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 Dystrophy
  - 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

Genomics
- Physical Mapping/
  Sequencing

Chromosome Region
- LOD

Candidate Gene Selection/
Polymorphism Detection

Association Study
- Mutation Characterization/
  Functional Annotation

| +C       |
| +C       |
| -G       |
| -C       |

..... GAG GGG GCC 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

Disease linked to ‘5’ allele in dominant inheritance pattern

Allele coded by CA copies
2 = CACA
6 = CACACACACACACA
Netherton Syndrome Linkage

Netherton Syndrome Haplotypes

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, MIM 256300). Most of these mutations predict premature termination codons. These results disclose a critical role of **SPINK5** on 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 (trichorhexis invaginata or bamboo hair), and severe atopic manifestations including atopic dermatitis and hayfever, with high serum IgE levels and hyperemorrhagia. Failure to thrive, infections and hypernatremic dehydration result in high postnatal mortality.

We recently localized the NS gene locus to 5q32 (ref. 2), in a region where the gene encoding the serine protease inhibitor LEKTI (for lympho-epithelial Kazal-type related inhibitor) has previously been mapped.

**Fig. 1.** Northern-blot analysis and **SPINK5** expression in NS families. **a**, Northern-blot analysis of **SPINK5** expression in cultured epidermal keratinocytes from a control individual (6) and two NS patients (7,8). A **SPINK5**-specific probe shows a 1.3-kb hybridization signal in the control and a reduction of signal in the keratinocyte line from a NS patient (arrowheads). Keratinocytes were isolated from each family at risk for the disease. **b**, Mutations detected in 11 different families (1:4) and 8 different families (1:7). **c**, Mutation N205D, which is a nonsense-mediated decay site, is observed in 14% of NS patients. **d**, Mutation N205D, which is a nonsense-mediated decay site, is observed in 14% of NS patients. **e**, Mutation N205D, which is a nonsense-mediated decay site, is observed in 14% of NS patients. **f**, Mutation N205D, which is a nonsense-mediated decay site, is observed in 14% of NS patients.
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:  $\lambda = \frac{\text{risk to relative of an affected individual}}{\text{risk in general population}}$
  – $\lambda$ 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
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

Inflammatory Bowel Disease Genome Screen

Linkage Outcomes for Complex Traits

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 Wjst1

Institutes of 1Epidemiology and 2Biomathematics and Biometrics, National Research Center for Environment and Health, Neuherberg, Germany; 3Channing Laboratory, Brigham and Women’s Hospital and Harvard Medical School, Boston; and 4Department 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

\[ \lambda_s = 1.5; \text{ Lods calculated using Kong & Cox (signed) procedure} \]
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...
The weakest link?

Genetics
Sib pairs

Genomics
Physical Mapping/
Sequencing

Chromosome Region

Association Study

Candidate Gene Selection/
Polymorphism Detection

Mutation Characterization/
Functional Annotation

...... 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

Disease linked to ‘5’ allele in dominant inheritance

Allele coded by CA copies
2 = CACA
6 = CACACACACACACA

affected
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

Controls

Cases

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

<table>
<thead>
<tr>
<th><strong>Linkage</strong></th>
<th><strong>Association</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Requires families</td>
<td>1. Families or unrelateds</td>
</tr>
<tr>
<td>3. Few markers for genome coverage (300-400 STRs)</td>
<td>3. Many markers for genome coverage ($10^5 – 10^6$ SNPs)</td>
</tr>
<tr>
<td>4. Allele-sharing weak design</td>
<td>4. Powerful design based on means</td>
</tr>
<tr>
<td>5. Yields coarse location</td>
<td>5. Yields fine-scale location</td>
</tr>
<tr>
<td>6. Good for initial detection; poor for fine-mapping</td>
<td>6. Good for fine-mapping, poor for initial detection</td>
</tr>
<tr>
<td>7. Powerful for rare variants</td>
<td>7. Powerful for common variants; rare variants generally impossible</td>
</tr>
</tbody>
</table>
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

Sometimes it’s hot, sometimes it’s not
Åke Lernmark¹ & Jurg Ott²

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

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

How many diseases does it take to map a gene with SNPs?
Kenneth M. Weiss¹ & Joseph D. Terwilliger²
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

Original study
- Deeb et al. (ref. 12)

Subsequent case/control studies
- Mancini et al. (ref. 20)
- Ringel et al. (ref. 22)
- Mehta et al. (ref. 21)
- Clement et al. (ref. 18)*
- Hara et al. (ref. 19)

Scandinavian case-control (this study)
SLSJ case-control (this study)

All subsequent case/control studies, pooled

Subsequent family-based studies
- Scandinavian trio (this study)
- Scandinavian sibship (this study)

All subsequent family-based studies, pooled

All data, this study
All subsequent data, including this study

Estimated risk (Aa allele) 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 1.1 1.2 1.3

ACE and MI

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

<table>
<thead>
<tr>
<th>Published studies</th>
<th>Number of cases</th>
<th>Cases</th>
<th>Controls</th>
<th>Risk ratio and 99% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECTIM</td>
<td>610</td>
<td>32%</td>
<td>27%</td>
<td></td>
</tr>
<tr>
<td>35 small studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&lt;200 cases each)</td>
<td>3578</td>
<td>34%</td>
<td>26%</td>
<td></td>
</tr>
<tr>
<td>14 larger studies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&gt;200 cases each)</td>
<td>6863</td>
<td>27%</td>
<td>27%</td>
<td></td>
</tr>
<tr>
<td>ISIS</td>
<td>4629</td>
<td>29%</td>
<td>28%</td>
<td></td>
</tr>
</tbody>
</table>

Altshuler et al *Nat Genet* 2000
Keavney et al *Lancet* 2000
Phenotypic Complexity

Weiss & Terwilliger, Nat Genet, 2000
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.
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

<table>
<thead>
<tr>
<th></th>
<th>Sample ‘A’</th>
<th></th>
<th>Sample ‘B’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>m</td>
<td>Freq.</td>
</tr>
<tr>
<td>Affected</td>
<td>50</td>
<td>50</td>
<td>.10</td>
</tr>
<tr>
<td>Unaffected</td>
<td>450</td>
<td>450</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>.50</td>
<td>.50</td>
<td></td>
</tr>
</tbody>
</table>

\[ \chi^2_1 \] is n.s.

\[ \chi^2_1 \] is n.s.

Neither has significant association
### Population Stratification

#### Sample ‘A’

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>m</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affected</td>
<td>50</td>
<td>50</td>
<td>.10</td>
</tr>
<tr>
<td>Unaffected</td>
<td>450</td>
<td>450</td>
<td>.90</td>
</tr>
<tr>
<td></td>
<td>.50</td>
<td>.50</td>
<td></td>
</tr>
</tbody>
</table>

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

#### Sample ‘B’

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>m</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affected</td>
<td>1</td>
<td>9</td>
<td>.01</td>
</tr>
<tr>
<td>Unaffected</td>
<td>99</td>
<td>891</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>.10</td>
<td>.90</td>
<td></td>
</tr>
</tbody>
</table>

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

#### Combined

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>m</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affected</td>
<td>51</td>
<td>59</td>
<td>.055</td>
</tr>
<tr>
<td>Unaffected</td>
<td>549</td>
<td>1341</td>
<td>.945</td>
</tr>
<tr>
<td></td>
<td>.30</td>
<td>.70</td>
<td></td>
</tr>
</tbody>
</table>

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

**Association induced by sample mixing**
Population Stratification: Real Example

Full heritage American Indian Population

<table>
<thead>
<tr>
<th>Gm&lt;sup&gt;3;5,13,14&lt;/sup&gt; haplotype</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>7.8%</td>
<td>29.0%</td>
</tr>
<tr>
<td>-</td>
<td>92.2%</td>
<td>71.0%</td>
</tr>
</tbody>
</table>

NIDDM Prevalence ≈ 40%

Caucasian Population

<table>
<thead>
<tr>
<th>Gm&lt;sup&gt;3;5,13,14&lt;/sup&gt; haplotype</th>
<th>Cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>66%</td>
<td>34%</td>
</tr>
<tr>
<td>-</td>
<td>34%</td>
<td>66%</td>
</tr>
</tbody>
</table>

NIDDM Prevalence ≈ 15%

Study without knowledge of genetic background:

OR=0.27
95%CI=0.18 to 0.40

Proportion with NIDDM by heritage and marker status

<table>
<thead>
<tr>
<th>Index of Indian Heritage</th>
<th>Gm&lt;sup&gt;3;5,13,14&lt;/sup&gt; haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>-</td>
</tr>
<tr>
<td>0</td>
<td>17.8%</td>
</tr>
<tr>
<td></td>
<td>19.9%</td>
</tr>
<tr>
<td>4</td>
<td>28.3%</td>
</tr>
<tr>
<td></td>
<td>28.8%</td>
</tr>
<tr>
<td>8</td>
<td>35.9%</td>
</tr>
<tr>
<td></td>
<td>39.3%</td>
</tr>
</tbody>
</table>

Reviewed in Cardon & Palmer, Lancet 2003
‘Control’ Samples in Human Genetics

< 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 $\rightarrow$ 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

<table>
<thead>
<tr>
<th>Genome-wide screen</th>
<th>Candidate gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Hypothesis-free</td>
<td>• Hypothesis-driven</td>
</tr>
<tr>
<td>• High-cost: large genotyping requirements</td>
<td>• Low-cost: small genotyping requirements</td>
</tr>
<tr>
<td>• Multiple-testing issues</td>
<td>• Multiple-testing less important</td>
</tr>
<tr>
<td>– Possible many false positives, fewer misses</td>
<td>– Possible many misses, fewer false positives</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Type of variant</th>
<th>Location</th>
<th>Functional effect</th>
<th>Frequency in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsense</td>
<td>Coding sequence</td>
<td>Premature termination of amino-acid sequence</td>
<td>Very low</td>
</tr>
<tr>
<td>Missense/ non-synonymous (non-conservative)</td>
<td>Coding sequence</td>
<td>Changes an amino acid in protein to one with different properties</td>
<td>Low</td>
</tr>
<tr>
<td>Missense/ non-synonymous (conservative)</td>
<td>Coding sequence</td>
<td>Changes an amino acid in protein to one with similar properties</td>
<td>Low</td>
</tr>
<tr>
<td>Insertions/deletions (frameshift)</td>
<td>Coding sequence</td>
<td>Changes the frame of the protein-coding region, usually with very negative consequences for the protein</td>
<td>Low</td>
</tr>
<tr>
<td>Insertions/deletions (in frame)</td>
<td>Coding or non-coding</td>
<td>Changes amino-acid sequence</td>
<td>Low</td>
</tr>
<tr>
<td>Sense/synonymous</td>
<td>Coding sequence</td>
<td>Does not change the amino acid in the protein — but can alter splicing</td>
<td>Medium</td>
</tr>
<tr>
<td>Promoter/regulatory region</td>
<td>Promoter, 5' UTR, 3' UTR</td>
<td>Does not change the amino acid, but can affect the level, location or timing of gene expression</td>
<td>Low to medium</td>
</tr>
<tr>
<td>Splice site/intron–exon boundary</td>
<td>Within 10 bp of the exon</td>
<td>Might change the splicing pattern or efficiency of introns</td>
<td>Low</td>
</tr>
<tr>
<td>Intronic</td>
<td>Deep within introns</td>
<td>No known function, but might affect expression or mRNA stability</td>
<td>Medium</td>
</tr>
<tr>
<td>Intergenic</td>
<td>Non-coding regions between genes</td>
<td>No known function, but might affect expression through enhancer or other mechanisms</td>
<td>High</td>
</tr>
</tbody>
</table>

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
### For

<table>
<thead>
<tr>
<th>Disease type</th>
<th>Locus</th>
<th>Allele</th>
<th>Trait</th>
<th>Frequency</th>
<th>Effect</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>AP01</td>
<td>T4</td>
<td>Alzheimer disease</td>
<td>0.16-0.75 (Caucasian)</td>
<td>Early onset</td>
<td>Allele present in prion and all world populations, positive interactions with obesity, may account for 20% of Alzheimer disease</td>
</tr>
<tr>
<td>Age-related</td>
<td>0.13-0.75</td>
<td>Increased risk</td>
<td>Well-established positive effect on age-related cognitive degeneration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiomyopathy</td>
<td>0.10-0.75</td>
<td>Increased risk</td>
<td>Accounts for 15-10% of plasma cholesterol variance (between populations); increases risk of cardiovascular disease; allele frequency approximately 1.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>RS6606</td>
<td>Venous thrombosis</td>
<td>0.02-0.60</td>
<td>Increased risk</td>
<td>Current have around 18% time risk for spontaneous venous thrombosis</td>
<td></td>
</tr>
<tr>
<td>Hypertension</td>
<td>P394G</td>
<td>P12A</td>
<td>Type 2 diabetes, hypertension,</td>
<td>0.05 (Caucasian)</td>
<td>Increased risk</td>
<td>Positive risk for thrombosis</td>
</tr>
<tr>
<td>G279C</td>
<td>Hypertension</td>
<td>0.02-0.29 (African)</td>
<td>Type 2 diabetes, hypertension, hyperlipidemia</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HFE</td>
<td>C282Y</td>
<td>Hemochromatosis</td>
<td>0.05 (Caucasian)</td>
<td>Increased risk</td>
<td>Accumulates 40% risk for hemochromatosis</td>
<td></td>
</tr>
</tbody>
</table>

### Against

<table>
<thead>
<tr>
<th>Disease type</th>
<th>Locus</th>
<th>Allele</th>
<th>Trait</th>
<th>Frequency</th>
<th>Effect</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular</td>
<td>LDLR</td>
<td>&gt; 75 alleles</td>
<td>Coronary artery disease</td>
<td>All new, except to be tested for rare alleles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP01</td>
<td>&gt; 24 alleles</td>
<td>Coronary artery disease</td>
<td>rs1049010 G/C polymorphism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>BRCA1</td>
<td>&gt; 343 alleles</td>
<td>Breast cancer</td>
<td>All new, except to be tested for rare alleles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA2</td>
<td>&gt; 343 alleles</td>
<td>Breast cancer</td>
<td>All new, except to be tested for rare alleles</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSH2</td>
<td>&gt; 343 alleles</td>
<td>Hereditary nonpolyposis colorectal cancer (HNPCC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSH4</td>
<td>&gt; 343 alleles</td>
<td>Hereditary nonpolyposis colorectal cancer (HNPCC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>&gt; 343 alleles</td>
<td>Multicancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neurodegenerative</td>
<td>ABC1D</td>
<td>&gt; 350 alleles</td>
<td>Stargardt disease, macular degeneration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APOE</td>
<td>&gt; 350 alleles</td>
<td></td>
<td>Macular, C677T, allele frequencies 0.014 (Europeans)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV0</td>
<td>&gt; 350 alleles</td>
<td>Inflammatory arthritis, conjunctivitis, ocular redness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G4</td>
<td>&gt; 350 alleles</td>
<td>Macular degeneration, ataxia, cerebral autosomal dominant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metabolism</td>
<td>ABC1D</td>
<td>&gt; 350 alleles</td>
<td>Cystic fibrosis</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data from the online Mendelian Inheritance in Man database [38].
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)

<table>
<thead>
<tr>
<th>Haplotype</th>
<th>Allele freq</th>
<th>D</th>
<th>$D'_{(marker,T)}$</th>
<th>$r^2_{(marker,T)}$</th>
<th>$OR_M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.30</td>
<td>0.21</td>
<td>1.0</td>
<td>1.0</td>
<td>2.00</td>
</tr>
<tr>
<td>T</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td>2.00</td>
</tr>
<tr>
<td>B</td>
<td>0.70</td>
<td>0.09</td>
<td>1.0</td>
<td>.18</td>
<td>1.43</td>
</tr>
<tr>
<td>C</td>
<td>0.90</td>
<td>0.03</td>
<td>1.0</td>
<td>.05</td>
<td>1.33</td>
</tr>
</tbody>
</table>

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
