How do we identify genes conferring probabilistic propensities?

**genome-wide linkage studies**
- establish statistically significant genome-wide evidence for linkage between a trait or disease state and a specific chromosomal location
- *apply the positional cloning strategy to a complex trait*

**genome-wide association studies** (GWAS)
- establish significant genome-wide evidence for a statistical association between a particular SNP and a disease state or trait
GWAS studies are based on the common disease-common variant hypothesis

- susceptibility factors have ancient origins
- common illnesses stem from the additive or multiplicative effects of combinations of common variants.
- in this model, each “risk variant” is postulated to confer only a small degree of risk, with no one variant sufficient to cause the disorder. Disease onset is postulated as the result of the combined effects of many such alleles.

Genome-wide association study: involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease or a particular trait.

What research tools are available?

- computerized databases that contain the reference human genome sequence
- a map of human genetic variation (International HapMap Project)
- technologies that can quickly and accurately analyze whole-genome sample for genetic variations that can or may contribute to a complex trait or disease susceptibility.
**How are genome-wide association studies conducted?**

- To carry out a genome-wide association study, researchers use **two groups of participants: people with the disease/trait being studied and similar people without the disease/trait.**
- Each participant's genome is scanned for strategically selected SNPs – for example, **representative SNPs for each haplotype block need to be selected**
- If certain SNPs are found to be significantly more frequent in people with the disease compared to people without disease, the variations are said to be **"associated"** with the disease.
- The associated genetic variations can serve as powerful pointers to the region of the human genome where the disease-causing problem resides.
- **HOWEVER** the associated variants themselves may not directly cause the disease. They may just be tagging along, that is **linked to**, the actual causal variants. For this reason, researchers often need to take additional steps, such as sequencing DNA base pairs in that particular region of the genome, to identify the exact genetic change involved in the disease/trait.
The genomewide association study is typically based on a case–control design in which SNPs across the human genome are genotyped. Panel A depicts a small locus on chromosome 9, and thus a very small fragment of the genome. In Panel B, the strength of association between each SNP and disease is calculated on the basis of the prevalence of each SNP in cases and controls. In this example, SNPs 1 and 2 on chromosome 9 are associated with disease, with P values of $10^{-12}$ and $10^{-8}$, respectively. The plot in Panel C shows the P values for all genotyped SNPs that have survived a quality-control screen, with each chromosome shown in a different color. The results implicate a locus on chromosome 9, marked by SNPs 1 and 2, which are adjacent to each other (graph at right), and other neighboring SNPs.

*Genomewide Association Studies and Assessment of the Risk of Disease*


*July 8, 2010*
Need to look at haplotypes and linkage disequilibrium again

In a population of individuals, if the association between alleles at two linked loci is random, then the two loci are said to be in linkage equilibrium.

If the association is non-random, then the two loci are said to be in linkage disequilibrium – in this case a particular allele at one locus is associated with a specific allele at the second locus more often than expected by chance.

Figure 18-17
Introduction to Genetic Analysis, Tenth Edition
© 2012 W. H. Freeman and Company
**Haplotype**: combination of specific alleles that are inherited together

**Representative SNPs for each haplotype block need to be selected**

Diagram of the distribution of SNPs and haplotypes for a chromosomal segment. **HAPLOTYPES** often occur in **BLOCKS** (regions of lower recombination) separated from one another by recombinational hotspots.

Determine whether two SNPs show disequilibrium by noting the color of the square where the rows for the markers intersect.

Within a **HAPLOTYPE BLOCK**, SNPs show strong linkage disequilibrium. SNPs in different haplotype blocks show weak or no linkage disequilibrium (are in linkage equilibrium).
The HapMap is a catalog of common genetic variants that occur in human beings. It describes what these variants are, where they occur in our DNA, and how they are distributed among people within populations and among populations in different parts of the world.

Every chromosome carries a unique combination of alleles that is known as a haplotype.

- However, within regions of about 500 kb and less it is possible to find combinations of SNPs that are found in multiple unrelated individuals.
- Such “blocks” point to regions that have not been broken up by recombination and are often separated from each other by short regions where there is evidence for considerable recombination (recombination hotspots).
- These observations suggest the human genome is a colorful mosaic of haplotype blocks delimited by recombination hotspots.
- Moreover, genome-wide association mapping using SNP-Chips means we can find regions that segregate with disease alleles in unrelated individuals.
Figure 2: The construction of the HapMap occurs in three steps. (a) Single nucleotide polymorphisms (SNPs) are identified in DNA samples from multiple individuals. (b) Adjacent SNPs that are inherited together are compiled into "haplotypes." (c) "Tag" SNPs within haplotypes are identified that uniquely identify those haplotypes. By genotyping the three tag SNPs shown in this figure, researchers can identify which of the four haplotypes shown here are present in each individual.

Example from your text: GWA study for adult-onset type 2 diabetes

Collect DNA samples from 4000 individuals:
- 2000 individuals with type 2 diabetes
- 2000 matched controls

Genotype each DNA sample for 300,000 SNPs that are distributed across the entire genome
- each of the haplotype blocks in the genome should be marked by one or more SNPs

<table>
<thead>
<tr>
<th>Table 19-8</th>
<th>Part of a Simulated Data Set for an Association-Mapping Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>SNP1</td>
</tr>
<tr>
<td>1</td>
<td>C/C</td>
</tr>
<tr>
<td>2</td>
<td>C/C</td>
</tr>
<tr>
<td>3</td>
<td>C/G</td>
</tr>
<tr>
<td>4</td>
<td>C/G</td>
</tr>
<tr>
<td>5</td>
<td>C/C</td>
</tr>
<tr>
<td>6</td>
<td>G/G</td>
</tr>
<tr>
<td>7</td>
<td>G/G</td>
</tr>
<tr>
<td>8</td>
<td>C/G</td>
</tr>
<tr>
<td>9</td>
<td>C/G</td>
</tr>
<tr>
<td>10</td>
<td>G/G</td>
</tr>
</tbody>
</table>

Such a study would result in 12 billion data points: to the left are some sample data

Perform a statistical test on each SNP locus to determine if one of its alleles is more frequently associated with diabetes than would be expected by chance
\[ P = \text{probability that the observed deviation in SNP allele frequency between the control and study groups is due to chance} \quad (\text{null hypothesis is that SNP is NOT associated with the trait}) \]

SNP 1 = either G or C
SNP 2 = either G or C

Example calculation illustrating the methodology of a case-control GWA study. The allele counts of each measured SNPs is evaluated, in this case with a chi-squared test, in order to identify variants associated with the trait in question. The numbers in this example are taken from a 2007 study of coronary artery disease (CAD) which showed that the individuals with the G-allele of SNP1 (rs1333049) were overrepresented amongst CAD-patients.

http://en.wikipedia.org/wiki/Genome-wide_association_study
A Manhattan plot is a type of scatter plot, usually used to display data with a large number of data-points.

In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, with the negative logarithm of the association P-value for each single nucleotide polymorphism displayed on the Y-axis. Because the strongest associations have the smallest P-values (e.g., $10^{-15}$), their negative logarithms will be the greatest (e.g., 15). [http://en.wikipedia.org/wiki/Manhattan_plot](http://en.wikipedia.org/wiki/Manhattan_plot)

An idealized Manhattan plot depicting several strongly associated risk loci for a hypothetical disease/trait
The statistical test indicates that a specific allele of an SNP is associated with a specific phenotype (that is, different genotypes at a specific SNP are associated with different phenotype)?

WHAT NEXT?
Association mapping does not prove that a gene or an SNP within a gene affects a trait

It only provides statistical evidence for the association

A molecular characterization of the gene or site tagged by the SNP is required
What types of predisposing alleles are we likely to see in GWAS studies?

The **common disease-common variant hypothesis** proposes that susceptibility factors have ancient origins. In this model, each “risk variant” is postulated to confer only a small degree of risk, with no one variant sufficient to cause the disorder. Disease onset is postulated as the result of the combined effects of many such alleles.

The **mutation-selection hypothesis** suggests that a heterogeneous collection of recent mutations accounts for most disease susceptibility.
(A) The oldest human alleles originated in Africa well before the diasporas of modern humans 50,000–60,000 years ago. These oldest alleles are common in all populations worldwide. Approximately 90% of the variability in allele frequencies is of this sort.

(B) Origins of common and rare alleles. KYA refers to “thousand years ago.” Horizontal arrows suggest continuing cross-migration between continental populations. Development of agriculture in the past 10,000 years and of urbanization and industrialization in the past 700 years has led to rapid populations growth and therefore to the appearance of vast numbers of new alleles, each individually rare and specific to one population or even to one family.

Cell 141, April 16, 2010
## Comparison of common disease-common variant hypothesis & the mutation-selection hypothesis

**NOTE:** these are not mutually exclusive scenarios

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Common disease-common variant hypothesis</th>
<th>Mutation-selection hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequencies of susceptibility alleles</td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>Effect size of susceptibility alleles</td>
<td>small</td>
<td>moderate</td>
</tr>
<tr>
<td>Locus heterogeneity (number of susceptibility loci for a given disease)</td>
<td>high</td>
<td>could be low</td>
</tr>
<tr>
<td>Allele heterogeneity (number of different susceptibility alleles at a given locus)</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td>Origin of susceptibility alleles</td>
<td>ancient common ancestor</td>
<td>relatively recent mutations</td>
</tr>
<tr>
<td>Technology to detect susceptibility alleles</td>
<td>association studies</td>
<td></td>
</tr>
</tbody>
</table>

from 4th edition of *Human Molecular Genetics*
Figure 1: This diagram shows two ancestral chromosomes being scrambled through recombination over many generations to yield different descendant chromosomes. If a genetic variant marked by the A on the ancestral chromosome increases the risk of a particular disease, the two individuals in the current generation who inherit that part of the ancestral chromosome will be at increased risk. Adjacent to the variant marked by the A are many SNPs that can be used to identify the location of the variant.

What are we likely to see in GWAS studies?

Genome Wide Association Studies (GWAS)   NRG  May 2008
ASO Hybridization-based approach for SNP genotyping

- Short synthetic allele-specific DNA probes are anchored to a solid surface (called SNP targets)
- A whole-genome PCR amplification with fluorescently-tagged primers is performed for each study subject – entire genome is represented by PCR products tagged with the B dye (below)
- The PCR products are allowed to hydrogen-bond (anneal or hybridize) to the probes under reaction conditions that destabilize interactions with even one mismatch
- The solid surface containing the probes is washed to remove mismatched targets while perfectly matched target-probe pairs are detected by fluorescence
SCALING UP

Scoring polymorphisms on a genome–wide scale using gene chips or SNP array technology

Genome-Wide Human SNP Array 6.0

The Genome-Wide Human SNP Array 6.0 features more than 1.8 million markers of genetic variation, including single nucleotide polymorphisms (SNPs) as well as probes for the detection of copy number variation. The SNP Array 6.0 allows researchers to perform association studies with large sample sizes in both initial scan and replication phases, thereby significantly increasing the overall genetic power of their studies.
USING DNA CHIPS to genotype each study subjects SNPs

Ordered Array of ASOs

Allele-specific oligonucleotide SNP probes synthesized directly on the solid surface at high density

...over a million different ASOs and controls can be gridded per cm².

http://www.youtube.com/watch?v=1_wDrqgS8w8&NR=1&feature=endscreen
http://www.youtube.com/watch?v=ui4BOtwJEXs&NR=1&feature=endscreen
http://www.youtube.com/watch?feature=endscreen&NR=1&v=3jX_08zdYCE