

# Strategies for finding disease genes

Dennis Drayna, PhD  
SSBCD/NIDCD/NIH

## Is the disease genetic?

- Many non-genetic traits cluster in families
- For diseases that have a genetic component, how much is genetic and how much non-genetic in origin?
  - Twin studies
  - Adoption studies
  - Segregation analysis

## Measures of genetic contribution

Heritability -  $h^2$

– Ranges from 0 to 1

- Examples

- Plasma cholesterol levels  $h^2 = .30 - .40$

- Bone mass density  $h^2 = .45 - .60$

- Stature  $h^2 = .80 - .90$

- High heritability does not imply simple underlying genetics

- Cannot distinguish one gene causing 100% of the disease from 100 genes each causing 1% of the disease

- Cannot estimate mode of inheritance

## Positional Cloning

- The ability to identify a gene based solely on its position in the genome
- Does not require knowledge of the underlying biology of the gene product
- Method used to identify the genes underlying the important Mendelian disorders in humans
- Begins with a linkage study

## Genetic linkage

- The non-random association of two traits as they are passed from parent to offspring
- Violations of Mendel's 2nd Law
- In humans, linkage studies are enabled by the use of polymorphic DNA markers
  - Variation in the DNA sequence itself is the heritable trait
- Several thousand well-characterized STRP/microsatellite markers available
- Several million well-characterized SNP markers available

## Steps in a linkage study

- Identify families carrying the disease
- Obtain DNA and diagnostic samples
  - $10^1$  to  $10^2$  individuals typically included
- Type each DNA at ~400 marker loci
- Analyze each marker for linkage to the trait
  - LOD score method
- Once linkage is found, evaluate all genes in the region to identify the causative mutation

## Proof a variant is causative

- The presumptive causative genotype exists only in affected individuals and not normals
- Different mutations in the same gene occur in different families with the same disease

## Success, then expansion

- Positional cloning strategies successfully identified the genes underlying all of the major Mendelian disorders in humans
  - Strategy still employed for very rare disorders that can generate significant biological insights
- Similar strategies were then applied to the major non-Mendelian disorders
  - Important health problems with significant but not exclusively genetic causes

## Then failure

- Linkage studies of dozens of important human medical disorders failed miserably
  - Psychiatric disease, diabetes, asthma, osteoporosis
- Low LOD scores
- Irreproducible results
- Inability to specify precise limits on the location of the gene
- No genes of major effect, akin to those underlying Mendelian disorders, were found

## What's the problem?

- Sample sizes were often small
- Analytical methods, such as the LOD score method are inappropriate for non-Mendelian patterns of inheritance
- Mendelian disorders may be a poor model in many respects for studies of non-Mendelian disorders
  - Penetrance, expressivity, mutation frequency

## The origins of common disease I.

- Rare variant - common disease hypothesis
  - Any given common disease is caused by a large number of Mendelian loci
  - Mutations at each locus have a relatively large effect
  - Each mutant allele is rare but all mutant alleles of all genes together give rise to a large number of affected individuals
  - Example thrombotic disease and Factor V Leiden

## The origins of common disease II.

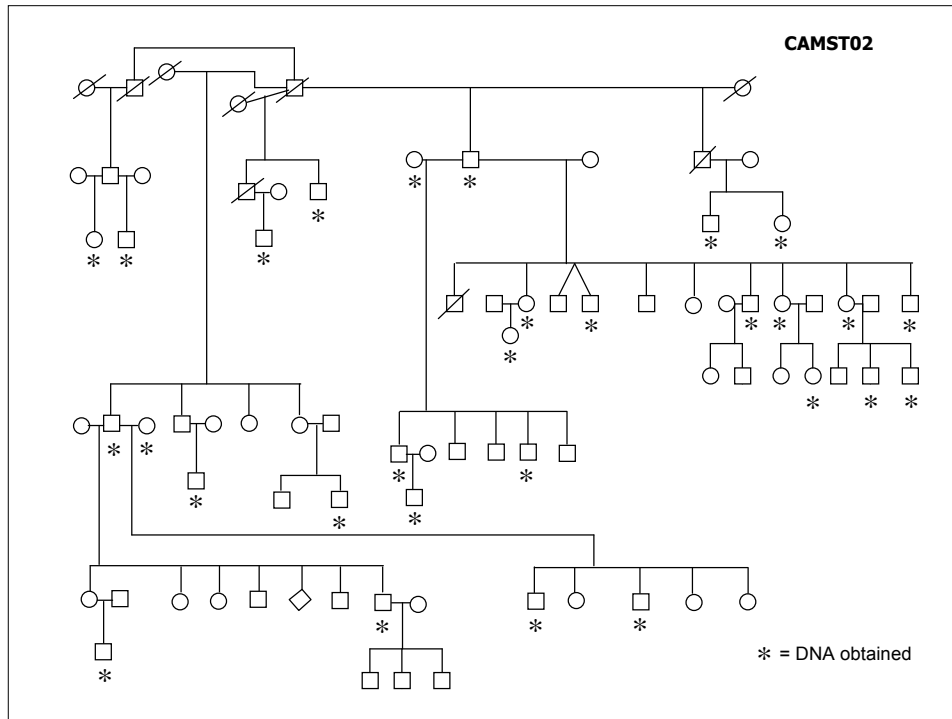
- Common variant - common disease hypothesis
  - Alleles that predispose individuals to common disease are common in the population
  - These alleles are of ancient origin
  - Each variant has a relatively small effect
  - The disease state is the additive or multiplicative result of the actions of multiple disease-predisposing alleles in an individual
  - Example: ApoE<sub>4</sub> and Alzheimer's disease

## The question remains unsettled

- Strategies are being developed to efficiently identify genes under both scenarios

## Rare variant - common disease hypothesis - strategies

- Find Mendelian forms of common disease
  - Familial Alzheimer's - presenilin loci
  - MODI - glukokinase, others
  - Stuttering



## Rare variant - common disease hypothesis - strategies

- Reduce genetic heterogeneity
  - Non-randomly mating populations
    - Polygamous populations
    - Inbred populations
  - Genetic isolates
    - Geographic isolates - islands, mountain villages
    - Cultural isolates - religion



## Common variant - common disease hypothesis - strategies

- Large sample sizes
- Association studies

## Association studies

- Association vs. linkage studies
  - Linkage = association within a family
  - Association studies employ affected individuals not within families
  - Association studies typically employ more individuals than linkage studies -  $10^2$ - $10^3$
- Association methods have been shown to be more powerful than linkage methods under a number of conditions that typically prevail in complex disease

## Association study caveat

- Linkage studies will find the location of the causative gene regardless of mutant allele
- Association studies require a common mutational origin in the study subjects
- Association studies WILL NOT find disease genes when a major cause of disease is repeated new mutations in the same gene

## Association study design

- Typical design is case-control
- Typical marker is SNP
- Typical measurement of association is chi-square test
- Criteria for significance are still being refined
  - Problem: correcting for large number of tests

## Association studies Mark I. - past

- Typically case-control
- $\geq 100$  cases,  $\geq 100$  controls
- Candidate genes
  - Chosen for biological/physiological/pathological relevance
  - Early studies focused on a single gene, often one SNP within or near that gene

## Association Studies I.

- Results - Most early SNP association studies were plagued by irreproducibility

## Association studies Mark II. - present

- Still case-control
- $\geq 1000$  cases,  $\geq 1000$  controls
- Recognition that cases and controls must be population matched to account for allele frequencies
- Recognition that haplotypes typically provide substantially more power to detect association than individual SNP's
- Candidate genes still used

## Association Studies Mark II.

- Genome-wide SNP association measurements
  - Enabled by technologies that reduce cost to \$.10 - \$.01 per genotype
  - Enabled by widespread SNP availability - dbSNP
  - 10K - 100K SNP's typed in each individual
- Reducing scale via sample pooling
  - Useful for screening
  - Not useful for statistical testing

## Association Studies Mark II.

- Tagging SNP's and the HapMap project
- Principle: Determine the number, location, and extent of all haplotype blocks. Type one marker each 20-40 kb in most populations, or one marker each 10-20 kb in African populations. Measure frequency of the haplotypes in cases and controls

## Association studies Mark II.

- Problems
  - Not all the human genome consists of such haplotype block structure. Some studies have shown that the more SNPs that are typed, the more haplotype blocks tend to become fragmented or disappear
  - arriving at reliable and robust statistical measures, accounting for huge number of multiple tests
- No successes yet

## Association studies Mark III. - future

- All SNP's are becoming known
- All functional elements in the genome are becoming known
- ~ 20% of the genome is comprised of:
  - Coding sequences
  - Micro RNA sequences
  - Evolutionarily conserved sequences

## Association studies Mark III.

- Principle: Perform large case-control studies with all SNP's residing in all the functional elements of the genome using very large study populations
  - 10,000 cases and 10,000 controls
- Additional statistical techniques will be required

## References

- Risch, NJ. Searching for genetic determinants in the new millennium. *Nature* 405:847-856 (2000)
- Lohmeuller, KE, Pearce, CL, Pike, M, Lander, ES, Hirschhorn, JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genetics* 33:177-182 (2003)
- Moore, JH and Ritchie, MD. The challenges of whole genome approaches to common diseases. *JAMA* 291:1642-1643 (2004)
- Carlson, CS, Eberle, MA, Kruglyak L, Nickerson D. Mapping complex disease loci in whole-genome association studies. *Nature* 429:446-452 (2004)

## References

[http://slack.ser.man.ac.uk/theory/association\\_chi.html](http://slack.ser.man.ac.uk/theory/association_chi.html)

- A simple and clear explanation of the use of the  $X^2$  (chi-square) test for the evaluation of case-control SNP association data