COMPLEX GENETIC DISEASES

Date: Sept 28, 2005*
Time: 9:30 am – 10:20 am*
Room: G-202 Biomolecular Building
Lecturer: David Threadgill
         4340 Biomolecular Building
         dwt@med.unc.edu
Office Hours: by appointment

*Please consult the online schedule for this course for the definitive date and time for this lecture.

Assigned Reading: This syllabus.

Key Concepts vs Supplementary Information: The syllabus is meant to accompany the lectures and the material covered (other than that in italics) is considered important to your understanding of the subject. Key items are marked in bold. Supplemental information is marked in italics.

Lecture Objectives: This lecture is to serve as an introduction to the genetic analysis of complex traits/diseases (those with polygenic inheritance). Emphasis is placed on the difference between Mendelian and polygenic traits. By the end of this lecture you should be familiar with (1) the types of traits and how their inheritance differs, (2) the principles of linkage analysis, (3) how these principles are applied to complex traits, and (4) the confounding problems of complex trait analysis.
Complex Genetic Diseases

Types of traits
The basic principle of genetics is the association of variation in phenotypic traits between individuals, such as height or disease susceptibility, with polymorphisms or mutations in specific genes. The type of trait and the structure of the population manifesting the variable phenotype must be taken into consideration when mapping the causative genes.

Complex or multifactorial inheritance - disease occurrences in families that do not match one of the simple Mendelian patterns of inheritance.

Discrete/qualitative trait - traits that are present or absent.

Continuous/quantitative trait - traits that have measurable characteristics across a range of values. This class includes the vast majority of diseases afflicting humans.

In addition to having differences in the mode of these traits, they are also associated with differences in the timing of their effects:

<table>
<thead>
<tr>
<th>Type</th>
<th>Incidence at Birth (per 1,000)</th>
<th>Prevalence at Age 25 Years (per 1,000)</th>
<th>Population Prevalence (per 1,000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diseases due to genome/chromosome mutations</td>
<td>6</td>
<td>1.8</td>
<td>3.8</td>
</tr>
<tr>
<td>Disease due to single gene mutations</td>
<td>10</td>
<td>3.6</td>
<td>20</td>
</tr>
<tr>
<td>Disease with multifactorial inheritance</td>
<td>~50</td>
<td>~50</td>
<td>~600</td>
</tr>
</tbody>
</table>
When genetically mapping genes causing specific traits, you also need to know whether the trait follows classical Mendelian inheritance or appears to be sporadic:

These two types of pedigrees will need very different approaches to identify the causative genes. Mendelian traits can use linkage analysis while ‘sporadic’ or non-Mendelian traits will need to have more complex approaches described below.
Linkage analysis

The foundation of linkage analysis is meiotic recombination that causes segregation of causative genes from surrounding DNA polymorphisms:

Linkage is used to identify a region of the genome (or a chromosome) that contains the causative gene. It is the first step in isolating the disease causing polymorphism or mutation. The subsequent steps of the molecular identification of the DNA differences has already been described in the molecular biology section:
The unit of measure in linkage is cM (centi Morgan). One cM equals 1% crossover.
--Crossovers represent the fraction of offspring with recombinations between markers to total offspring.
--The larger the measure, the farther apart two markers are.
--If markers are $\geq 50$cM, then they are unlinked.
--Linkage is a genetic distance, so two markers can be on the same chromosome but be unlinked.

For unlinked genes, the frequency of all combinations should be equal:

The example above is for two unlinked loci. The loci can be two molecular markers or one might be a disease trait and the other a molecular marker. For linked markers, the farther apart the markers are (A), the weaker the linkage (and the greater the genetic distance between them). Likewise, the stronger the linkage (B), the closer the two markers are to each other (and the lesser the genetic distance between them). Remember, when we talk about markers, this can be a molecular marker or a phenotypic trait:
In the example above, markers in A are not linked, even though they are on the same chromosome. Markers in this example are syntenic (on the same chromosome) but not linked. Those in B are tightly linked and in C linked, but farther apart.

How can two markers on the same chromosome be not linked?

This is because once genes are greater than 50cM apart, they have 50% recombination making them appear like they are on different chromosomes. Furthermore, double recombinations can occur, but this still maintains the distribution of recombinants to non-recombinants.
Example:

Let’s assume we are genetically mapping the gene for a recessive trait. We have one large pedigree with 20 offspring from two non-affected parents. Five offspring are affected. We genotype the individuals for two markers that are on different chromosomes and get these genotypes:

```
Marker 1: aa aa Aa Aa aa Aa AA AA Aa aa AA Aa aa AA AA AA Aa
Marker 2: bb Bb BB bb BB Bb bb BB Bb Bb Bb BB bb Bb Bb Bb BB bb Bb
```

Is either of these markers linked to the recessive gene? If so, what is the genetic distance between the linked marker and the disease causing gene?

First, determine the number of recombinant (R) between two markers and the number of non-recombinants (NR). Our hypothesis is that b is linked to the disease causing allele:

```
Marker 2: bb Bb BB bb BB Bb bb BB Bb Bb Bb BB bb Bb Bb Bb BB bb Bb
R  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X
NR  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X  X
```

10 recombinants/20 total offspring = 50% crossovers = 50cM

Therefore, b is not linked to the disease since linkage is ≥ 50cM
What about a?

As a second example, consider two markers at opposite ends of a chromosome carrying the autosomal dominant mutation for retinitis pigmentosa:
**Multipoint linkage analysis** can be used to establish marker order by minimizing the number of apparent multiple crossovers.

By comparing the frequency of genotypes with all possible marker orders, a marker order can be determined that minimizes the double recombinants that are needed to explain the results. Multipoint linkage analysis is important in very complex mapping studies that usually involve dozens of marker loci. Maps created by this approach are also very important as frameworks to provide diagnostic information for genetic counseling where causative genes (particularly those associated with multigenic traits are not known) have not been identified but there approximate location in the genome is known.
For example, by knowing the marker order, it is possible to predict whether the offspring will be affected or not with NF1.

Approaches for complex trait mapping
Linkage studies as described above are best applied to simple Mendelian traits or traits with high penetrance and expressivity.

**Penetrance**: the proportion of individuals of a specified genotype that manifest a particular phenotype or disease.

**Expressivity**: the range of phenotypes manifested by individuals with a given genotype.

**Polygenic traits**, those that have many genes contributing to them where no one gene causes the disease or phenotype, require alternative approaches. These types of diseases are typically detected through familial clustering.

A primary characteristic of diseases with complex inheritance is that affected individuals tend to cluster in families (**familial aggregation**). This is because relatives share a greater proportion of their alleles than with unrelated individuals in the population.

The reverse is not necessarily true - familial aggregation of a disease does not mean that a disease must have a genetic contribution. Non-genetic factors could have the same effect - besides sharing alleles, families share culture, behavior, diet and environmental exposures.

**Concordant** – when two individuals in a family have the same disease. Affected individuals may not share the same pre-disposing alleles, in which case they are phenocopies (**genetic heterogeneity**).
**Discordant** - when some family members have a disease and others not. Discordancy could be due to one individual having the disease allele and the other not. Alternatively, the disease may have incomplete penetrance.

Phenocopies and incomplete penetrance contribute to the difficulty of identifying genes that are responsible for polygenic and multifactorial diseases.

**Case-control studies** - An approach to assess familial aggregation (and the basis for association studies described below) is to compare patients with a disease (the cases) with suitably chosen individuals without the disease (the controls) with respect to family history of disease (it is also important to randomize other variables such as environmental exposures, occupation, geographical location, parity, previous illnesses, ethnicity, etc).

To assess a possible genetic contribution to familial aggregation of a disease, the frequency with which the disease is found in the extended families of the cases (positive family history) is compared with the frequency of positive family history among suitable controls that do not have the disease.

These studies are subject to many different kinds of errors, the most troublesome being ascertainment bias.

**Ascertainment bias** - a difference in the likelihood that affected relatives of the cases will be reported to the epidemiologist as compared with the affected relatives of controls.

Choice of controls is also a confounding factor. Controls should differ from the cases only in their disease status and not in any other variable.

This is also why it is important to known the population structure as described previously so that appropriate controls can be selected for the cases.

Associations found in case-control studies do not prove causation. Many variables are not independent of each other.

**Separating Genetic and Environmental Factors** - Disease concordance and allele sharing among relatives can be used to dissect the relative contribution of genetics and environment to disease. When genetics is important, the frequency of disease concordance increases as the degree of relatedness increases.

The most extreme examples of two individuals having alleles in common are identical (monozygotic-MZ) twins. The next most closely related are first degree relatives, including fraternal (dizygotic-DZ) twins.
Two major approaches have been developed to map genes associated with complex diseases (with many variants of these for special disease models).

**Affected pedigree method** - type of linkage analysis that relies on pairs of family members who are concordant for the trait. Sibling-sibling or parent-sibling pairs are based upon the concept of identity-by-descent. The pairs share stretches of DNA in common through a direct inherited lineage.
**Association method** - is based upon the presumption that there is an increased frequency of particular alleles in affected compared with unaffected individuals in the population (in essence a case control study).

Origin of **linkage disequilibrium** used in association studies:

Limitations of complex trait analysis:
Complex epistatic interactions—many genes, usually with only a small effect individually, contribute to complex traits (heart disease, cancer susceptibility, diabetes, etc) but cumulatively, they often have unusual and unpredictable interactions that are greater than the sum of their individual effects.
**Genetic heterogeneity** - multiple different genetic differences between individuals may contribute to similar phenotypes complicating the association between genotype and phenotype.

**Pleiotropy** - individual genes may have multiple effects and may contribute to many different traits.
History of mutations/polymorphisms-genetic differences may be associated with multiple linked marker combinations or have arisen multiple times during past generation.

**Population stratification** -unknown population structures will have a significant impact on results, particularly case control (association) studies, if they are not properly controlled.