Questions and Answers for Genetics and Genomics in Medicine Chapter 8

Question 1

Candidate gene approaches have allowed the identification of human disease genes on the basis of prior information about the cause of a related phenotype. Give an example of successful disease gene identification based on prior knowledge of (a) a related human phenotype and (b) a related mouse phenotype.

Answer

(a) Identification of the gene for congenital contractural arachnodactyly (OMIM 121050). This disorder shows overlapping features with Marfan syndrome (OMIM 154700), a dominantly inherited disorder of fibrous connective tissue. After the *FBN1* fibrillin gene was shown to be a locus for Marfan syndrome, the related *FBN2* fibrillin gene was investigated and shown to be mutated in congenital contractural arachnodactyly.

(b) Identification of the gene for Waardenburg–Hirschsprung disease (Waardenburg syndrome, type 4A; OMIM 277580). The *Dominant megacolon* (*Dom*) mouse is a model of Hirschprung disease in which there is a congenital absence of ganglion cells in regions of the gastrointestinal tract. The mutant mice were observed to have pigmentary abnormalities resembling those in Waardenburg syndrome. After the *Sox10* gene was identified as the disease gene in the *Dom* mouse mutant, the corresponding human gene, *SOX10*, was screened for mutations and shown to be the gene mutated in Waardenburg–Hirschsprung disease.

Question 2

What is a lod score? In standard genomewide linkage analyses the cutoff for statistical significance of linkage is a lod score of +3. How was this limit set?

Answer

A lod score is a statistical estimate of the likelihood that two loci, such as a disease locus and a marker locus, are located near to each other on the same chromosome. It gets its name from logarithm of the odds, because a lod score is expressed as the logarithm to the base 10 of a likelihood ratio: the likelihood of linkage divided by the likelihood of nonlinkage. A lod score of 3 means that the $\log_{10}$(likelihood of linkage/likelihood of nonlinkage) = 3.0. That means that the likelihood of linkage is 1000 times greater than the likelihood of nonlinkage.

Why should such an apparently high burden of proof be needed when for many statistical studies a *P* value of 0.05 or 0.01 suffices? The answer stems from the inherent improbability of linkage at the outset. For any two loci to be linked, they need to be located not just on the same chromosome out of more than 20 chromosomes but also comparatively close together on the same chromosome. A rough calculation might suggest that the prior odds would be of the order of 50:1 against linkage or only 1:50 in favor of linkage. If something has a very low starting probability, the burden of proof needs to be higher. A likelihood ratio of 1000:1 in favor of linkage multiplied by a prior odds of 1:50 in favor of linkage gives a final odds of only 20:1 in favor of linkage. That is, a single lod score of 3 is

not proof of linkage; there is a 1 in 20 chance that the loci are not linked, and in practice genomewide claims for linkage based on a single lod score should be treated as provisional if the lod score is less than 5.

Question 3

The table below shows the percentage phenotype concordance in monozygotic (MZ) and dizygotic (DZ) twins in four hypothetical genetic diseases A to D. Which disease would you estimate to have the highest heritability and which the lowest heritability, and why?

Disease	**Concordance in MZ twins (%)**	**Concordance in DZ twins (%)**
A	37.5	16.2
B	15.2	11.7
C	19.2	7.9
D	17.2	1.6

Answer

Diseases in which there is high heritability show a much higher percentage concordance in MZ twins when compared with that in DZ twins. When a ratio is taken of the two concordances, as shown in the table below, the MZ/DZ ratio is highest in disease D, indicating that it has the highest heritability, and lowest in disease B (lowest heritability).

Disease	**$\frac{\text{Concordance in MZ}}{\text{Concordance in DZ}}$**
A	37.5/16.2 = 2.31
B	15.2/11.7 = 1.30
C	19.2/7.9 = 2.43
D	17.2/1.6 = 10.75

Question 4

The risk of developing a disease is sometimes expressed as a risk ratio, λ. What is meant by this ratio? Disorder A has a λ_S of 600 and disorder B has a λ_S of 6. What types of disease are A and B?

Answer

The risk ratio, λ, is the disease risk for a relative of an affected person divided by the disease risk for an unrelated person. The risk ratio is a measure of the contribution made by genetic factors to the etiology of the disease.

λ_S is the ratio of the disease risk of a sib of an affected person and the disease risk of an unrelated person. A λ_S of 600 means that a brother or sister of an affected person has 600 times the risk of developing that disease than does an unrelated person. Such a very high risk can be explained only if A is a single-gene disorder. But a modest λ_S of 6 means that a person has 6 times the risk of

developing the same disease as his affected brother or sister than does an unrelated person. That is compatible with B being a complex common disease for which siblings may have some genetic susceptibility factors in common.

Question 5

Genomewide linkage studies can often be carried out with just a few hundred DNA markers, but genomewide association studies often use hundreds of thousands of markers. Explain why this difference exists by explaining the very different designs of these two approaches.

Answer

Essentially, association is an effect that can be observed over very short distances only, whereas linkage between two loci can be observed over quite long regions of a chromosome. Using, say, 400 markers on a genomewide linkage analysis means that there is on average 1 marker for each 8 Mb of DNA. That is possible because linkage tracks the co-segregation of alleles in *families* over just a few generations. There will be only a few meioses in which recombination might separate a disease locus from a marker locus.

Association, however, is a statistical property that looks at the co-segregation of alleles at different loci in *populations*. The DNA analyzed comes from individuals who last had a common ancestor many generations ago. There will have been a comparatively large number of meioses in which there will have been an opportunity for recombination to separate a marker locus from a locus where there is a genetic susceptibility factor. To have any chance at all of detecting association, therefore, a marker locus and a disease susceptibility locus have to be very closely linked.

It might help to visualize this by imagining a founder effect where a small European immigrant population with a common disease susceptibility allele arrive in North America at the beginning of the sixteenth century. Imagine that the disease susceptibility locus is very tightly linked to a marker locus where allele 2 had a high frequency in the immigrant population. If we take an average generation time of 25 years, there could be expected to have been 20 generations and very many meioses in which there has been an opportunity for recombination to separate the association of allele 2 at the marker locus with the disease-susceptibility variant. Although allele 2 will be found at the marker locus in many normal people in the population, and other alleles may be found at the marker locus in affected individuals, there may still be a small excess of allele 2 in affected individuals which can be detected statistically. The requirement that the marker locus must be very tightly linked to the disease susceptibility locus means that association is normally detected over kilobases rather than megabases. That is why so many marker loci need to be used in genomewide association analysis.

Question 6

What is meant by the odds ratio in case-control studies? Calculate the odds ratio from the table below.

	Number of cases with disease X	**Number of unaffected controls**
Possessing genetic variant X	850	297
Lacking genetic variant X	150	703

Answer

The odds ratio is the odds of being affected if you have a specific genetic variant divided by the odds of being affected if you lack the genetic variant. For the population group with genetic variant X, the odds of being affected are 850/297 = 2.862. For the group lacking X, the odds of being affected are 150/703 = 0.213. The odds ratio is therefore 2.862/0.213 = 13.44.

Question 7

What is linkage disequilibrium? Which of the following haplotypes shows evidence of linkage disequilibrium, given the individual allele frequencies?

(a) haplotype *A*1-B*3* with a population frequency of 0.101 (frequencies of 0.231 for *A*1* and 0.431 for *B*3*)
(b) haplotype *C*2-D*1* with a population frequency of 0.071 (frequencies of 0.311 for *C*2* and 0.225 for *D*1*)
(c) haplotype *E*1-F*1* with a population frequency of 0.205 (frequencies of 0.236 for *E*1* and 0.289 for *F*1*)
(d) haplotype *X*2-Y*3* with a population frequency of 0.101 (frequencies of 0.532 for *X*2* and 0.434 for *Y*3*).

Answer

Linkage disequilibrium is the nonrandom association of alleles from different loci in a population. In practice, the alleles need to be from very closely linked loci. It can be inferred when the population frequency of a haplotype is unexpectedly significantly different from the product of the individual population frequencies of the alleles.

Haplotypes (c) and (d) both show evidence of linkage disequilibrium.

Haplotype (c) is substantially more frequent than would be expected from the allele frequencies. The combination of alleles might possibly confer some selective advantage to the person that carries it so that they have increased fitness. Alternatively, the linked loci might by chance be in a region where recombination is suppressed compared with the average.

Haplotype (d) is substantially less frequent than would be expected from the allele frequencies. The combination of alleles might conceivably be selectively disadvantageous. Alternatively, the linked loci might be in a region with greatly enhanced recombination.

Question 8

In genomewide association studies, the threshold for statistical significance, *P*, is often set at a very high value, often about 5×10^{-8}. Why so high?

Answer

Genomewide association studies use SNP microarray hybridization, which typically involves many hundreds of thousands of parallel DNA hybridizations, one for each of the fixed oligonucleotides. Because such huge numbers of hybridization tests are being carried out, there is a greatly increased probability of obtaining a positive result for individual DNA hybridization assays simply by chance. Stringent statistical significance thresholds are therefore required to assess the significance of individual hybridization results. For example, a more stringent genomewide significance threshold is to divide the standard *P* value of 0.05 by the number of tests. If the SNP hybridization involved 1,000,000 hybridization assays, a stringent *P* value would then be $0.05/1{,}000{,}000 = 5 \times 10^{-8}$.