

# Strategies for finding disease genes

---

Dennis Drayna  
Section on Systems Biology of  
Communication Disorders  
LMG, NIDCD, NIH

## Why find disease genes?

---

- Provide understanding of pathophysiology of disease
- Provide understanding of biology of specific organs/systems
- Improve diagnosis
- Identify targets for improved therapeutics (drugs)

## How do we know a disease is due to genetic causes?

### Nature vs. Nurture

- Familial aggregation
- Twin studies
- Adoption studies
- Segregation analysis
  
- Heritability estimates
  - $h^2$
  - 0.0 – 1.0

## Positional cloning

- Identification of a gene solely on the basis of its position in the genome
- Begins with a linkage study
- Linkage identifies the approximate chromosomal region in which the disease-causing gene resides

## Genetic linkage

- Violations of Mendel's 2<sup>nd</sup> Law
  - Law of independent assortment
- Occurs because the genes that cause two different traits reside very close to each other on the same chromosome
- Can be observed over distances from 1 bp to ~30 million bp
- Now studied with genetic markers

## Genetic markers

- Simple sequence repeat markers - STRP
  - 2-,3-,4-nucleotide repeated sequence
  - (CA)<sub>n</sub> most common in genome
  - Tri- and tetra-nucleotide sequences most commonly used for genome wide linkage
  - Differences in the number of repeat units
  - Genotyping measures length differences
    - Typically many alleles segregating



## Genetic markers

---

- Single nucleotide polymorphisms-SNPs
- Differ in the base pair at a single position
  - Typically only 2 alleles segregating
- Assayed by a variety of technologies
- Focus on low cost/unit genotype



## A linkage study in practice

---


- Ascertain and enroll families
- Obtain DNA samples – 10's to 100's
- Obtain complete phenotype/diagnosis information
- Genotype
  - 400 STRP markers, many thousands of SNPs
- Perform statistical analysis



## Statistical analysis

---


- Parametric analyses
  - Traditional LOD scores – well accepted
  - Extract the maximum amount of linkage information from a pedigree
  - Requires precise specification of model
- Non-parametric analyses
  - Do not require specification of a model
  - p values, NPL scores



## Positional cloning – from linkage to gene

---

- Greatly aided by the Genome Projects
- Highly dependent on re-sequencing
- Goal – To find mutations in (or near) a gene that can be proven to cause disease
- Traditional gold standard criteria for identification of the disease gene is the observation of different mutations in the same gene in different families with the same disease



## Mendelian disorders – the success story

---

- Genes underlying all major human inherited disorders now identified
  - Neurofibromatosis
  - Cystic Fibrosis
  - Hemochromatosis
  - Muscular Dystrophy
- Opportunities for important discoveries based on Mendelian disorders still remain



But,

---

- Many mutations cannot be found, even in well-understood Mendelian diseases
  - Only 1 mutation identified in individuals with recessive disorders



## Gene identification - problems

---

- Mutations in or near a gene but outside the coding sequence
- Mutations in genes that are duplicated in the genome
- Copy number polymorphisms
  - 8% of genome
- Non-protein coding regions
  - Small RNA's



## Linkage studies of complex traits – into the swamp

---

- Complex traits are due to a combination of genetic and non-genetic causes
- Make up most of the major human health problems worldwide
- Heritabilities range from  $\sim 0.30$  to  $\sim 0.70$
- Multiple large linkage studies failed
  - Minimally significant linkage scores, poor replicability, no overlapping locations

## Linkage studies of complex traits – the problems

- Who's affected?
  - Dichotomous vs. continuous traits
  - Assignment of cut-off values
- Phenocopies
- Locus heterogeneity
- Incomplete penetrance
- Variable expressivity

## Two alternative hypotheses

- 1. Common disease-common variant hypothesis
  - Common diseases are caused by variants that are common in the population
    - $10^{-1}$
  - Each individually exerts a small effect
  - Probably evolutionarily ancient





## Two alternative hypotheses

---

- 2. Common diseases are caused by uncommon mutations of large effect
  - Present at low frequencies
    - $10^{-3}$  to  $10^{-2}$
  - Near Mendelian effects
  - Factor V Leiden and thrombosis



## Linkage studies of complex traits – new strategies

---

- Mendelian forms of otherwise non-Mendelian disorders
- Meta-analyses of linkage studies
- Model organisms
- Limiting genetic diversity

## Meta-analyses

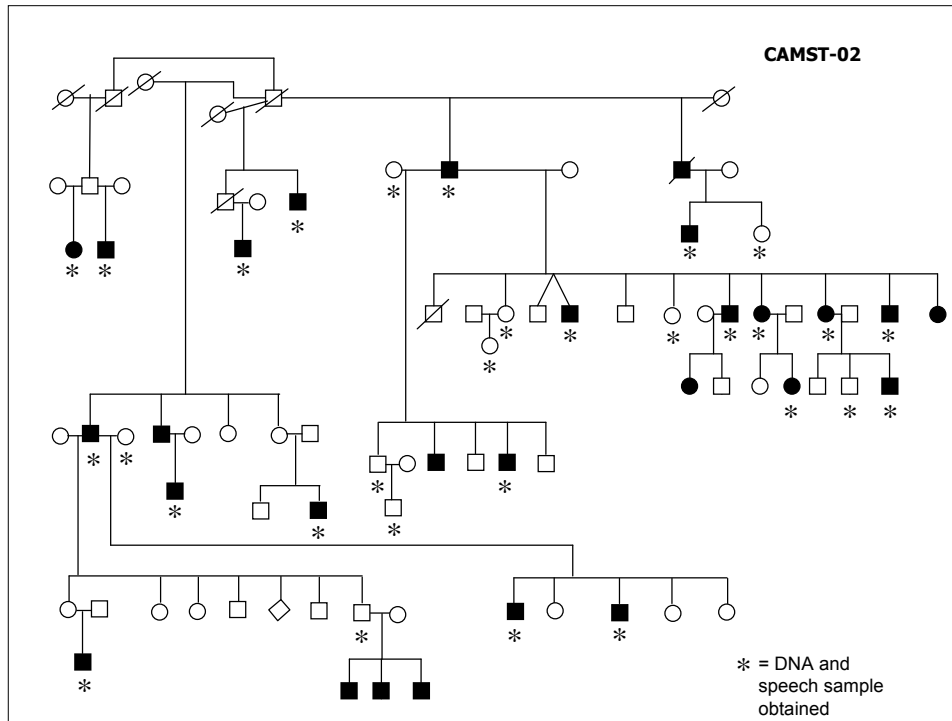
---

- Statistical analysis of statistical analyses
- What do many different linkage analyses, when analyzed as a group, tell us?
- Frequently find evidence of one (or a small number of) significant linkage loci
  - Chromosome 16 in Crohn's Disease
  - Chromosome 10 in macular degeneration

## Mendelian forms of otherwise non-Mendelian disorders

---

- MODY – Maturity onset diabetes of the young
  - Glukokinase gene
- Stuttering
  - West African kindreds

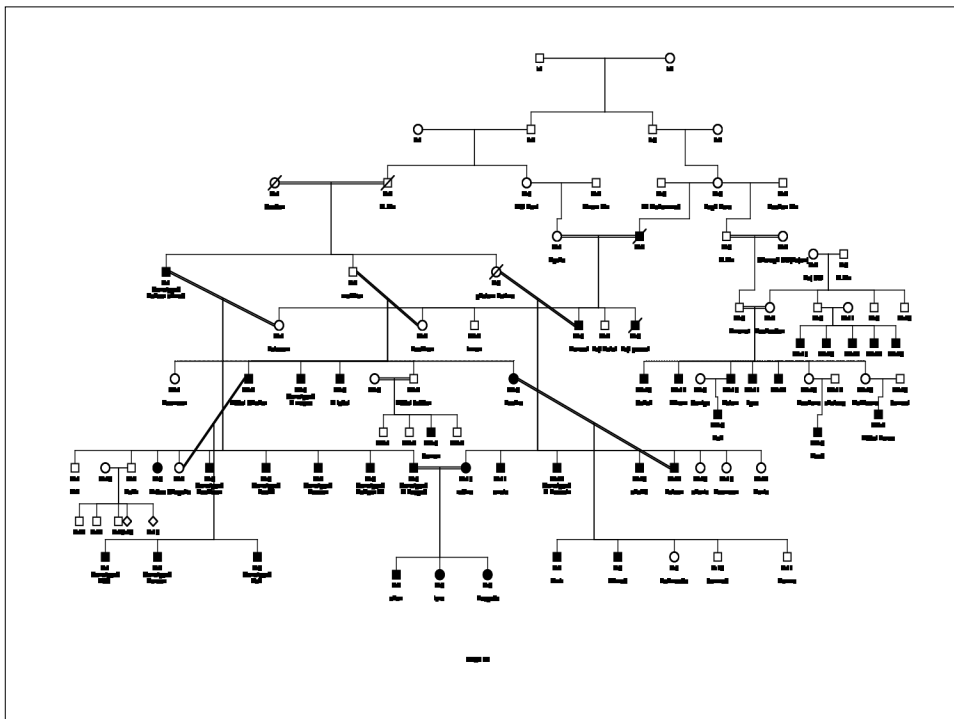


## Model organisms

- Mice
  - Can be bred to homogeneous lines in which typically complex traits segregate as Mendelian disorders
    - Ob mouse/leptin gene
  - Typically provide only one piece of the puzzle

## Limiting genetic diversity

- Reduced the number of genes (and loci) that cause the disease in the study population
- Inbred populations
  - Cousin marriages, polygamous marriages
- Geographic isolates
  - Islands, valleys
- Religious isolates, cultural isolates
  - Samaritans, Navajo, Amish



## But, linkage will likely never work for some traits

---

- Skip linkage and families altogether
- Perform genome-wide, population-based association studies

## Association studies

---

- Population based, rather than family based
- Typically case-control design
- Shown to be more powerful, in theory, than linkage for diseases caused by common variants with small effects



## Association studies

---

- Single gene association studies
  - The first try
- Disease-related gene association studies
  - Collections of genes
- Genome-wide association studies
  - Current state of the art



## Genome-wide association studies

---

- Large subject populations
  - 1000's of cases and controls
- Large numbers of markers
  - 500,000 or more
  - Affymetrix vs. Illumina technologies
- Statistical analysis
  - Under development
- Reliance on the HapMap



## The human HapMap

---

- Haplotype map
- Haplotypes
  - The arrangement of specific alleles along the length of one region of one chromosome
- Problems
  - Not all the human genome exists in such a block structure
  - Not all populations exhibit the same block structure



## Genome-wide association studies

---

- Statistical analysis
  - Assignment of fractions of the variance
  - Risk calculations
- Case-control design
- Methods under development

## Given uncertainties, two strategies

---

- Tagging SNPs
  - One SNP per haplotype block
- "Picket fence" SNPs
  - Evenly spaced

## Whole-genome association - problems

---

- Multiple test correction
- Fails for "new" mutations, mutation hotspots, multiple ( $> \sim 5$  ?) founder mutations, extremely old mutations
- Costly





## The hybrid strategy

---

- Find evidence for linkage
- Perform targeted case-control association study



## Example: Crohn's disease (IBD)

---

- 7 independent genome-wide linkage studies
  - 16 different loci identified
- Chromosome 16
- Association study across short arm
- SNPs in CARD15/NOD2 gene associated with disease



## Example: Macular degeneration

---

- Multiple genome-wide linkage scans
- Chromosome 10
- SNP association study in case-control
- Identified complement factor H
- Illuminated the role of inflammation in the disease



## Future trends

---

- Large-scale population association studies
  - 500,000 individuals in Britain
  - Longitudinal
  - Co-variates
- Individual genome sequencing
  - The \$1000 genome