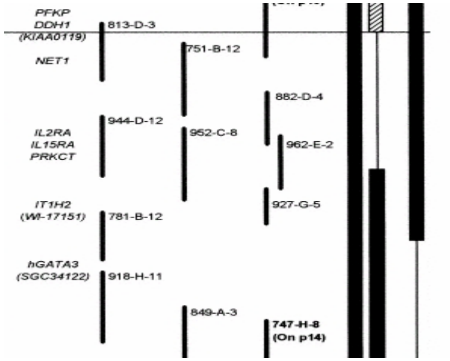


Association Study

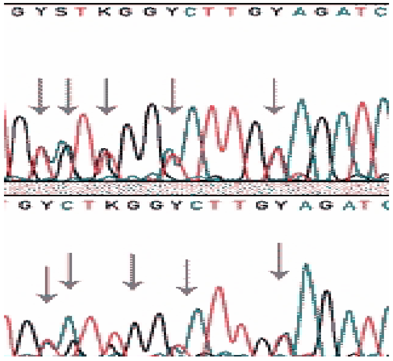


## Genomics

Physical Mapping/
Sequencing



Candidate Gene Selection/
Polymorphism Detection



Mutation Characterization/
Functional Annotation

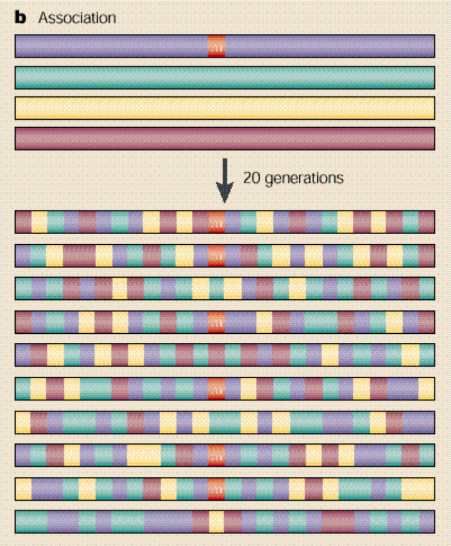
```
                                   +C
                                   +C
                         -G
                                   -C
..... GAG GGG GGC ACC CCC CCC ATG GAT .....
      Glu Gly Gly Thr Pro Pro Met Asp .....
              321 322
```

# Association Analysis

- Simple genetic basis

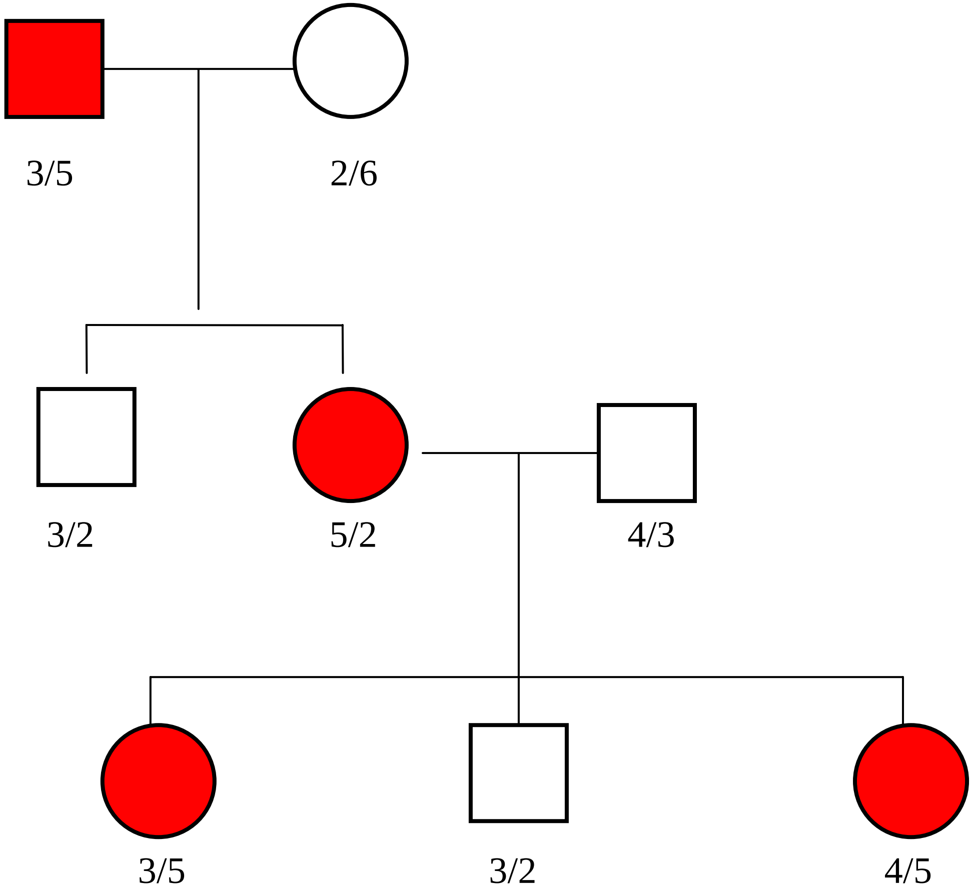
Short unit of resemblance
Population-specific

- One of easiest genetic study designs



Correlate allele frequencies with traits/diseases
At core of monogenic & oligo/polygenic trait models

# Linkage: Allelic association WITHIN FAMILIES



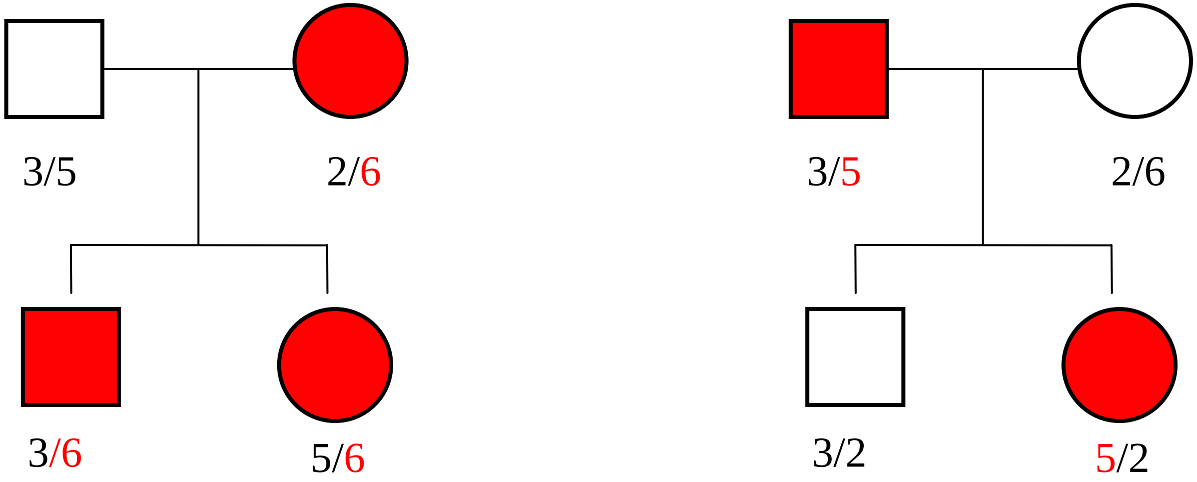
🟥 affected

○

Allele coded by CA copies
2 = CACA
6 = CACACACACACA

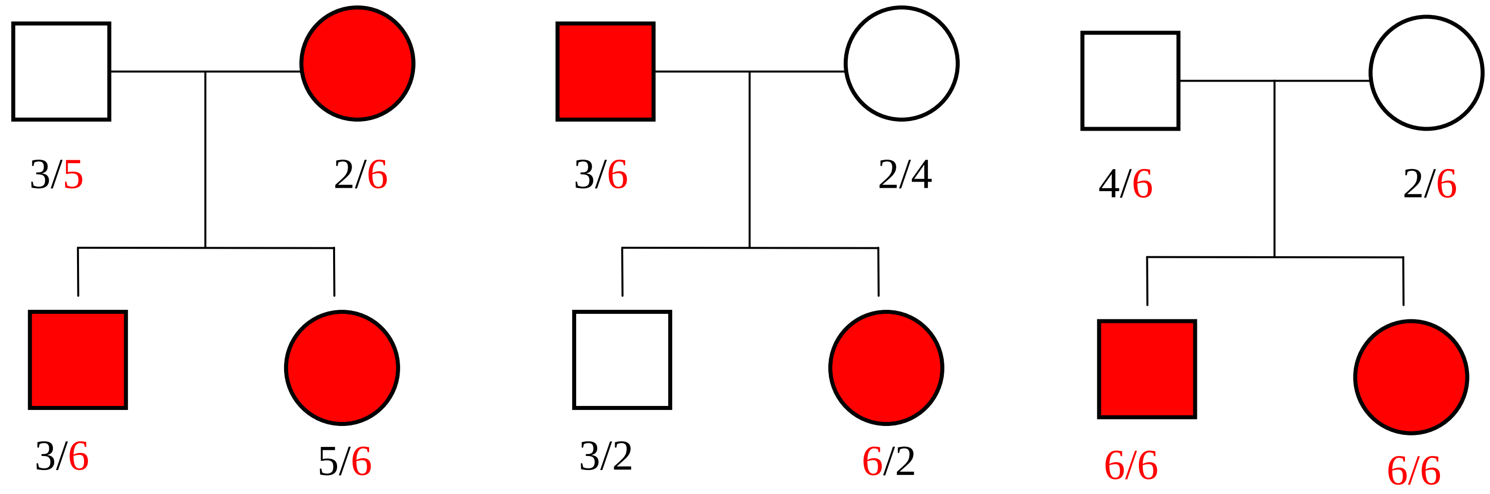
Disease linked to ‘5’ allele in dominant inheritance

# *Allelic Association:*
# *Extension of linkage to the population*



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

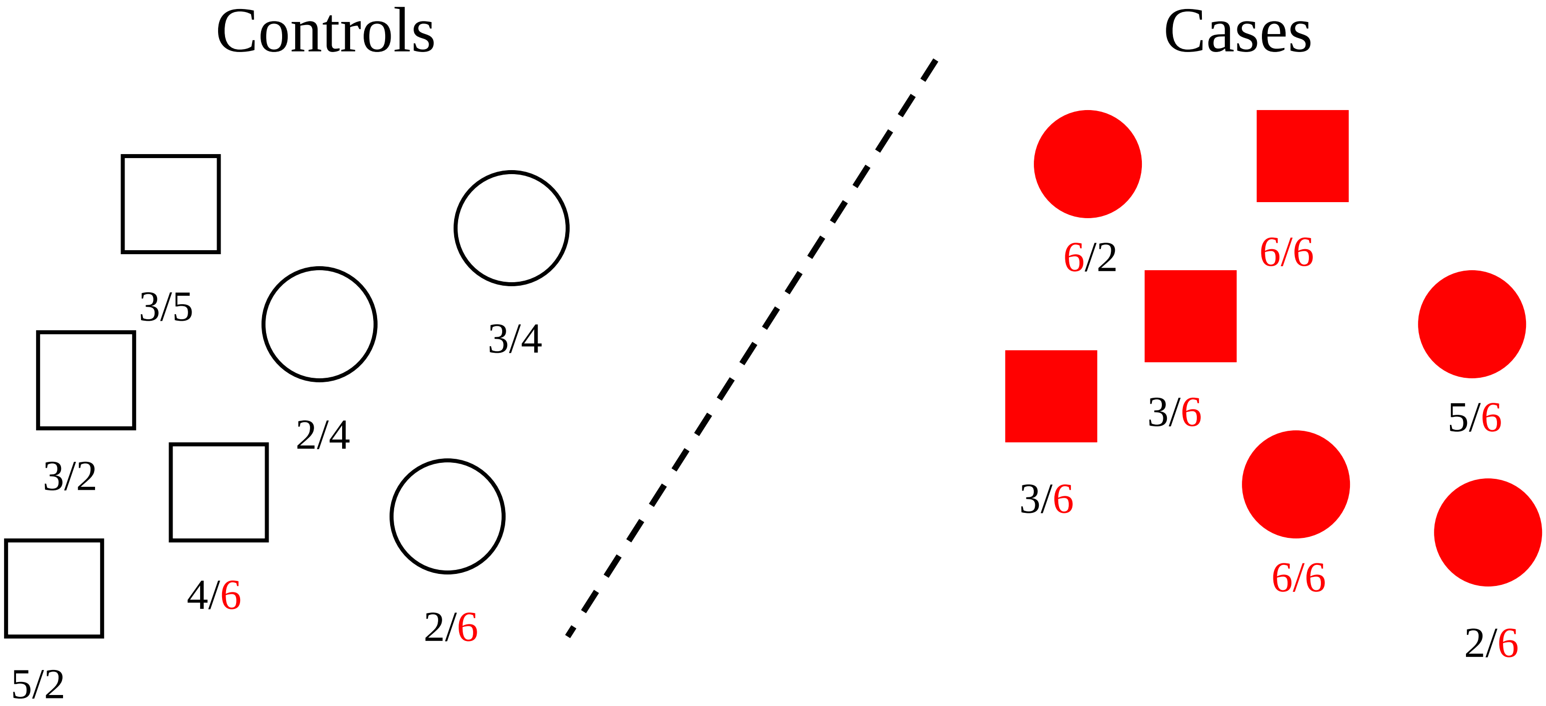
# Allelic Association
# Extension of linkage to the population



*All* families are ‘linked’ with the marker

Allele 6 is ‘associated’ with disease

# Allelic Association



Allele 6 is ‘associated’ with disease

# Association – identical ancestral origin

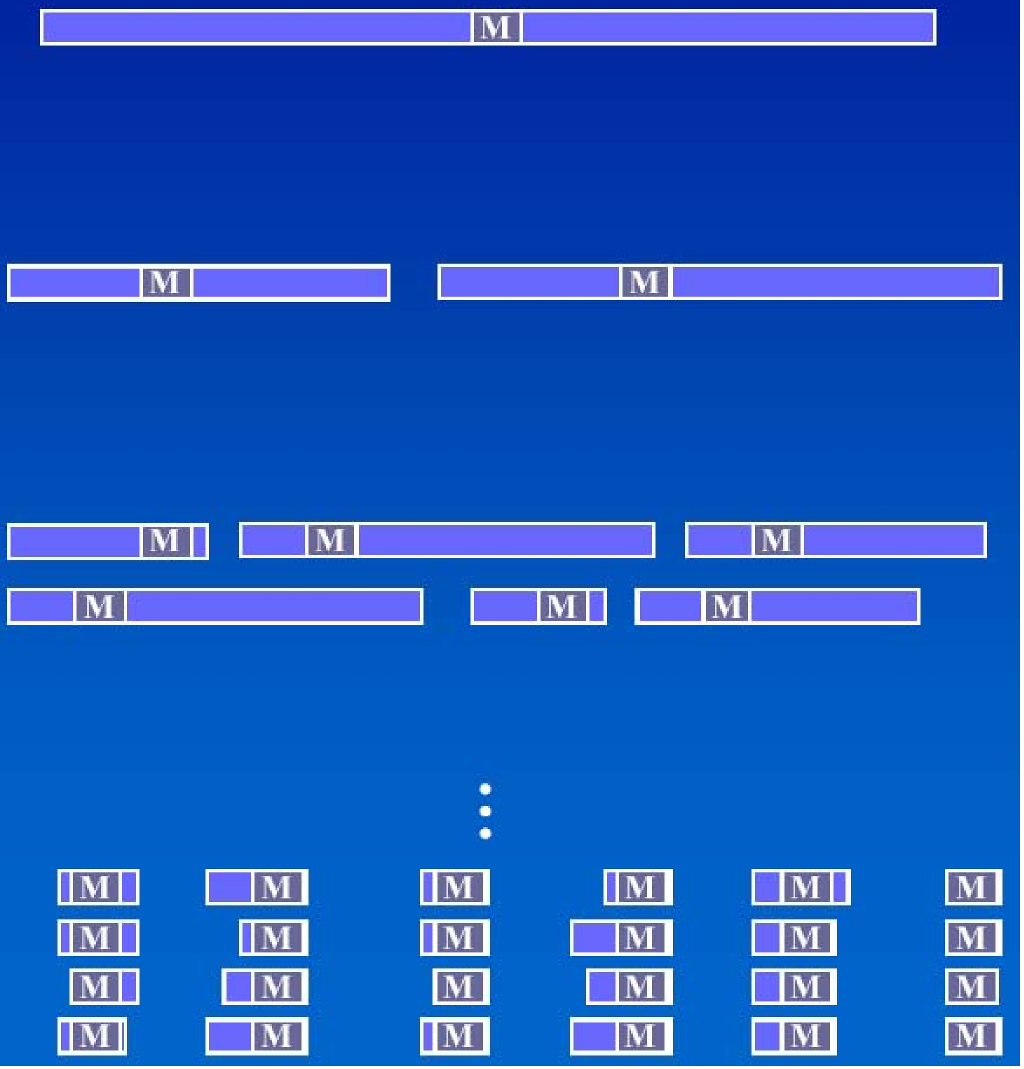
**Generation I - a disease-causing mutation occurs on a chromosome**

**Generation II - about 50% of the children receive the mutation and a surrounding chromosomal segment from the mutated founder**

**Generation III - the lengths of the segments originating from the mutated founder chromosome are shorter than or equal to those in GII.**

⋮

**Generation n - very short segments around the mutated locus conserved**



# Linkage vs Association

### Linkage

1. Requires families
2. Matching/ethnicity generally unimportant
3. Few markers for genome coverage (300-400 STRs)
4. Allele-sharing weak design
5. Yields coarse location
6. Good for initial detection; poor for fine-mapping
7. Powerful for rare variants

### Association

1. Families or unrelateds
2. Matching/ethnicity important
3. Many markers for genome coverage ($10^5$ – $10^6$ SNPs)
4. Powerful design based on means
5. Yields fine-scale location
6. Good for fine-mapping, poor for initial detection
7. Powerful for common variants; rare variants generally impossible

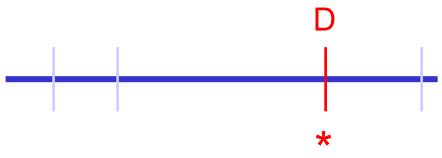
# *Allelic Association*

## *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

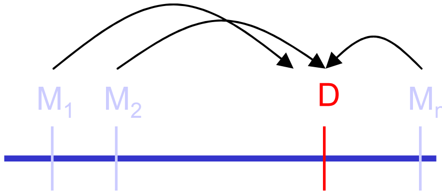
# Indirect and Direct Allelic Association

## 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

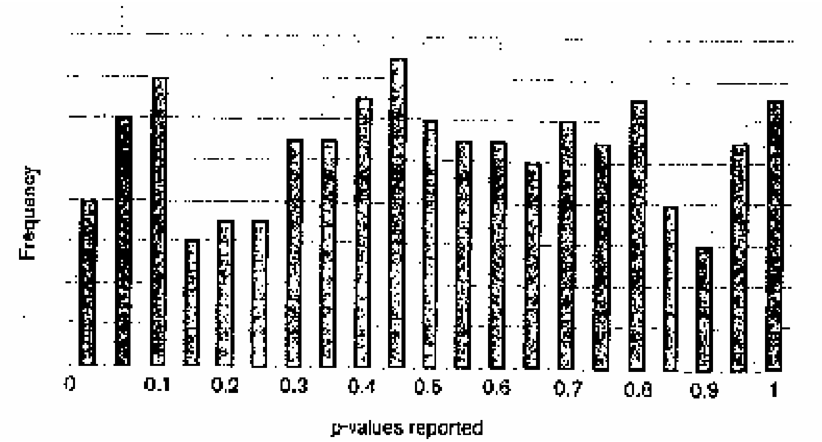
Allelic Association: correlation between marker allele and trait

# How many association studies have been conducted?

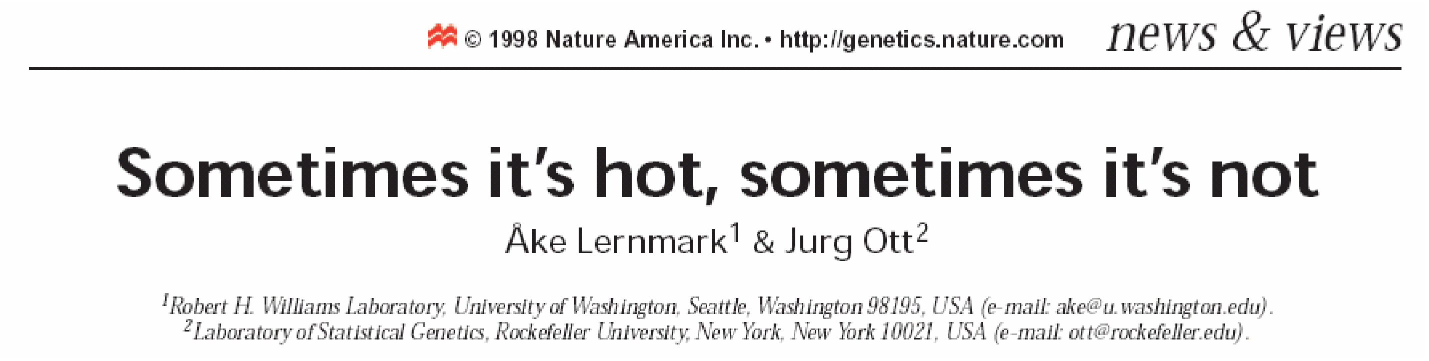
- Pubmed: 1 Mar 2004. “Genetic association” gives 23,467 hits
- > 10% hits in HLA alone
- Probably ~ 20 confirmed associations for complex traits

# Association Study Outcomes

## Reported p-values from association studies in *Am J Med Genet* or *Psychiatric Genet* 1997



Terwilliger & Weiss, *Curr Opin Biotech*, 9:578-594, 1998

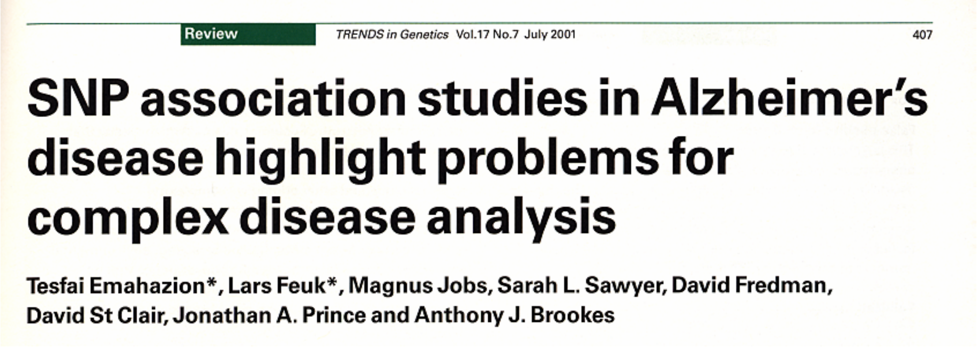


© 1998 Nature America Inc. • http://genetics.nature.com *news & views*

## Sometimes it's hot, sometimes it's not

Åke Lernmark[1] & Jurg Ott[2]

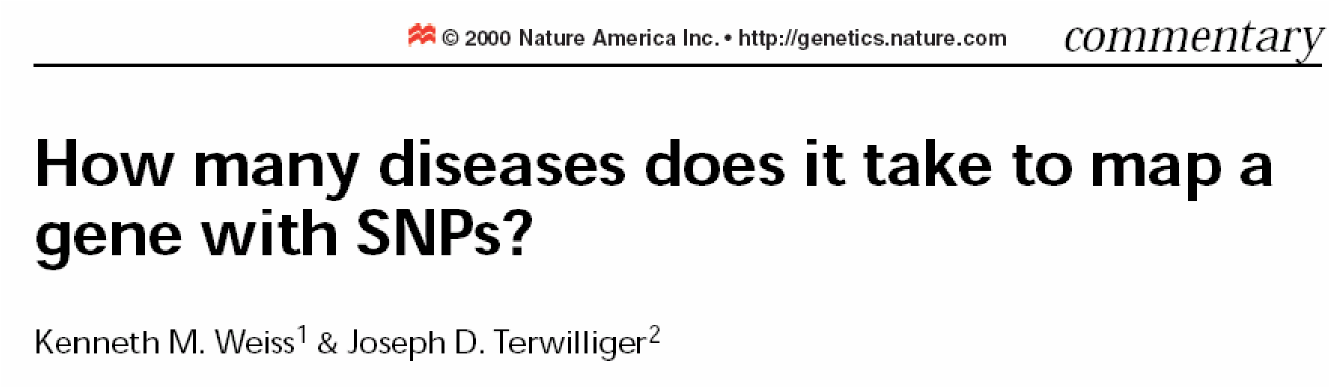
*[1]Robert H. Williams Laboratory, University of Washington, Seattle, Washington 98195, USA (e-mail: ake@u.washington.edu).*
*[2]Laboratory of Statistical Genetics, Rockefeller University, New York, New York 10021, USA (e-mail: ott@rockefeller.edu).*



Review *TRENDS in Genetics* Vol.17 No.7 July 2001 407

## SNP association studies in Alzheimer's disease highlight problems for complex disease analysis

**Tesfai Emahazion*, Lars Feuk*, Magnus Jobs, Sarah L. Sawyer, David Fredman, David St Clair, Jonathan A. Prince and Anthony J. Brookes**



© 2000 Nature America Inc. • http://genetics.nature.com *commentary*

## How many diseases does it take to map a gene with SNPs?

Kenneth M. Weiss[1] & Joseph D. Terwilliger[2]

# Why limited success with association studies?

1. Small sample sizes → results overinterpreted

2. Phenotypes are complex.  Candidate genes difficult to choose

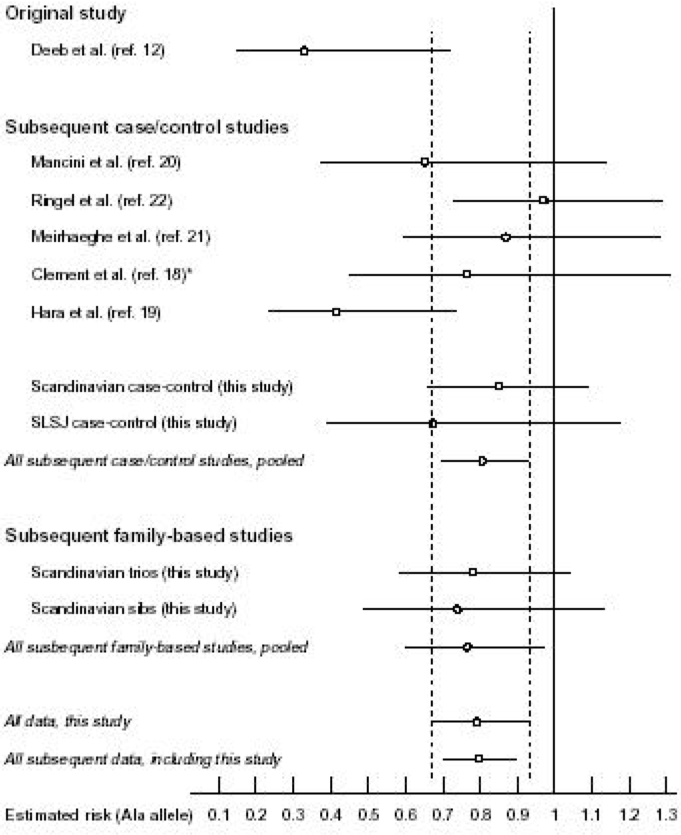
3. Allelic/genotypic contributions are complex.  Even true associations difficult to see.

4. Background patterns of LD are unknown.  Difficult to appreciate signal when can’t assess noise.

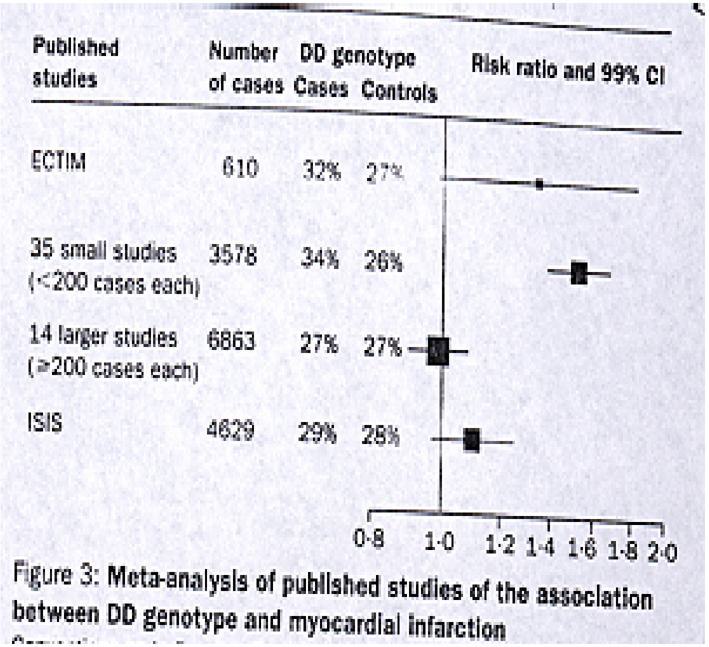
5. Population stratification has led clouded true/false positives

# Sample Size Matters

## PPARγ and NIDDM



## ACE and MI



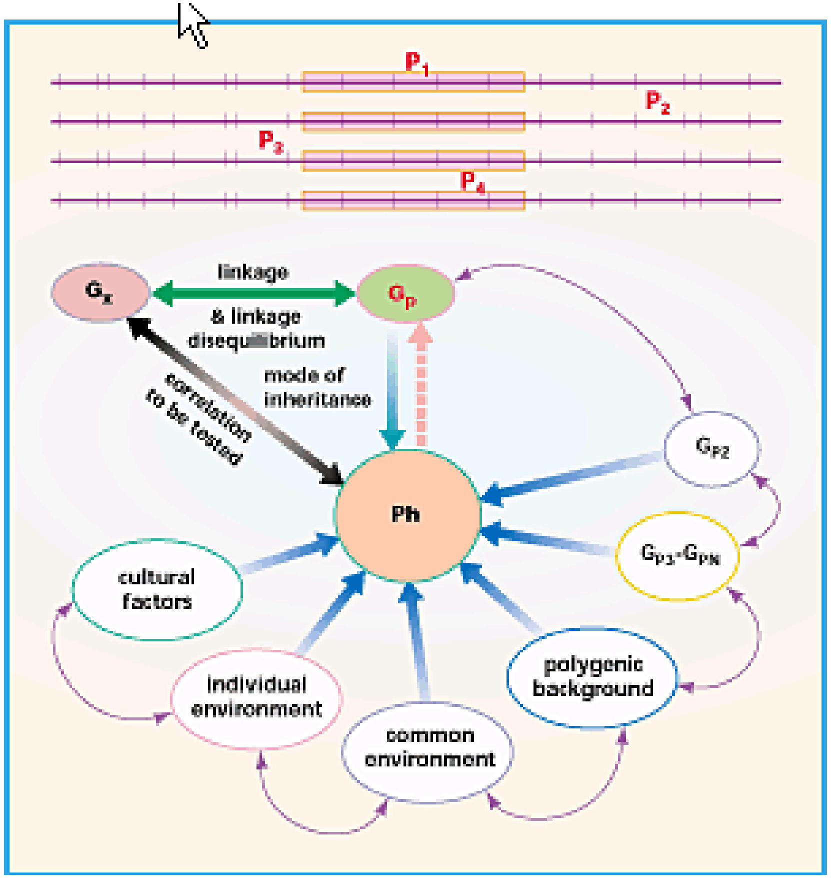
| Published studies | Number of cases | DD genotype Cases | DD genotype Controls |
|---|---|---|---|
| ECTIM | 610 | 32% | 27% |
| 35 small studies (<200 cases each) | 3578 | 34% | 26% |
| 14 larger studies (≥200 cases each) | 6863 | 27% | 27% |
| ISIS | 4629 | 29% | 28% |

Figure 3: Meta-analysis of published studies of the association between DD genotype and myocardial infarction

Altshuler et al *Nat Genet* 2000

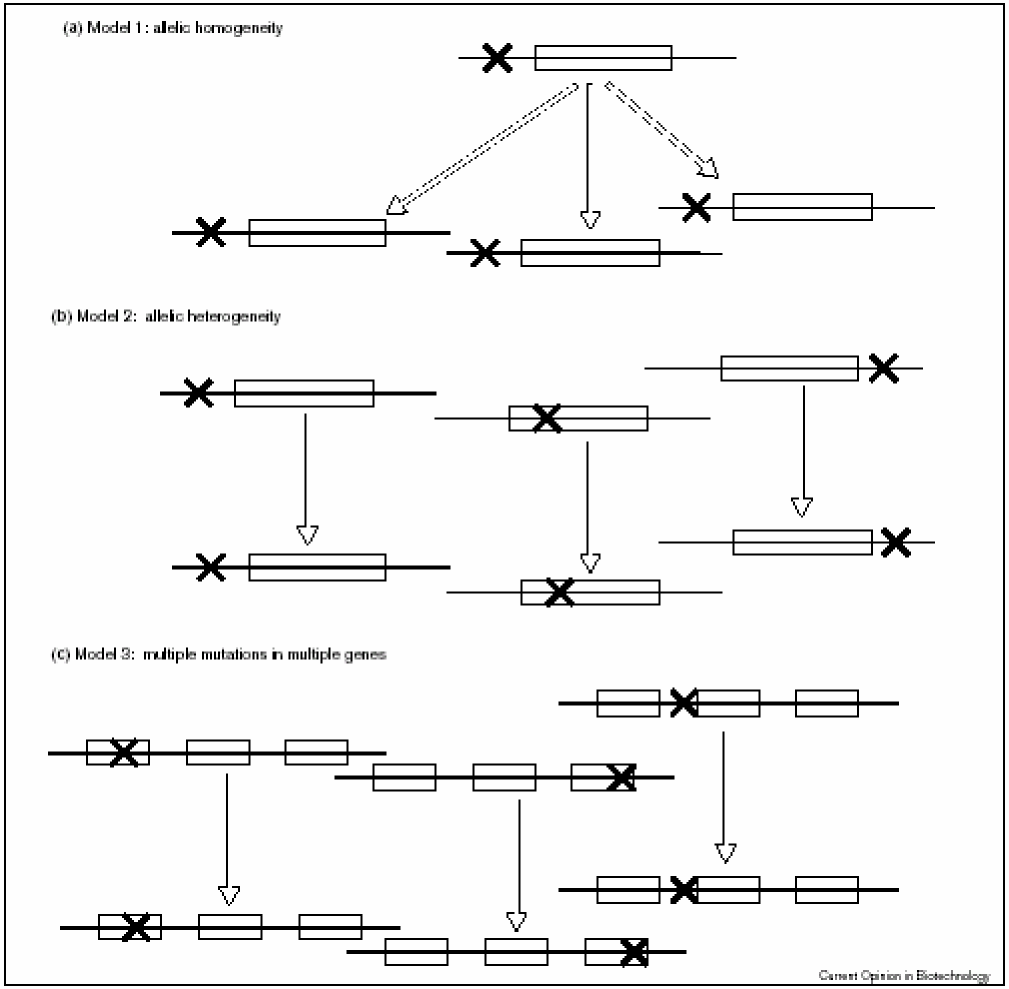
Keavney et al *Lancet* 2000

# Phenotypic Complexity



Weiss & Terwilliger, Nat Genet, 2000

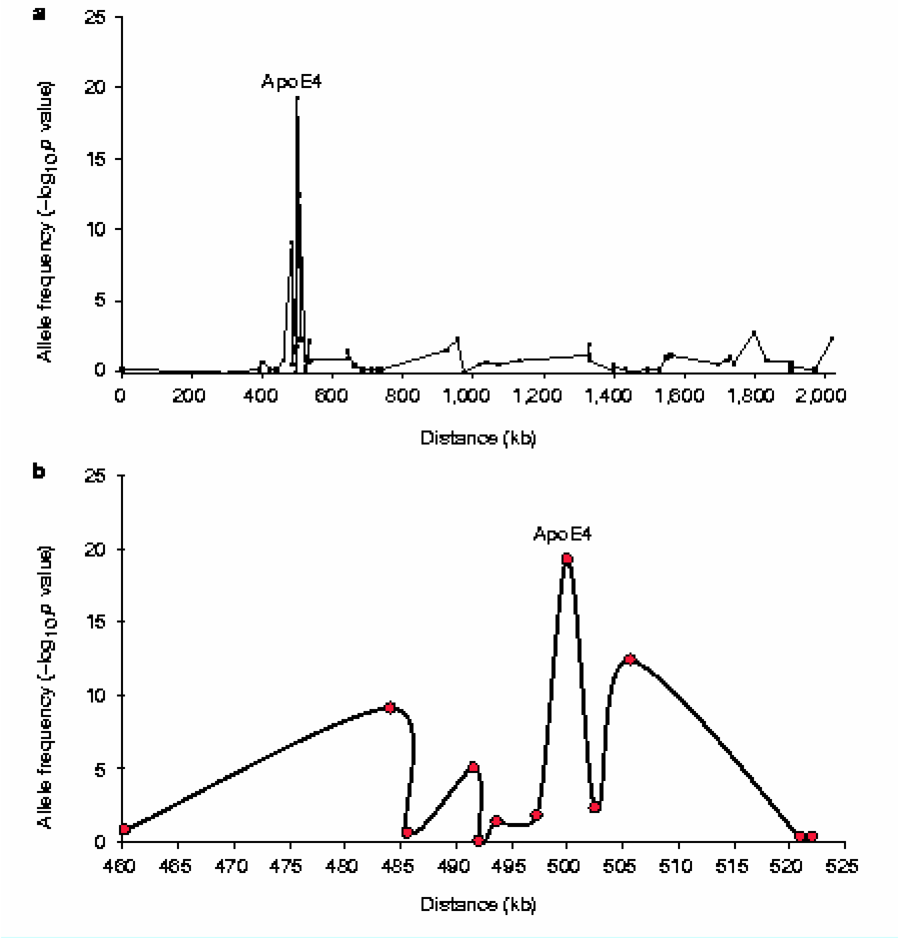
# Heterogeneity


(a) Model 1 : allelic homogeneity
(b) Model 2: allelic heterogeneity
(c) Model 3: multiple mutations in multiple genes
Current Opinion in Biotechnology

Three simple models for the allelic complexity of genetic disease are shown. (a) In Model 1, all disease-predisposing alleles at a given locus are identical by descent in the population – having derived from some common ancestor. In this situation, there is expected to be a conserved haplotype around the disease allele, which is shared by all carriers in the population many generations later. (b) Model 2 shows the case of allelic heterogeneity, in which multiple different allelic variants can each predispose to the phenotype. Thus among individuals with one of these 'D' alleles, there will be an assortment of haplotype backgrounds. The more heterogeneity, the less LD. (c) Model 3 shows the situation for multiple 'D' alleles in different genes. These genes may be linked (as shown) or unlinked.

Terwilliger & Weiss, Curr Opin Biotechnol, 1998

# Effects of linkage disequilibrium



Roses, *Nature* 2000

# Main Blame

*Why do association studies have such a spotted history in human genetics?*

**Blame: 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.

# Population Stratification

- Recent admixture of populations
- Requirements:
  - Group differences in allele frequency
  - Group differences in outcome
- Leads to spurious association

- In epidemiology, this is a classic matching problem, with genetics as a confounding variable

Most oft-cited reason for lack of association replication

# Population Stratification

- Consider two case/control samples, A and B, genotyped at a marker with alleles M and m

Sample ‘A’

|  | M | m | Freq. |
|---|---|---|---|
| Affected | 50 | 50 | .10 |
| Unaffected | 450 | 450 | .90 |
|  | .50 | .50 |  |

$\chi^2_1$ is n.s.

Sample ‘B’

| M | m | Freq. |
|---|---|---|
| 1 | 9 | .01 |
| 99 | 891 | .99 |
| .10 | .90 |  |

$\chi^2_1$ is n.s.

Neither has significant association

# Population Stratification

Sample ‘A’

|  | M | m | Freq. |
|---|---|---|---|
| Affected | 50 | 50 | .10 |
| Unaffected | 450 | 450 | .90 |
|  | .50 | .50 |  |

$\chi^2_1$ is n.s.

+

Sample ‘B’

|  | M | m | Freq. |
|---|---|---|---|
| Affected | 1 | 9 | .01 |
| Unaffected | 99 | 891 | .99 |
|  | .10 | .90 |  |

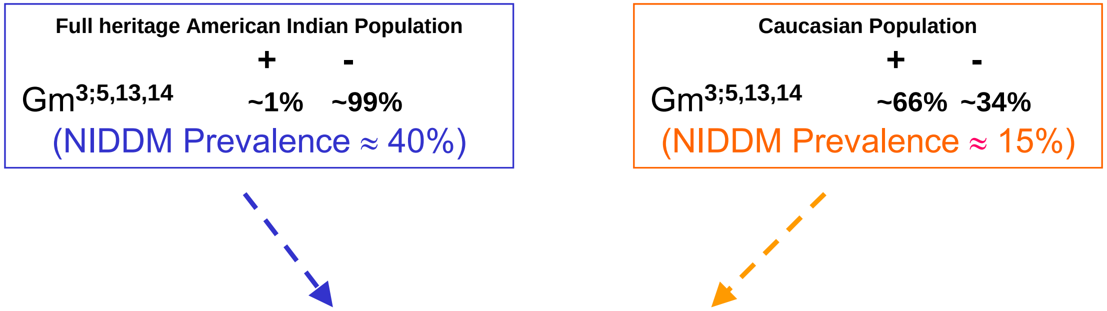
$\chi^2_1$ is n.s.

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

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

**Association induced by sample mixing**

# 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

| ***Index of Indian Heritage*** | **Gm$^{3;5,13,14}$ haplotype** |  |
|---|---|---|
|  | + | - |
| 0 | 17.8% | 19.9% |
| 4 | 28.3% | 28.8% |
| 8 | 35.9% | 39.3% |

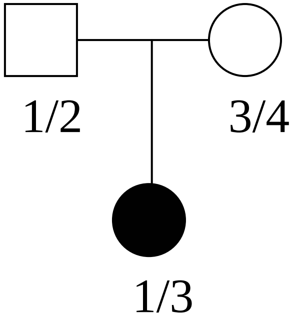
Reviewed in Cardon &  Palmer, Lancet 2003

# ‘Control’ Samples in Human Genetics $\leq$ 2000

- Because of fear of stratification, complex trait genetics turned away from case/control studies

  - *fear may be unfounded*

- Moved toward family-based controls (flavour is TDT: transmission/disequilibrium test)



“Case” = transmitted alleles

= 1 and 3

“Control” = untransmitted alleles

= 2 and 4

# TDT Advantages/Disadvantages

## Advantages

Robust to stratification
Genotyping error detectable via Mendelian inconsistencies
Estimates of haplotypes possible

## Disadvantages

Detection/elimination of genotyping errors causes bias (Gordon et al., 2001)
Uses only heterozygous parents
Inefficient for genotyping

3 individuals yield 2 founders: 1/3 information not used

Can be difficult/impossible to collect

Late-onset disorders, psychiatric conditions, pharmacogenetic applications

# Association studies < 2000: TDT

- TDT virtually ubiquitous over past decade

  Grant, manuscript referees & editors mandated design

- View of case/control association studies greatly diminished due to perceived role of stratification

## Association Studies ~ 2000:
## Return to population

- Case/controls, using extra genotyping
- Traditional trial design, augmented by genotyping

# Detecting and Controlling for
# Population Stratification with Genetic Markers

## Idea

- Take advantage of availability of large N genetic markers

- Use case/control design

- Genotype genetic markers across genome
  (Number depends on different factors)

- Look if any evidence for background population substructure exists and account for it

- Different approaches/different assumptions, models
  - GC (Devlin & Roeder, 1999)
  - Structured Association (Pritchard, Donnelly and others, 2000+)

# Why limited success with association studies?

1. Small sample sizes → results overinterpreted
2. Phenotypes are complex. Candidate genes difficult to choose
3. Allelic/genotypic contributions are complex. Even true associations difficult to see.
4. Background patterns of LD are unknown. Difficult to appreciate signal when can’t assess noise.
5. Population stratification has led to many false positives and misses

# Upcoming association studies have real promise

- Large, epidemiological-sized samples emerging
  - ISIS, Biobank UK, Million Women's Study, …

- Availability of millions of genetic markers
  - Genotyping costs decreasing rapidly
    - Cost per SNP: 2001 ($0.25) → 2003 ($0.10) → 2004 ($0.05)

- Methods for dealing with population structure advancing

- Background LD patterns being characterized
  - International HapMap and other projects (see McVean lecture)

Could argue that association studies haven't failed: they have yet to be conducted properly.

Key elements now in place to do so.

# Current Association Study Challenges

## 1) Genome-wide screen or candidate gene

### Genome-wide screen

- Hypothesis-free
- High-cost:  large genotyping requirements
- Multiple-testing issues
  - Possible many false positives, fewer misses

### Candidate gene

- Hypothesis-driven
- Low-cost: small genotyping requirements
- Multiple-testing less important
  - Possible many misses, fewer false positives

# Current Association Study Challenges

## 2) What constitutes a replication?

*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

### Extreme case (recently reported):

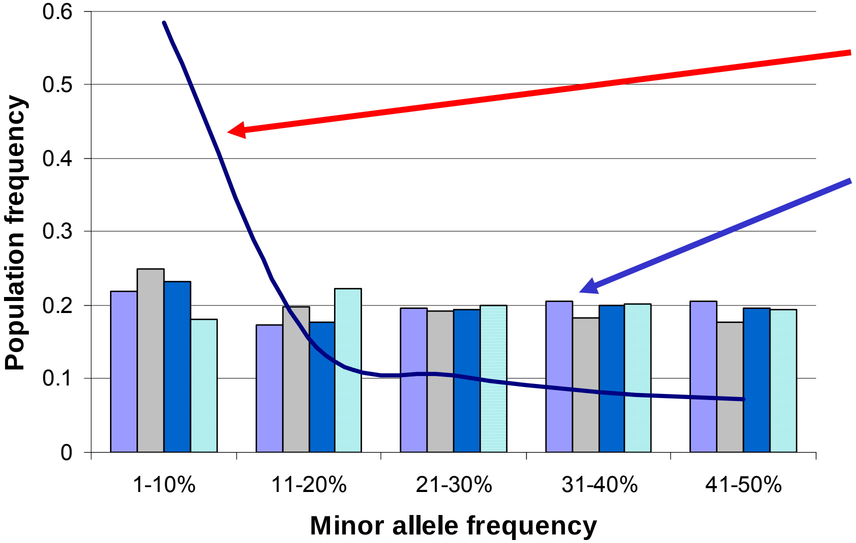
- "Replication" to correlated phenotype (asthma vs atopy).
- Different study design and selection strategies
  ("outcomes must attest to the robustness of the findings")
- Same gene region, different markers ("they're in LD, so must be okay")
- Opposite alleles/haplotype associated ("heterogeneity")

# Current Association Study Challenges

## 3) 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



Expected frequency in population

Frequency of public markers

# Current Association Study Challenges

## 3) Do we have the best set of genetic markers

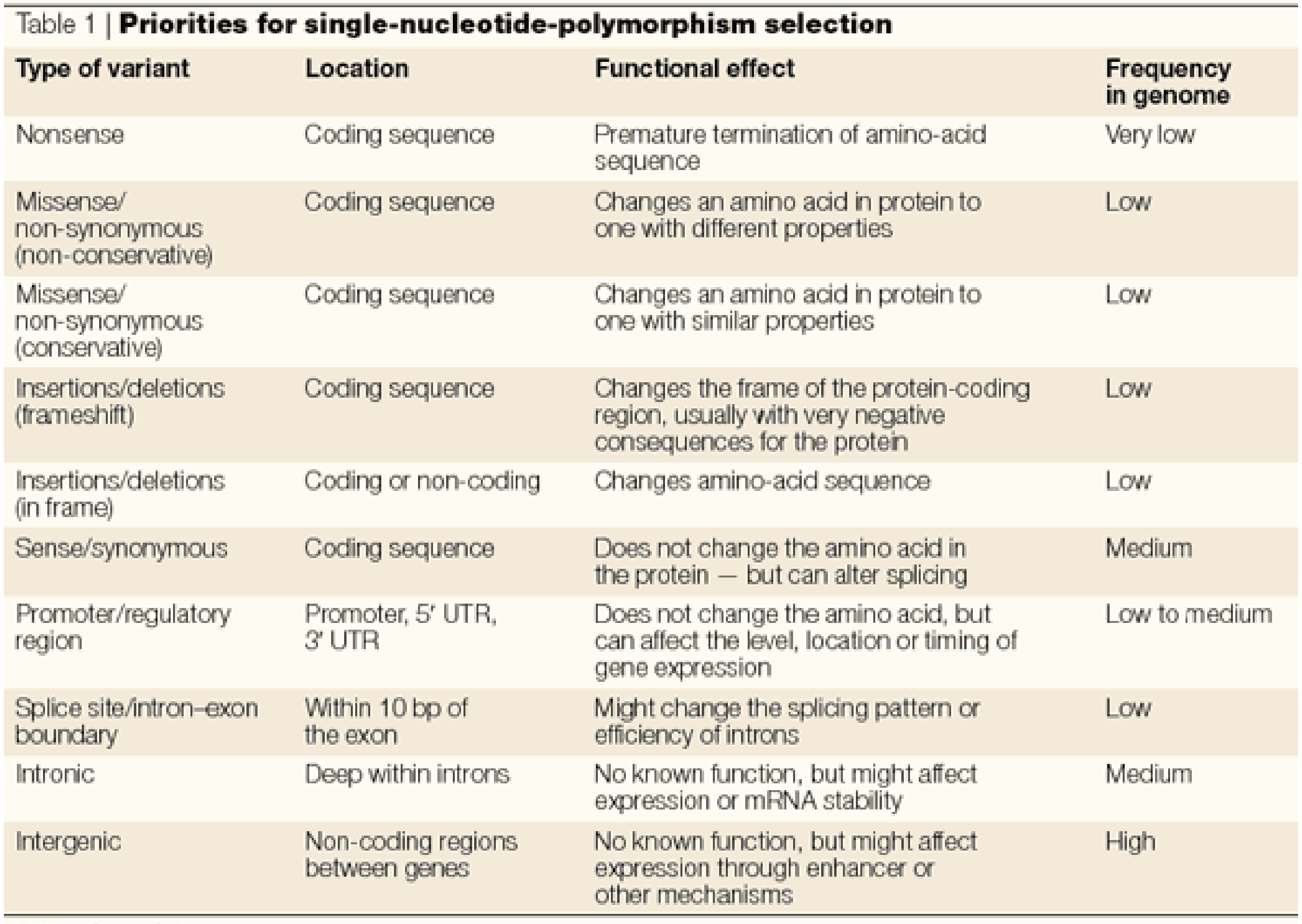


Table 1 | **Priorities for single-nucleotide-polymorphism selection**

| 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 |

Tabor et al, Nat Rev Genet 2003

# Current Association Study Challenges

## 4) Common-Disease Common-Variant Hypothesis

Common genes (alleles) contribute to inherited differences in common disease

Given recent human expansion, most variation is due to old mutations that have since become common rather than newer rare mutations.

Highly contentious debate in complex trait field

# Common-Disease/Common-Variant

## For

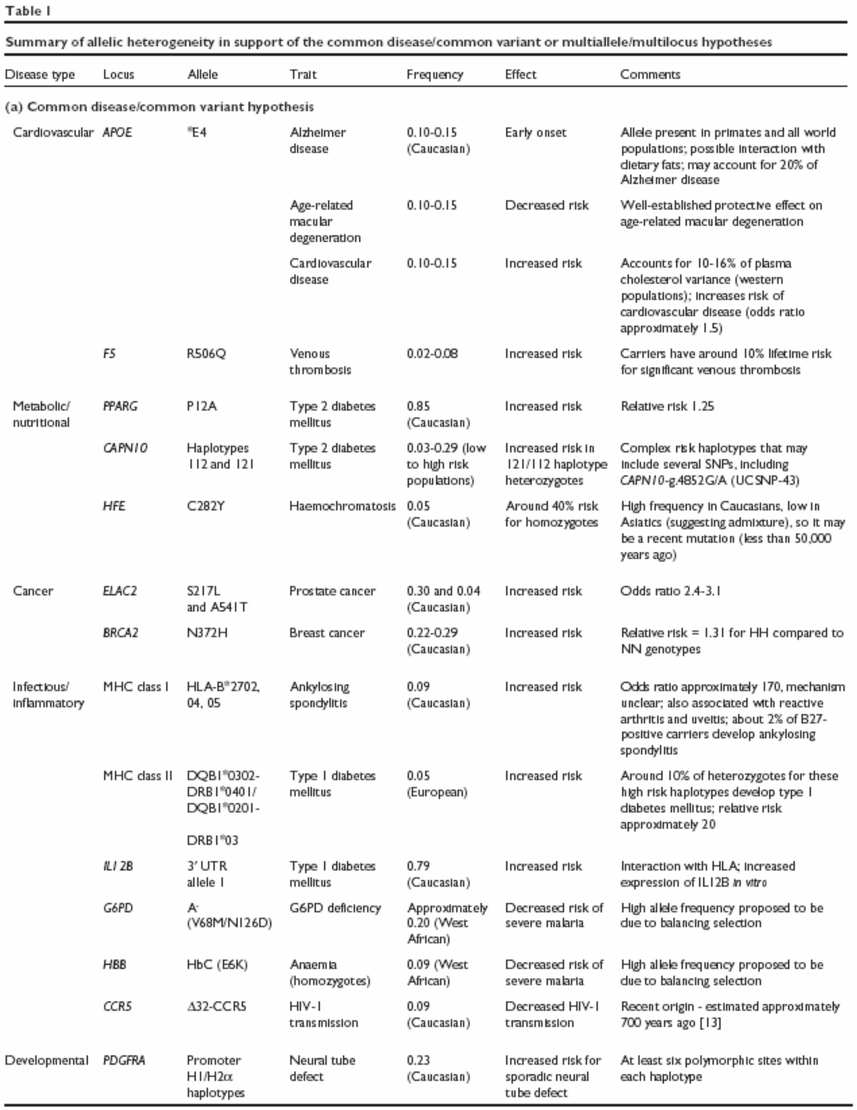


**Table 1**

**Summary of allelic heterogeneity in support of the common disease/common variant or multiallele/multilocus hypotheses**

| Disease type | Locus | Allele | Trait | Frequency | Effect | Comments |
|---|---|---|---|---|---|---|
| **(a) Common disease/common variant hypothesis** | | | | | | |
| Cardiovascular | *APOE* | *E4 | Alzheimer disease | 0.10-0.15 (Caucasian) | Early onset | Allele present in primates and all world populations; possible interaction with dietary fats; may account for 20% of Alzheimer disease |
| | | | Age-related macular degeneration | 0.10-0.15 | Decreased risk | Well-established protective effect on age-related macular degeneration |
| | | | Cardiovascular disease | 0.10-0.15 | Increased risk | Accounts for 10-16% of plasma cholesterol variance (western populations); increases risk of cardiovascular disease (odds ratio approximately 1.5) |
| | *F5* | R506Q | Venous thrombosis | 0.02-0.08 | Increased risk | Carriers have around 10% lifetime risk for significant venous thrombosis |
| Metabolic/ nutritional | *PPARG* | P12A | Type 2 diabetes mellitus | 0.85 (Caucasian) | Increased risk | Relative risk 1.25 |
| | *CAPN10* | Haplotypes 112 and 121 | Type 2 diabetes mellitus | 0.03-0.29 (low to high risk populations) | Increased risk in 121/112 haplotype heterozygotes | Complex risk haplotypes that may include several SNPs, including *CAPN10*-g.4852G/A (UCSNP-43) |
| | *HFE* | C282Y | Haemochromatosis | 0.05 (Caucasian) | Around 40% risk for homozygotes | High frequency in Caucasians, low in Asiatics (suggesting admixture), so it may be a recent mutation (less than 50,000 years ago) |
| Cancer | *ELAC2* | S217L and A541T | Prostate cancer | 0.30 and 0.04 (Caucasian) | Increased risk | Odds ratio 2.4-3.1 |
| | *BRCA2* | N372H | Breast cancer | 0.22-0.29 (Caucasian) | Increased risk | Relative risk = 1.31 for HH compared to NN genotypes |
| Infectious/ inflammatory | MHC class I | HLA-B*2702, 04, 05 | Ankylosing spondylitis | 0.09 (Caucasian) | Increased risk | Odds ratio approximately 170, mechanism unclear; also associated with reactive arthritis and uveitis; about 2% of B27-positive carriers develop ankylosing spondylitis |
| | MHC class II | DQB1*0302-DRB1*0401/ DQB1*0201-<br>DRB1*03 | Type 1 diabetes mellitus | 0.05 (European) | Increased risk | Around 10% of heterozygotes for these high risk haplotypes develop type 1 diabetes mellitus; relative risk approximately 20 |
| | *IL12B* | 3′ UTR allele 1 | Type 1 diabetes mellitus | 0.79 (Caucasian) | Increased risk | Interaction with HLA; increased expression of IL12B *in vitro* |
| | *G6PD* | A- (V68M/N126D) | G6PD deficiency | Approximately 0.20 (West African) | Decreased risk of severe malaria | High allele frequency proposed to be due to balancing selection |
| | *HBB* | HbC (E6K) | Anaemia (homozygotes) | 0.09 (West African) | Decreased risk of severe malaria | High allele frequency proposed to be due to balancing selection |
| | *CCR5* | Δ32-CCR5 | HIV-1 transmission | 0.09 (Caucasian) | Decreased HIV-1 transmission | Recent origin - estimated approximately 700 years ago [13] |
| Developmental | *PDGFRA* | Promoter H1/H2α haplotypes | Neural tube defect | 0.23 (Caucasian) | Increased risk for sporadic neural tube defect | At least six polymorphic sites within each haplotype |

## Against

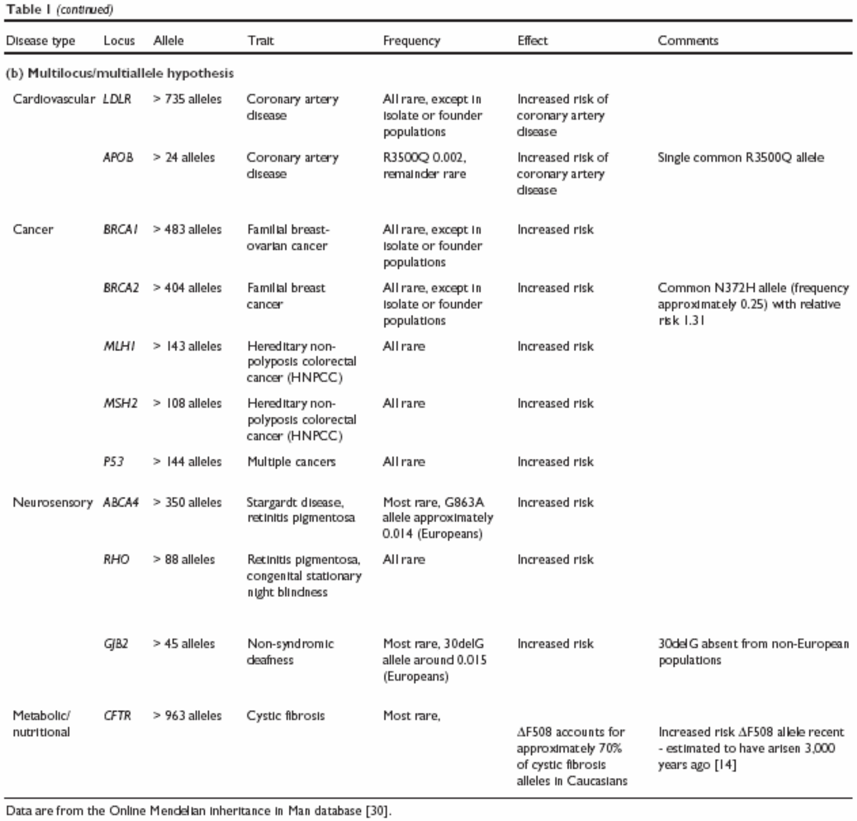
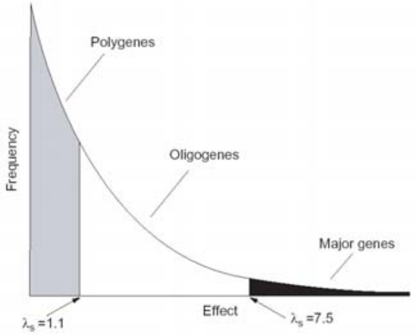


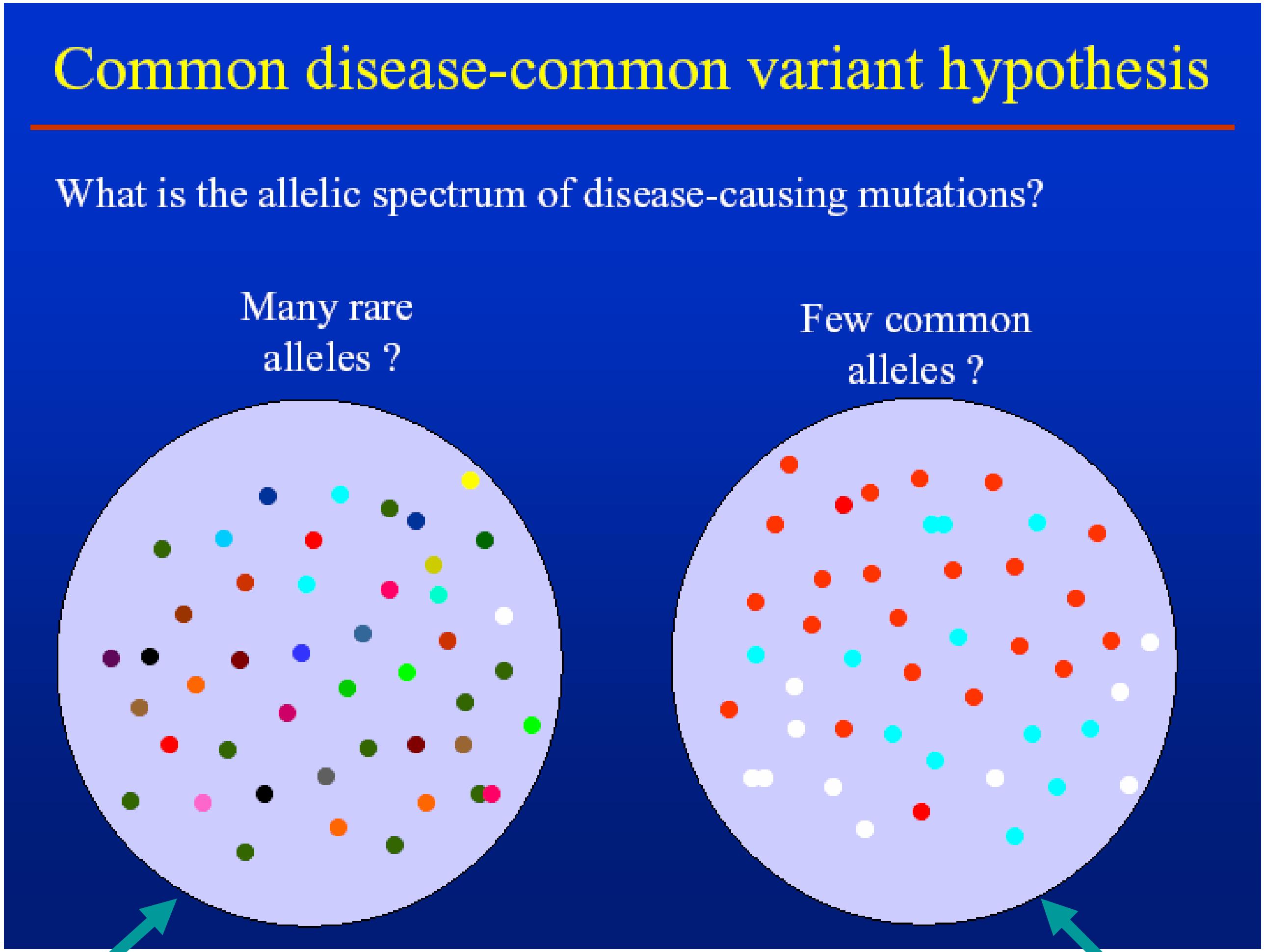
**Table 1** *(continued)*

| Disease type | Locus | Allele | Trait | Frequency | Effect | Comments |
|---|---|---|---|---|---|---|
| **(b) Multilocus/multiallele hypothesis** | | | | | | |
| Cardiovascular | *LDLR* | > 735 alleles | Coronary artery disease | All rare, except in isolate or founder populations | Increased risk of coronary artery disease | |
| | *APOB* | > 24 alleles | Coronary artery disease | R3500Q 0.002, remainder rare | Increased risk of coronary artery disease | Single common R3500Q allele |
| Cancer | *BRCA1* | > 483 alleles | Familial breast-ovarian cancer | All rare, except in isolate or founder populations | Increased risk | |
| | *BRCA2* | > 404 alleles | Familial breast cancer | All rare, except in isolate or founder populations | Increased risk | Common N372H allele (frequency approximately 0.25) with relative risk 1.31 |
| | *MLH1* | > 143 alleles | Hereditary non-polyposis colorectal cancer (HNPCC) | All rare | Increased risk | |
| | *MSH2* | > 108 alleles | Hereditary non-polyposis colorectal cancer (HNPCC) | All rare | Increased risk | |
| | *P53* | > 144 alleles | Multiple cancers | All rare | Increased risk | |
| Neurosensory | *ABCA4* | > 350 alleles | Stargardt disease, retinitis pigmentosa | Most rare, G863A allele approximately 0.014 (Europeans) | Increased risk | |
| | *RHO* | > 88 alleles | Retinitis pigmentosa, congenital stationary night blindness | All rare | Increased risk | |
| | *GJB2* | > 45 alleles | Non-syndromic deafness | Most rare, 30delG allele around 0.015 (Europeans) | Increased risk | 30delG absent from non-European populations |
| Metabolic/ nutritional | *CFTR* | > 963 alleles | Cystic fibrosis | Most rare, | ΔF508 accounts for approximately 70% of cystic fibrosis alleles in Caucasians | Increased risk ΔF508 allele recent - estimated to have arisen 3,000 years ago [14] |

Data are from the Online Mendelian inheritance in Man database [30].



Wright & Hastie, Genome Biol 2001



## Common disease-common variant hypothesis

What is the allelic spectrum of disease-causing mutations?

Many rare alleles ?

Few common alleles ?

Taken from Joel Hirschorn presentation, www.chip.org

If this scenario, association studies will not work

If this scenario, properly designed association studies should work well

# Current Association Study Challenges

## 5) Integrating the sampling, LD and epidemiology principles

Unanswerable questions in indirect association studies:

How much LD is needed to detect complex disease genes?

What effect size is big enough to be detected?

How common (rare) must a disease variant(s) be to be identifiable?

What marker allele frequency threshold should be used to find complex disease genes?

# Main Point

•In any indirect association study, we measure marker alleles that are *correlated* with disease variants…

We do not measure the disease variants themselves

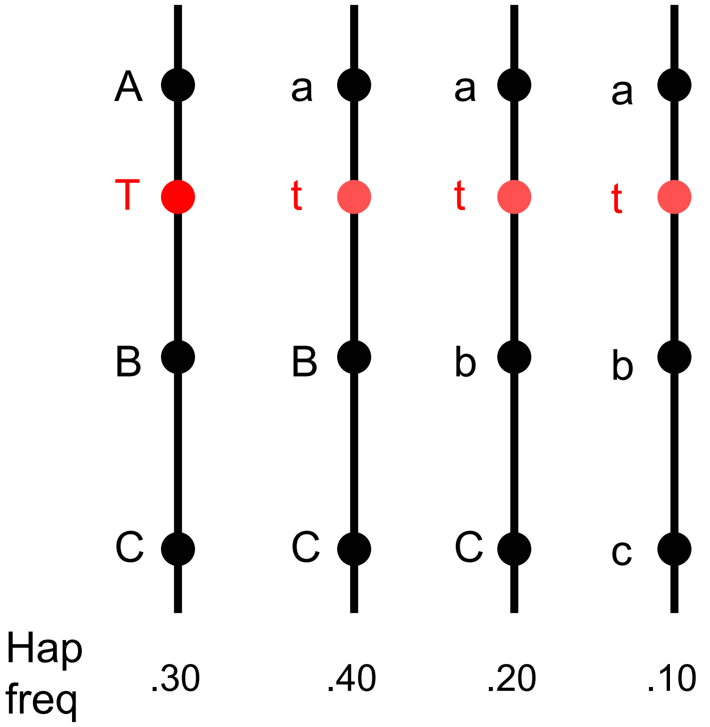
•But, for study design and power, we concern ourselves with frequencies and effect sizes *at the disease locus*….

This can only lead to underpowered studies and inflated expectations

•We should concern ourselves with the apparent effect size at the marker, which results from

1) difference in frequency of marker and disease alleles
2) LD between the marker and disease loci
3) effect size of disease allele

# Single Trait allele (or multiple alleles on same haplotype)



| Allele freq | D | $D'_{(marker,T)}$ | $r^2_{(marker,T)}$ | $OR_M$ |
|---|---|---|---|---|
| A = 0.30 | 0.21 | 1.0 | 1.0 | 2.00 |
| T = 0.30 | | | | 2.00 |
| B = 0.70 | 0.09 | 1.0 | .18 | 1.43 |
| C = 0.90 | 0.03 | 1.0 | .05 | 1.33 |

Zondervan & Cardon, Nat. Rev. Gen. 2004; 5: 89-100

# Integrating sampling, LD and epi…

- **‘Rare’ variants (0.001<x<0.1):**
  - with small effect sizes (OR <1.5) → *not detectable* in large studies (X000s)
  - with moderate - large effect sizes (OR > 2.0) → detectable

- **‘Common’ variants (>0.1):**
  - will have modest effect sizes (OR <2.0) → detectable **ONLY in large studies (X000s) and iff MAF ≈ DAF and LD is high**

⇒***Strongest argument for using common markers is not CD-CV; it is practical. For small effects, common markers are only ones for which we have sufficient sample sizes.***

# Future

Better samples, larger marker sets, improved statistical measures, greater understanding of LD, …*hold real promise for association*

1) Some important disease genes will emerge
2) Not all important disease genes will be identified

The diseases are severe enough to warrant the effort, even if it yields only some of the answers

# Suggested Reading

- Balding, D. J., M. Bishop and C. Cannings (Editors), 2001 *Handbook of Statistical Genetics*. John Wiley & Sons, Chichester.
- Elston, R. C., L. J. Palmer and J. E. Olson (Editors), 2002 *Biostatistical Genetics and Genic Epidemiology*. John Wiley & Sons, Chichester.
- Falconer, D. S., 1981 *Introduction to quantitative genetics*. Longman Group, Ltd., Harlow, U.K.
- Lynch, M., and B. Walsh, 1998 *Genetics and analysis of quantitative traits*. Sinauer Associates, Sunderland, Mass.
- Terwilliger, J. D., and J. Ott, 1994 *Handbook of human genetic linkage*. Johns Hopkins University Press, Baltimore.
- Sham, P., 1997 *Statistics in Human Genetics*. Hodder Arnold, London.

- Botstein, D., and N. Risch, 2003 Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet **33 Suppl:** 228-237.
- Cardon, L. R., and J. I. Bell, 2001 Association study designs for complex diseases. Nat Rev Genet **2:** 91-99.
- Devlin, B., K. Roeder and L. Wasserman, 2001 Genomic control, a new approach to genetic-based association studies. Theor Popul Biol **60:** 155-166.
- Pritchard, J. K., M. Stephens and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics **155:** 945-959.
- Spielman, R., R. McGinnis and W. Ewens, 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics **52:** 506-516.
- Risch, N. J., 2000 Searching for genetic determinants in the new millennium. Nature **405:** 847-856.
- Weiss, K. M., and J. D. Terwilliger, 2000 How many diseases does it take to map a gene with SNPs? Nat Genet **26:** 151-157.