IDENTIFYING DISEASE GENES

Nature, 15 Feb 2001

Science, 16 Feb 2001
IDENTIFYING DISEASE GENES

• Identifying genes that contribute to disease risk is one of the main objectives of molecular research.

• Such findings have contributed to improvements in diagnosis, prognosis and therapy.

• With the successful identification of disease genes for many single-gene disorders, the focus has shifted to diseases with a complex, multifactorial aetiology.

• What are the approaches available?
GENETIC INHERITANCE: 3 MAIN TYPES

1. Chromosome disorders
   • Many genes duplicated or deleted

2. Direct cause and effect
   • A mutation in a single gene is enough to cause disease

3. Common complex disorders
   • Genes and environment influence risk for disease
   • This is the case for most common chronic diseases
SINGLE GENE VS. COMPLEX DISEASE (I)

Normal red blood cells

Sickle red blood cells
WHAT WORKED FOR SINGLE-GENE DISORDERS?

1. Define disease phenotype
2. Collect multiplex families
3. Perform linkage genome screen
4. Obtain initial localization
5. Define minimum candidate region
6. Test candidate genes for mutations
ASSOCIATION STUDIES

• Association refers to the co-occurrence of an allele, genotype or haplotype with a disease trait, more frequently than can be readily explained by chance.

• Allele B is associated with Disease D, if persons with D have B more often than expected in the general population.

• The principle of an association study is also simple-- gather some people with a disease and some people without a disease, and look to see if a certain allele (or genotype) is present more often in the cases than the controls.
CANDIDATE GENE VS. GENOME-WIDE

• Hypothesis driven
• More cost-effective (?)
• Statistical burden is reduced
• Sample size
  • Fewer samples compel fewer comparisons
  • DNA is a non-renewable resource
• Fishing expedition
APPROACHES TO GWAS

• Identify all 12 million common SNPs
• Collect 1,000 cases and 1,000 controls
• Genotype all DNAs for all SNPs
• That adds up to 24 billion genotypes

• Imagine, this approach cost 50 cents a genotype.
• That’s R12 billion for each disease – completely out of the question!!
2007 SCIENTIFIC BREAKTHROUGH OF THE YEAR
Science Magazine, December 21, 2007

“It’s all about me!”

Single Nucleotide Polymorphisms (SNPs)

<table>
<thead>
<tr>
<th>Individual 1</th>
<th>Individual 2</th>
<th>Individual 3</th>
<th>Individual 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACAGCCA.....</td>
<td>TTCGGGGTC....</td>
<td>AACAGCCA.....</td>
<td>TTCGAAGTC....</td>
</tr>
<tr>
<td>AACAGCCA.....</td>
<td>TTCGAAGTC....</td>
<td>AACATGCCA.....</td>
<td>TTCGGGGTC....</td>
</tr>
<tr>
<td>AACAGCCA.....</td>
<td>TTCGGGGTC....</td>
<td>AACAGCCA.....</td>
<td>TTCGGGGTC....</td>
</tr>
<tr>
<td>AACAGCCA.....</td>
<td>TTCGGGGTC....</td>
<td>AACAGCCA.....</td>
<td>TTCGGGGTC....</td>
</tr>
</tbody>
</table>
The International HapMap Project is a partnership of scientists and funding agencies from Canada, China, Japan, Nigeria, the United Kingdom and the United States to develop a public resource that will help researchers find genes associated with human disease and response to pharmaceuticals. See "About the International HapMap Project" for more information.

Project Information
- About the Project
- HapMap Publications
- HapMap Tutorial
- HapMap Mailing List
- HapMap Project Participants
- HapMap Mirror Site in Japan

Project Data
- HapMap Genome Browser (Phase 1, 2 & 3 - merged genotypes & frequencies)
- HapMap Genome Browser (Phase 3 - genotypes, frequencies & LD)
- HapMap Genome Browser (Phase 1 & 2 - full dataset)
- GWAs Karyogram
- HapMart
- HapMap FTP
- Bulk Data Download
- Data Freezes for Publication
- ENCODE Project
- Guidelines For Data Use

News
- **2009-12-14: Notice to Haplovie users**
  Recently, there are several questions about Haplovie data format errors, and these errors were observed when users tried to analyze HapMap release 27 data dumped from HapMap. The current Haplovie version (4.1) does not work with release 27 data. Haplovie will generate a software error similar to "Hapmap data format error: NA09884" when trying to open the data.

  The r27 data format will be supported by next Haplovie version. There is a beta test version that is supposed to work and it can be obtained from http://www.broadinstitute.org/haplovie/haplovie-downloads. But since it is NOT an official release version, please use it base on your own judgment.

- **2009-12-10: Corrected HapMap3 phased haplotypes available for chromosome X**
  Phased haplotypes for consensus HapMap3 release 2 data for chromosome X has been corrected and the new data are now available for bulk download. Sorry for any inconvenience this might have caused.

- **2009-12-02: HapMap3 phased haplotypes available for chromosome X**
  Phased haplotypes for consensus HapMap3 release 2 data has been phased for chromosome X and are now available for bulk download. [Update: The downloading was disabled because several users have found that there are repeating data in some of the chrX phasing data files. The data source is being contacted and the downloading will be enabled as soon as the problem is cleared.]

http://www.hapmap.org/
APPROACHES TO GWAS: TAG SNPS

Diagram showing approaches to GWAS focusing on tagging SNPs.
Published Genome-Wide Associations through 12/2009.
658 published GWA at $p \leq 5 \times 10^{-8}$

NHGRI GWA Catalog
www.genome.gov/GWAStudies
GWAS – GENOTYPE DATA

THE BAD & THE UGLY....

[Graphs showing scatter plots with data points and numbers associated with each plot.]

rs72096

rs7498865

rs8713532
GENOTYPING RESULTS - COPY NUMBER VARIATION

The deletion appears as loss of heterozygotes in the 1<sup>st</sup> plot, and the duplication splits the heterozygous cluster in the 2<sup>nd</sup> plot.
RESULTS FROM GWAS STUDIES
QQ PLOTS

a. Observed data conforms closely to expectation – little evidence for association
b. Inflation of observed findings across distribution – population stratification
c. Population substructure and evidence of strong association
d. Little evidence of substructure, compelling evidence of strong association
RESULTS FROM GWAS STUDIES
MANHATTAN PLOTS
RESULTS FROM GWAS STUDIES
REGIONAL PLOTS

---

**a**

- **rs1801274**
- *FCGR3A*
- *FCGR2A*
- *FCGR2C*
- *FCGR3B*
- *FCGR2B*
- *FCRLA*
- *FCRLD*
- *USP12*

- **P = 10^{-6}**
- **P = 10^{-12}**

- 98-Kb region without GWAS SNP coverage

**Physical distance: 273.6 kb**

**LD map type:** $D^*$

**0 0.2 0.4 0.6 0.8 1**

---

**a**

- **Observed (-log P)**
- **Recombination rate (cM/Mb)**

- **P = 2.24 × 10^{-13}**

- Chromosome 21 position (kb):
  - 29,325
  - 29,625
  - 29,325

- Genes:
  - RPL12PA
  - C21orf17
  - BACH1
  - GRIK1
RESULTS FROM GWAS STUDIES
ALLELE SCORING

![Graph showing the relationship between the number of weighted risk alleles and mean BMI. The graph displays a histogram of the number of individuals with different numbers of weighted risk alleles, along with a line indicating the mean BMI for each category. The x-axis represents the number of weighted risk alleles, ranging from <21 to ≥88, and the y-axis represents the number of individuals. The mean BMI values are plotted with error bars, showing the variability around the mean.]
GWAS ISSUES: POPULATION STRATIFICATION

CASES

CONTROLS
GWAS ISSUES: POPULATION STRATIFICATION

Population Group 1

CASES

CONTROLS
GWAS ISSUES:
POPULATION STRATIFICATION

Population group 2

CASES

CONTROLS
GWAS ISSUES: POPULATION STRATIFICATION
POPULATION STRATIFICATION: PRINCIPAL COMPONENT ANALYSIS
PITFALLS OF GWAS – MISSING HERITABILITY

The case of the missing heritability
PITFALLS OF GWAS – MISSING HERITABILITY

• Too stringent statistical cut-off
• Larger sample sizes
• Rare variants
• Structural variants
• Epigenetics
• Gene-Environment interaction

• Future – whole-genome sequencing of individuals
PITFALLS OF GWAS – AFRICAN SAMPLES

RSA

EUR

YRI

CHB/JPT
SINGLE GENE DISORDERS

1. Define disease phenotype
2. Collect multiplex families
3. Perform linkage genome screen
4. Obtain initial localization
5. Define minimum candidate region
6. Test candidate genes for mutations
COMPLEX DISEASE

DEFINE PHENOTYPE
- CLINICAL DEFINITION
- GENETIC CONTRIBUTION

APPROACH
- GENOME-WIDE
- CANDIDATE GENE

STUDY DESIGN
- FAMILY-BASED
- POPULATION-BASED

ANALYSIS
- STATISTICAL ANALYSIS
- LABORATORY ANALYSIS

VERIFICATION
- REPLICATION
- FUNCTIONAL STUDIES

Sample size
- Linkage Analysis
- Association Study

Association Study
- LABORATORY ANALYSIS
- REPLICATION
- FUNCTIONAL STUDIES
23ANDME KIT