31-33. Mapping disease genes in humans
Reading 408-421

- One of the major goals in modern genetics is understanding the genetic basis of phenotypic differences. The first step is locating genes that carry alleles responsible for differences in phenoype, i. e. mapping them. In chapter 5 you learned map genes using controlled crosses in a Mendelian experimental design. The basic idea of mapping autosomal genes is described in Fig. 5.4. You start with pure breeding parental lines, one homozygous for *b* and one homozygous for *c*. The F_1 is doubly heterozygous. Furthermore, you know the **haplotype phase** of every F_1 individual: *b* and c^+ are on one chromosome and b^+ and *c* are on the other. With two loci, there are two haplotype phases. The experimental design tells you which one you have. In the testcross, the phenotype of an offspring indicates the haplotype phase of its maternal chromosome. Because of the experimental design, each offspring informs you about the result of meiosis. Another way of saying this is that there was an **informative meiosis** in the mother. In controlled crosses of this type, the number of offspring in the test cross is the number of informative meioses, from which you estimate the recombination frequency (RF).
- The sample size is needed to provide accurate estimates of RF is **inversely proportional to RF**. Roughly speaking, you need 10/RF informative meioses to give a reasonably accurate estimate. In other words, the degree of accuracy with which you can map a trait is limited by your sample size. If your sample size is 100, you cannot expect to estimate RFs smaller than about 0.05 or 5 cM.
- When it is not possible to do controlled crosses, it may be possible to infer haplotype phase. For example, if an AABB father marries and aabb mother, the phase of the offspring has to be AB/ab, just as in a Mendelian experiment. Even if the mother were AaBb, a doubly heterozygous offspring would still have to have phase AB/ab. In doing this kind of analysis, you assume the loci are closely enough linked that there is no crossing over in the parental meioses.
- In some cases, mapping of genes causing some human disease is possible because the protein responsible for the disease is known. Biochemical analysis showed that a common form of hemophilia is caused by an absence of Factor VIII (Fig. 11.6). The amino acid sequence was then used to obtain the DNA sequence which was then used to probe a **genomic library**, which was then sequenced in cases to determine the particular sequence error.
- The problem of **linkage mapping in humans** is difficult because you cannot perform controlled crosses. You take the pedigree and marker loci you have and try to identify informative meioses. An example is given in the box on pp. 416-7 [Note that M1 and M2 in the box are different alleles. M1 and M2 in the text on p. 390 are different loci.]. The result from the analysis of a single family and a single marker locus is a **LOD** score. LOD from unrelated families scores are added. The goal of such a study is to find one (or more) marker loci for which the LOD score is high enough to indicate genetic linkage, and to indicate how far the causative locus is from the marker locus.

- The first step of a mapping study is usually done with ~300 SSR markers with an average spacing of 10 cM. With 10 cM spacing of markers, an overall LOD score of 3 is often taken as good evidence of linkage to a particular marker. If significant linkage is found, the next step is to screen additional markers in region with a goal of getting 1 cM resolution (i. e. within 1 mb). Figure 11.18a shows an example illustrating the linkage of alleles at marker locus G8 at the tip of 4p to Huntington's disease, which is a dominant condition.
- The average gene density is roughly 10 genes per mb. After you map a gene to within a 1 mg region, you try to identify **candidate genes** by using what is known about the types of genes in the region and their expression patterns. You would then sequence the candidate genes to determine whether there is a perfect association between the condition and variant sequence.
- How well this procedure works depends on what nature provides you. Huntington's disease (HD) was especially simple because the mode of inheritance (Mendelian dominance) was known and there is a single site, the location of an SSR that is perfectly associated with the HD. The phenotype indicates whether there is a mutant at that site in the causative locus. It is no accident that locus that causes HD was the first disease-associated locus found by this method. In that study RFLPs, not microsatellites were used as markers.
- For other Mendelian diseases, **genetic heterogeneity**, meaning that different variant cause the disease in different families. A typical and important example is the phenylalanine hydroxylase gene (PAH) at 12q23.2. Mutants of this gene are associated with phenylketonuria (PKU). Individuals homozygous for mutants of PAH cannot metabolize phenylalanine. Unless phenylalanine is removed from their diet, they develop severe mental retardation. There are more than 500 mutations of PAH known, http://www.pahdb.mcgill.ca/. **Compound heterozygotes** have PKU. The frequency of all together is ~1% in European and Asian populations and 0.1% in sub-Saharan African populations. Different alleles are common in different parts of the world.
- In PAH, there is evidence of **recurrent mutation**, based on the observation that the same mutant is found on chromosomes with different haplotypes. The basic idea is that a mutant arises on a chromosome carrying particular alleles at closely linked loci. That forms the ancestral haplotype. Recombination will not break up the ancestral haplotype at loci that are closely enough linked. For example, the mutant R408W is found on chromosomes with two different haplotypes. Haplotype 1 has allele 8 at a VNTR and 244 bp allele at an SSR; haplotype 2 has allele 3 and the 240 bp SSR. PAH was mapped in the same way that Factor VIII was. The amino acid sequence of the protein was used to probe the genome.
- Cystic fibrosis (CF) is an intermediate case. In people with European ancestry, it is the most common Mendelian disease, with a frequency of 1/2500 live births. CF is much less common in other ethnic groups in the US (1/8000-1/9000 in Hispanics, 1/15,000 in African-Americans, and 1/30,000 in Asians). Roughly 70% of the cases Europeans are caused by a single allele, ΔF508, which is a 3 base in-frame deletion that results in the loss of the 508th amino acid, a phenylalanine (F). But there at least 1603 other mutations of CFTR known to cause CF. They are of all types and found in all exons

and in the promotor and 5′ UTR regions. (http://www.genet.sickkids.on.ca/cftr/StatisticsPage.html).

- Other ways of mapping relies on the slow breaking down of the **ancestral haplotype** that carried the first copy of mutant that causes the disease. The mutant tends to remain associated to alleles at closely linked loci on the ancestral haplotype. The association between alleles at different loci on a chromosome is called **linkage disequilibrium (LD).** The method this association is called **LD mapping**. It works well in relatively isolated populations that were founded by a few individuals because it is likely that there was only one disease allele among the ancestors. One of the first examples of LD mapping was by Hästbacka et al. (1992). They mapped the gene causing diastrophic dysplasia (DTD) (an autosomal recessive disorder causing short stature and dysplasia of the joints) in Finland. It is a rare condition that is more common in Finland than elsewhere. A gene associated with DTD had already been mapped to 5q31. Two marker genes were found that had the following haplotypes on normal and DTD chromosomes: Normal (1-1 (4); 1-2 (28); 2-1 (7); 2-2 (84)) DTD (1-1 (144); 1-2 (1); 2-1 (0); 2-2 (7). The idea is that there was a founder effect when ancestors of the current population of Finland arrive about 2000 years ago (100 generations). One founded carried the DTD gene on a chromosome with the 1-1 haplotype at these two markers. This haplotype is very rare on normal chromosomes. The probability of 7 recombination events on 152 chromosomes in 100 generations is 0.00046, so the causative locus can estimated to be about 0.046 cM from the first marker locus. If 1 cM=1 mb, this translates to about 46,000 bases. It was found later to be about 70 kb away.
- Most single-gene or Mendelian diseases in humans have been mapped and the causative genes have been cloned and sequenced. One of the main challenges in human genetics is finding genes associated with **complex inherited diseases**, i. e. diseases that have a substantial genetic component but that do not have a simple Mendelian pattern of inheritance. Various types of heart disease, most cancers, diabetes, autoimmune diseases such as multiple sclerosis, and psychiatric diseases such as schizophrenia and bipolar disorder are all complex inherited diseases. They are the leading causes of mortality in developed countries and create the heaviest burden on health care resources in developed countries.
- Evidence for a genetic component for complex diseases comes from studies of disease risk in families. Empirical estimates of risk are obtained from relatives of a **proband**, i. e. an affected individual. Schizophrenia is a complex inherited disease with **prevalence** in the US of about 0.83%. The **recurrence risk** is the risk to a relative of the **proband**. The **risk ratio** (λ) is the ratio of the recurrence risk to the prevalence, and indicates the proportional increase in risk to relatives. For schizophrenia, the λ values are as follows: MZ 52.1 (44.3%), DZ 14.2 (12.7%), OF 10.0 (8.5%), FS 8.6 (7.3%), HS 3.5 (3%), N 3.1 (3%), G 3.3 (3%), C 1.8 (1.5%).
- The principle method for mapping genes affecting complex diseases is **case-control study**. The idea is simple. You compare the frequency of a marker allele in unrelated individuals with a disease (cases) with the frequency in unrelated individuals (controls). Cases and controls have to be matched in other ways, including age, sex,

and ethnic background. If you find a significant difference in frequency, you say there is a significant association. The marker may be causative itself (but probably not) or it may be closely linked to the causative allele because it is in LD with it.

- Suppose you have 500 cases and 500 controls, and you find the genotype an A/T SNP. You find 657 As in the cases and 543 As in the controls. You would like to know if this difference is significant, so you compute a chi-square statistic, which is 27.075. Table 5.1 tells you that with 1 df, P<0.001. The actual value is P=0.00023. If you had tested only 1 marker, you would say there is a significant association. The problem is that you do not pick one SNP to test but many. In recent **genome-wide association studies (GWAS)** commercially available chips determine the genotypes at 500,000 SNPs. The way currently used to test for significant association is to scan a large number of SNPs in a small group cases and controls. Then chose a list of SNPs highly associated with the disease and test them is a new and large group. Then do it again. The idea is to find SNPs that are consistently associated with the disease. Once a SNP is identified in one such GWAS, it will be tested on other groups. If it is found to be significantly associated, then there is much greater confidence that the association is not accidental.
- One of several successful GWAS is by Easton et al. who looked for additional genes associated with familial breast cancer, meaning cancer found in two or more close relatives. BRCA1 and BRCA2 were already known to cause familial breast cancer but the two loci together account for breast cancer in only about 25% of affected families. Easton et al. did their study in 3 stages. In stage 1, they compared the genotypes of 227,876 SNPs in 390 cases and 364 controls. In stage 2, they picked the 12,711 SNPs that showed the most significant differences in stage 1 and typed them in 3990 cases and 3916 controls. In stage 3, they took the 20 most significant SNPs and typed them in a total of 21,860 cases and 22,578 controls. They found significant evidence of association with 6 SNPs. Five are close to known genes, including genes cell growth and cell signaling which are important in the development of cancer. One, however, rs13281615, is not near any known gene. This study does not prove the involvement of these genes in breast cancer but they are being subject to intensive further study. Still, they account for only 3.6% of the risk of familial breast cancer. This is an example of what has been called the “problem of missing heritability.”

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nature Genetics 2:204-11
http://www.nature.com/ng/journal/v2/n3/abs/ng1192-204.html

Easton DF, Pooley KA, Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, Struewing JP et al. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087-1093
http://www.nature.com/nature/journal/v447/n7148/abs/nature05887.html

Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18-21. http://www.nature.com/news/2008/081105/full/456018a.html

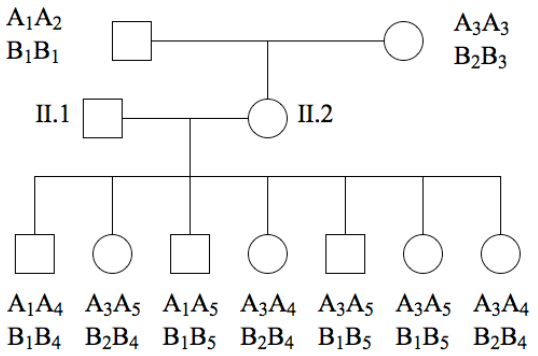
Problems: Ch. 11—I, 20-24, 28, 29, 32.

31.1. In many genetic studies, it is important to know the haplotype phase of individuals who are heterozygous at two or more loci. In human populations, the haplotype phase of an individual can sometimes be inferred if the genotypes of both parents is known. For each of the two following families, find the phase of the offspring and the haplotype contributed by each parent. Assume that mutation has not occurred and that the loci are autosomal. [Hints: Write a list of the pairs of haplotypes that the mother and father can produce. For part a, there is a single answer. For b, there is more than one answer.]

Mother	Father	Offspring
a. AaBB	aaBb	AaBb
b. AABbcc	aaBbCc	AaBbCc

Answer

a. Mother's gametes are AB and aB. Father's gametes are aB and ab. Only AB from the mother and ab from the father can create the double heterozygote so the phase is AB/ab.
b. Mother's gametes are ABc and Abc. Father's gametes are aBC, aBc, abC and abc. The A has to come from the mother. The C has to come from the father. But you cannot determine whether the B comes from the mother or the father. One possibility is ABc/abC and the other is Abc/aBC.



31.2 In the above pedigree, you are given the genotypes but not the haplotype phases of two codominant genes in grandparents and their 7 grandchildren, but not of their daughter (II.2) or her husband (II.1). [This problem is from last year's midterm.]

a. What is the genotype of II.1?

Ans. $A_4A_5B_4B_5$

b. What is the genotype and haplotype phase of II.2?

Ans. A_1B_1 / A_3B_2

c. How many of the grandchildren have the parental haplotype from II.2?

Ans. 5

d. What is the odds ratio in favor the hypothesis that RF=2/7 compared to the hypothesis that RF=1/2?

Ans. odds ratio= $\left(\frac{5}{7}\right)^5\left(\frac{2}{7}\right)^2 \Big/ \left(\frac{1}{2}\right)^7$