Strategies for finding disease genes

Dennis Drayna, PhD
NIDCD/NIH

Current Topics in Genome Analysis
Spring, 2008

Purpose of the Genome Project

- Develop technologies
- Understand human biology
- Understand evolutionary human history

- Improve human health

Human Disease - Is it genetic?
Nature vs. nurture

- Many diseases (and other traits) run in families
- How do we determine the origin of such traits?

Concept of heritability

- h^2 - The fraction of variance due to genetic factors
- Ranges from 0 to 1
- Is distinguished from unique environment, common environment
  - Includes things as diverse as diet, education, geographic location

Estimating heritability

- Adoption studies
  - Open vs. closed adoption records
    - Scandinavia vs. U.S. vs. rest of the world
  - One-sided vs. two-sided analyses
    - Adoptive parents - adopted children
    - Biological parents - adopted children
    - Biological parents - biological children

Estimating heritability

- Segregation analysis
  - Compare occurrence in families to expected patterns under Mendelian models
  - Can provide support for mode of inheritance
  - Can make estimates of single major gene effects
  - Requires a large group of randomly ascertained families with the disorder

Estimating heritability

- Family studies
  - Estimate based on degree of genotype and phenotype sharing
    - First-degree relatives share half their genes
    - Second degree relatives share 1/4 of their genes

Estimating heritability

- Twin studies
  - Many study designs
  - Currently popular
    - Large twin cohorts developed
      - Twin Research Unit, St. Thomas' Hospital
      - FinnTwin
      - Others
  - Well-developed analytical methods
    - Structural equation modeling
    - Successive best-fit methods

Twin Studies

- Existing resources
  - Twin cohorts
  - Twinsburg, Ohio Twins Festival
    - Held annually in August
    - > 2000 pairs of twins attend
    - Research-friendly administration and environment

Heritability estimates from twin studies

- Stature (height)
- Psychiatric diseases
- Bone mass density
- Plasma cholesterol level
- Salt taste sensitivity

- Religiosity
- Social attitudes

Highest heritabilities

- Are observed for Mendelian disorders
- Even in Mendelian disorders, uncertainties arise from:
  - Reduced penetrance
    - Have the disease genotype but not the disease
  - Variable expressivity
    - Severe vs. mild (sometimes subtle) disease
  - Locus heterogeneity
    - Mutations in genes at more than one locus cause same disease
  - Allelic heterogeneity
    - Different mutations in the same gene cause different symptoms, or even different diseases

Mendelian Disorders

- An early confirmation of Mendel
  - Garrod - alkaptonuria
- Are typically considered rare by physicians
  - 10^{-4} - 10^{-6}
- But, can amount to a significant number of patients
  - >20,000 patients with Huntington Disease
- Traditionally a focus of pediatrics

Finding Mendelian disease genes

- Brought human genetics into the molecular realm - 1980s - 1990s
- The genes underlying the common Mendelian disorders now known
- Relied on Positional Cloning

Positional Cloning

- The ability to identify a gene causing a trait based solely on its position in the genome
- Information about the biochemistry, physiology, or pathology not required
- Agnostic regarding disease mechanism

The positional cloning process

1. Establish linkage to a marker at a known location
2. Search the linked region to identify all resident genes
3. Examine each of these genes to identify causative mutation(s)

Positional cloning shortcut

- Identify a cytogenetic abnormality associated with the disorder
  - Often very rare
- The disorder of interest can be present as part of a syndrome
- Test DNA probes (now DNA microarrays) across region of the abnormality to identify rearrangement in affected, but not in unaffected, individuals

1. Finding linkage

- Performed in families in which the disease is segregating
- Linkage occurs because the genes encoding the two traits reside in close proximity to each other
  - Based on violations of Mendel's Second Law (independent assortment)
- Uses genetic markers of known location
  - The phenotype used is naturally-occurring inherited variation in DNA sequence itself
  - Markers assembled into panels that optimize efficiency in the lab and in subsequent analysis

Getting started

- Assemble families
  - 10^1 - 10^2 individuals required
  - Obtain DNA - blood is traditional source, other sources gaining popularity
    - How much of the DNA is human?
  - Obtain clear, consistent, detailed phenotype information
  - Ongoing contact with families is essential

DNA-based genetic markers

- simple sequence repeats
  - Di-, tri-, and tetra-nucleotide repeats
  - differences based on length across repeat
  - gttatcttagagctcagtcacacacacacacacacacacatccaggtattggatcaact
  - single copy variable repeat single copy
  - Linkage panel contains ~400 markers

- SNPs - single nucleotide polymorphisms
  - ggattacctgaccctgAccgcttaatcattgatt
  - ggattacctgaccctgGccgcttaatcattgatt
  - Linkage panel contains 5,000-10,000 markers

Genotype individuals

- Microsatellites
  - Assayed by PCR followed by gel electrophoresis
    - Weber Marshfield panels
  - Utilizes same instruments as DNA sequencing
  - Many alleles at each locus - highly informative
- SNPs
  - Non-electrophoretic methods
    - Affymetrix, Illumina, Sequenom
  - Hybridization-based
  - Genotypes at many SNP sites gathered simultaneously
  - Only 2 alleles at each locus - often uninformative

Analyze data for linkage

- Parametric methods
  - LOD score method
    - Maximum likelihood estimation
    - used for Mendelian traits
    - Logarithm of the odds that: markers are linked, at a particular distance (θ), divided by the odds that they're linked at 50% co-inheritance, i.e., they're not linked at all
    - Classic LINKAGE package - MLINK, ILINK

LOD scores

- Historically defined
  - LOD of 1 = suggestive
  - LOD of 2 = probable
  - LOD of 3 = proof
- Each non-recombinant informative meiosis (= parent to offspring co-inheritance) contributes a LOD score of + 0.3
- LOD scores can also be negative, indicating a lack of linkage
  - LOD of (-) 2 is accepted as proof of non-linkage

Non-parametric methods

- Used for non-Mendelian traits
- Evaluate deviations from expected degree of allele-sharing in affected family members
- Typically applied to a large collection of affected relative pairs
- Support for linkage reported as p-values, NPL values, others
- GENEHUNTER, Allegro, Merlin, others

2. Search the linked region to identify all resident genes

- www.ncbi.nih.gov
  - NCBI (UCSC Genome Browser)
- ENCODE
- Large-scale genomic re-sequencing of target region

3. Examine each of these genes to identify causative mutation(s)

- Perform DNA sequencing to identify mutations which exist in affected individuals but not in normal individuals
- Gold standard proof:
  - No mutations in that gene observed in normal individuals
  - Different mutations observed in the same gene in different families with the same disease

Perspective

- The surprises - causative gene for many disorders was completely new…
  - Polycystic Kidney Disease
- Or completely unexpected
  - Hemochromatosis
- Take-home lesson: assigning genes as candidates based on knowledge of the biology of their gene products is of limited value

Does linkage have a future?

- All the common Caucasian Mendelian disease genes have now been identified
  - Cystic fibrosis, hemochromatosis, Muscular dystrophy, neurofibromatosis
- But, the genes for a very large number of rare disorders remain unidentified
  - The genes underlying less than half of the known Mendelian disorders in humans have been found
  - “Niche disorders” - although medically rare, can provide important insights into biology, both normal and pathologic
  - Deafness, Familial Mediterranean Fever

Linkage

- Very good at finding variant genes:
  - that have large effects
  - that show Mendelian inheritance
  - that contain mutations that are rare in the normal population

But, linkage has limitations

- From the perspective of clinical medicine, Mendelian disorders are not a significant portion of the total disease burden in the population
- Linkage has been spectacularly unsuccessful at identifying genes containing mutations that:
  - Confer small or moderate effects
  - Are common in the population
- These are the genes that underlie the important common diseases
  - Psychiatric disease
  - Metabolic disease
  - Cardiovascular disease

The current goal

- To realize the ultimate promise of the Genome Project, which is to help solve common diseases

A possible solution?

- Copy number variation (CNV)
- Approximately 10% of the genome exists in different copy number in different individuals
- CNV was invisible to traditional candidate gene evaluation methods
- Many efforts to examine the role of CNV in human disease currently underway

The more promising solution

- Association studies
- Shown to have theoretically greater power to detect disease genes when they:
  - Exert small effects
  - Are common in the population

Association studies

- Typically case-control design
  - Long history of use in medical research
- Typically enroll large numbers of subjects
  - 10^3 - 10^4
- Typically employ measurement of association at a large number of (SNP) loci
  - 10^5 - 10^6
  - Enabled by recent new technologies

Issues with association studies

- Studies are finding small risk factors
- Importance of a variant that explains 1% of the disorder - diabetes and the insulin gene
- Relative risk vs. population attributable risk
  - Individual risk may double, but the variant may account for only 2% of the risk in the population
  - “little diagnostic and no prognostic value”
- Requires shared ancestral variant
  - Recurrent mutation in the same gene will obliterate any association with neighboring SNPs
    - Would fail for achondroplasia, hemophilia