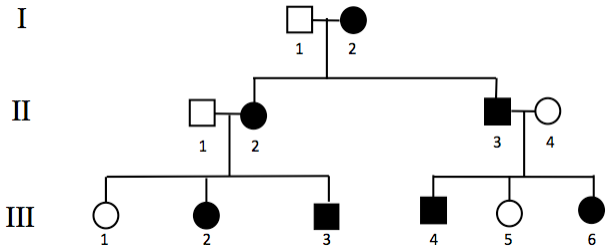
Genetics and Genomics in Medicine Chapter 8

Questions & Answers

Linkage Analysis Question

Question 8.1



Affected members of the pedigree above have an autosomal dominant disorder, and cytogenetic analyses using conventional chromosome banding did not identify a disease-associated chromosomal abnormality. However, recent previous studies in other families mapped the disease locus to a region on the short arm of chromosome 17 and showed that four polymorphic microsatellite markers, *P, Q, R* and *S* are closely linked to the unidentified disease locus. (The order of the markers from most distal to most proximal is given by the following sequence: *P-Q-R-S*). When the same four markers were used to genotype each member of the family above, the alleles obtained were as listed below.

	I-1	I-2	II-1	II-2	II-3	II-4	III-1	III-2	III-3	III-4	III-5	III-6
P	1,4	1,2	1,4	1,2	2,4	1,4	1,4	1,2	2,4	2,4	4	1,2
Q	2,3	1,4	2,4	1,3	1,2	3	2,3	1,4	1,2	1,3	2,3	1,3
R	1	1,3	2	1,3	1,3	2,4	1,2	2,3	1,2	3,4	1,4	2,3
S	1,3	2,4	1,2	3,4	1,4	2	1,3	2,4	1,3	2,4	1,2	2,4

a) If we were to assume that the disease locus is also at the same 17p region in this family, what is the disease haplotype originating from I-2 (that is, the sequence of alleles at the four marker loci on the chromosome that has the disease allele)?
b) Has the disease haplotype been conserved in all family members?
c) Is it likely that the disease in this family is linked to 17p?
d) If the disease locus were to be at 17 p in this family, what do you deduce about the position of the disease locus with respect to the marker loci?

Answer 8.1

a) 2 (*P*) – 1 (*Q*) – 3 (*R*) – 4 (*S*).

b) No. III-3 has inherited a maternal recombinant haplotype: 2 (*P*) – 1 (*Q*) – 1 (*R*) – 3 (*S*).
c) The disease locus does seem to segregate with 17p markers through most meioses – there is the exception of the maternal chromosome inherited by III-3 where alleles for markers *R* and *S* are not the expected ones.
d) Because of the presumed recombination in the maternal chromosome 17 inherited by III-3, the disease locus can be inferred to be distal to marker loci *R* and *S*.

Explanation 8.1
The deduced haplotypes for all family members are as follows (using the order *P–Q–R–S*):
I-1: 4-2-1-1 and 1-3-1-3
I-2: 2-1-3-4 and 1-4-1-2
II-1: 4-2-2-1 and 1-4-2-2
II-2: 1-3-1-3 (paternal) and 2-1-3-4 (maternal)
II-3: 4-2-1-1 (paternal) and 2-1-3-4 (maternal)
II-4: 1-3-2-2 and 4-3-4-2
III-1: 4-2-2-1 (paternal) and 1-3-1-3 (maternal)
III-2: 1-4-2-2 (paternal) and 2-1-3-4 (maternal)
III-3: 4-2-2-1 (paternal) and 2-1-1-3 (maternal)
III-4: 2-1-3-4 (paternal) and 4-3-4-2 (maternal)
III-5: 4-2-1-1 (paternal) and 4-3-4-2 (maternal)
III-6: 2-1-3-4 (paternal) and 1-3-2-2 (maternal)

Multiple Choice Questions

Question 8.2

The same genomic DNA samples recovered from family members in Question 8.1 have been genotyped with an additional polymorphic marker T that has been identified in the same region of 17q. The alleles recorded were as follows:

	I-1	I-2	II-1	II-2	II-3	II-4	III-1	III-2	III-3	III-4	III-5	III-6
T	2,3	2	2	3	2	3	2,3	2	2	3	2,3	3

Which of the following statements is most likely to apply?
a) The genotypes have been recorded wrongly, possibly because samples were switched, and there is a need to repeat the genotyping.
b) There is something inherently unstable about marker *T*; there is no point in repeating the genotyping, and a different marker on 17q should be used instead.
c) The assumed biological relationships shown in the pedigree are wrong.
d) The disease-causing mutation has affected the marker locus.

Answer 8.2

d) The disease-causing mutation has affected the marker locus.

Explanation 8.2

Most likely, the disease in this family is caused by a deletion that is large at the DNA level (but too small to be detected by chromosome banding), and the deletion spans the *T* marker locus and the gene that causes the disease. All affected individuals will therefore be hemizygous for marker locus *T*, but II-1 and II-4 would be expected to be homozygous at the *T* locus, with genotypes 2-2 and 3-3 respectively.

Question 8.3

Which, if any, of the following statements is true?

a) The heritability of a disease can vary from one human population to another, but remains fairly constant within any individual population.
b) The risk ratio in complex diseases is high compared to in monogenic disorders.
c) In complex diseases the concordance in phenotype between monozygotic twins is higher than between dizygotic twins
d) Disorders in which genetic factors have a large role always have a significantly higher concordance in phenotype between monozygotic twins and between dizygotic twins than do disorders where genetic factors play a less significant role.

Answer 8.3

c) In complex diseases the concordance in phenotype between monozygotic twins is higher than between dizygotic twins

Question 8.4

With regard to affected sib pair (ASP) analysis which, if any, of the following statements is true.

a) ASP analysis is a type of parametric linkage analysis that is popularly used to study complex diseases.
b) It is highly convenient because samples are required from the affected sibs only and samples from just 100 affected sib pairs are usually enough to obtain decent results.
c) It is less suited to studying complex disease where the relative risk of disease is low.
d) A lod score of 3 is highly significant evidence for linkage in ASP analysis.

Answer 8.4

c) It is less suited to studying complex disease where the relative risk of disease is low.

Explanation 8.4
It is a nonparametric linkage analysis that often requires many hundreds of sib pairs. A lod score of 3 would be at most suggestive evidence of linkage; for highly significant evidence of linkage a lod score above 5.4 would be required.

Question 8.5

Regarding two-point linkage analysis, which, if any, of the following statements is true?

a) A lod score of 3 means that the likelihood of the data, given that the two loci are linked, is 1000 times greater than is the likelihood of the data, if the two loci are unlinked.
b) A lod score of 3 is highly significant evidence for linkage.
c) A lod score of -2 is highly significant evidence against linkage.
d) A lod score of 3 is 100,000 times more convincing evidence of linkage than a lod score of -2.

Answer 8.5

a) A lod score of 3 means that the likelihood of the data, given that the two loci are linked, is 1000 times greater than is the likelihood of the data, if the two loci are unlinked.
d) A lod score of 3 is 100,000 times more convincing evidence of linkage than a lod score of -2.

Explanation 8.5
A lod score of +3 and -2 are at the very threshold of statistical significance, respectively in favor, or against linkage.

Question 8.6

With respect to genomewide association (GWA) studies, which, if any, of the following statements, is false?

a) The vast majority of genetic variants identified in GWA studies of complex association studies have been of weak effect, with odds ratios of 1.2 or less.
b) Even where GWA studies have identified disease susceptibility factors with odds ratios greater than 1.2, getting from disease variant to identifying a disease susceptibility locus is essentially impossible.
c) GWA studies have generally been of very limited use in predicting disease risk.
d) GWA studies have been very valuable.

Answer 8.6

b) Even where GWA studies have identified disease susceptibility factors with odds ratios greater than 1.2, getting from disease variant to identifying a disease susceptibility locus is essentially impossible.

Explanation 8.6
GWA studies have been very important in providing new disease biomarkers and new insights into the pathogenesis of complex diseases.

Question 8.7

With respect to genomewide association (GWA) studies, which, if any, of the following statements, is true?

a) GWA studies to determine genetic susceptibility to human complex disease usually use 1000 polymorphic SNPs markers that are distributed across the genome
b) Unlike linkage analyses, GWA studies are ideally suited to identifying genetic susceptibility in genetically heterogeneous diseases.
c) GWA studies have been very important in elucidating the pathogenesis of complex diseases and in providing new disease biomarkers.
d) GWA studies have transformed our ability to predict disease risk.

Answer 8.7

c) GWA studies have been very important in elucidating the pathogenesis of complex diseases and in providing new disease biomarkers.

Question 8.8

Genomewide association (GWA) studies have identified common single nucleotide DNA variants that are very rarely of even moderate effect and that are collectively insufficient to explain the genetic variance of complex diseases. Which of the following can at least partly explain the genetic variance?

a) Rare single nucleotide variants.
b) *de novo* variants.
c) Copy number variants.
d) Gene-gene interactions

Answer 8.8

a) Rare single nucleotide variants.
c) Copy number variants.
d) Gene-gene interactions

Question 8.9

The human *APOE* gene has three common alleles, *APOE*e2, APOE*e3*, and *APOE*e4*, that gives rise, respectively, to the common alleles apoE2, apoE3 and apoE4 at the protein level. Which, if any, of the following statements is true?

a) The **e4* allele confers a high risk of Alzheimer disease and people with two **e4* alleles have twice the disease risk of people with one **e4* allele.
b) The **e3* allele is a protective factor, conferring reduced risk of Alzheimer disease while the **e2* allele has a disease risk that is intermediate between those of **e3* and **e4*.
c) The **e4* allele is considered the ancestral *APOE* allele because although the chimp and gorilla also have three apoE proteins the apoE4 protein is the most frequent allele.
d) The **e4* allele has reached a high frequency because Alzheimer disease has such a late age at onset that reproductive success rates are not diminished.

Answer 8.9
None.

Explanation 8.9
All are false.
a) People with two **e4* alleles have a roughly five times risk of Alzheimer disease than people with one **e4* allele.
b) **e2* is a protective factor; **e3* has intermediate risk.
c) Chimp and gorilla have a non-polymorphic apoE protein that is more closely related to human apoE4.
d) The **e4* allele also predisposes to cardiovascular disease: People who carry at least one copy of the *APOE* e4 allele have an increased chance of developing atherosclerosis (accumulation of fatty deposits and scar-like tissue in the lining of the arteries; the progressive narrowing of the arteries increases the risk of heart attack and stroke, and so biological fitness is reduced.

Question 8.10
Which, if any of the descriptions below is false. Our personal microbiomes
a) are microorganisms that are mostly located within our guts.
b) contain 10 times more cells than we have.
c) are tolerated, despite being foreign microorganisms, by down-regulating innate immune responses, mostly those that require Toll-like receptors.
d) include microbes that are highly beneficial to us.

Answer 8.10
None.

Explanation 8.10
All are true. In the case of (d), many of the microbes are of active benefit to us in different ways, helping us to derive additional energy through the fermentation of undigested carbohydrates, break down xenobiotics, and synthesize vitamins B and K.

Fill in the Blanks Questions

Question 8.11

Fill in the blanks

Before we had a human genetic map, disease gene identification was difficult. Sometimes, however, knowledge of the gene product provided a path to a gene underlying a Mendelian disorder. For example, haemophilia A was long known to be a deficiency of a specific blood clotting protein, ___1____ ___2___, and using that information it was possible to purify large amounts of ___1____ ___2___ from pig blood and then to design ____3_____ oligonucleotides that corresponded to all possible codon interpretations of an optimal part of the amino acid sequence of the pig ___1____ ___2___ protein. The resulting oligonucleotides were then used as ____4_____ probes to screen first a human ____5____ library and then a ____6____ ___ 7____ library to identify the underlying human gene.

Answer 8.11

1. factor. 2. VIII. 3. degenerate. 4. hybridization. 5. cDNA. 6. genomic. 7. DNA

Question 8.12

Fill in the blanks

Once we had a human genetic map, a general method, ____1____ ____2____, could be applied to identifying genes underling single gene disorders. In order to carry this through, there was first the need to identify families with multiple affected individuals and to obtain samples of ____3____ ____4____ from both affected and unaffected family members. The individual samples could then be tested by assaying each of several hundred DNA ____5_____ . The DNA ____5____ were selected because they were known to be ____6____ and because they were each known to map to a ____7____ subchromosomal location. The object was to identify ____5____ that must map close to the disease gene because alleles from these markers tended to ____8____ with disease through generations in families.

Answer 8.12

1. linkage. 2. analysis. 3. genomic. 4. DNA. 5. markers. 6. polymorphic. 7. unique (or specific). 8. segregate.

Question 8.13

Fill in the blanks

Poor family structure means that it may not be easy, or even feasible, to use ____1____ ___2____ in order to map the underlying gene for certain types of single gene disorders. The disorder may be extremely rare so that insufficient samples can be obtained, or the vast majority of affected individuals are sporadic cases that arise through a ____3____ ___4____ mutation, such as in severe congenital ____5____ disorders. In the past, gene identification in these cases was sometimes possible using cytogenetic analyses to identify a disease-associated chromosome abnormality. Two or more affected individuals might be identified to have a translocation or ____6____ with a common ____7_____ or a recurring interstitial or terminal ____8_____ might be identified. Once the subchromosomal location of the gene had been identified it was possible to use ____9_____ ____10_____ or ____9_____ ___11____ gene approaches to identify the underlying gene. More recently, a more direct approach, ___12____ ____13____ makes it possible to quickly get to the underlying gene without knowing its chromosomal location.

Answer 8.13
1. linkage. 2. analysis. 3. de. 4. novo. 5. dominant. 6. inversion. 7. breakpoint. 8. deletion. 9. positional. 10 cloning. 11. candidate. 12. exome. 13. sequencing.

Question 8.14
Fill in the blanks

In a _____1____-control study affected individuals (___1___) are compared with control subjects that are suitably ___2___ to the ____1___ with regard to age, sex and so on. In genetic studies of complex diseases the ___1___ and controls are genotyped for genetic ___3____ in the hope of identifying ___3____ that may confer significantly increased disease risk (____4____ ___5____) or significantly reduced disease risk (____6____ ____5____). A popular way of assessing the data is to calculate an ___7___ ___8_____, that is the ___7____ of being affected if the genetic ____3____ is present divided by the odds of being affected if the genetic ____3____ is absent.

Answer 8.14
1. case(s). 2. matched. 3. variant(s). 4. susceptibility 5. factors. 6. protective . 7. odds. 8. ratio.

Question 8.15
Fill in the blanks

While linkage is a ____1_____ phenomenon, association is a _____2_____ property. Linkage is concerned with seeking to identify the map relationship between two or more ___3___, such as disease and marker ___3____, by analysing samples from individuals within ____4____.

Association, on the other hand, is a relationship between ____5____and is studied by analysing samples from individuals within ____6____. Linkage works over ____7____ distances whereas, in practice, association works over very ____8_____ distances. If allele *A*1* is shown to be positively associated with a complex disease it would be described as a disease ____9____ factor, and if allele *A*2* is negatively associated with the disease it would be described as a ____10____ factor

Answer 8.15
1. genetic. 2. statistical. 3. loci. 4. families. 5. alleles. 6. populations. 7. long. 8. short. 9. susceptibility. 10. protective.

Question 8.16
Fill in the blanks

In an autoimmune disorder, cells in the body come under attack from certain antibodies known as ____1_____ , and also from certain types of ____2_____ T cell that may attack specific host cells (such as insulin-producing ____3_____ ____4_____ cells in the case of type I diabetes). In ____2_____ T cell responses the host cell peptides serve as ____5_____ and they are presented to T cells after they have been bound by ____6_____ proteins. ____6_____ proteins differ in their ability to bind individual peptide ____5____ and that is the primary basis for ____6____-disease associations.

Answer 8.16
1. autoantibodies. 2. autoreactive. 3. pancreatic. 4. beta. 5. autoantigens. 6. HLA.

Question 8.17
Fill in the blanks

Imagine a genetic variant that is tightly linked to a disease susceptibility allele. On a single chromosome, the ____1____ containing the genetic variant and the linked disease susceptibility allele will have a higher ____2____ than would be expected (that is, it would be higher than the ____2____ of the genetic variant multiplied by the ____2_____ of the disease susceptibility allele. This is an example of _____3_____ ____4_____ , the non-random _____5_____ of alleles at two or more loci. Although _____3_____ ____4_____ describes *any* non-random _____5_____ of alleles at different loci, in practice, the alleles are at very closely _____6_____ loci. Although _____3_____ _____4_____ can occur if a particular combination of alleles offer some advantage and is _____7_____ selected, it may often simply reflect reduced _____8_____ between loci (certain regions of the genome, such at the HLA complex, show significantly reduced _____8_____). When a new DNA variant is created by mutation it will

show very tight ____3_____ with alleles at neighboring loci. However, the _____3_____ ____4_____ will gradually be eroded by _____8_____ but that will take a very _____9_____ time for any locus that is physically very close to the locus with the new mutation.

Answer 8.17
1. haplotype. 2. frequency. 3. linkage. 4. disequilibrium. 5. association. 6. linked. 7. positively. 8. recombination. 9. long.

Question 8.18

Fill in the blanks with single words or single letters.

Each of us carries our personal ____1____ that shares our body space (but is principally distributed within our ____2____) and that contains ____3____ times more cells than our body. These cells are foreign ____4_____ that are nevertheless tolerated by the body, largely by suppressing _____5_____ ____6_____ responses, notably those that depend on _____7____ receptors. Our personal _____1_____ is normally beneficial to us because some of the _____4_____ are beneficial to us in different ways. They can help us derived additional energy through the fermentation of undigested _____8_____, help us break down _____9_____ , and they synthesize vitamins _____10_____ and _____11_____ for us. In ____12____ ____13____ diseases, however, an abnormal ____6____ response is directed against antigens carried by foreign ____4_____ within our ___2____ and that leads to accumulation of ____14____ blood cells within the linings of the _____15_____ , producing chronic _____16_____.

Answer 8.18
1. microbiome. 2. gut. 3. ten. 4. microorganisms. 5. innate. 6. immune. 7. Toll-like. 8. carbohydrates. 9. xenobiotics. 10. B. 11. K. 12. inflammatory. 13. bowel. 14. white. 15. intestines. 16. inflammation.

Essay and List Questions

Question 8.19

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 one has the lowest heritability, and why?

Disease	**% Concordance in MZ twins**	**% Concordance in DZ twins**
A	25.5	5.6
B	15.2	11.3
C	7.0	3.0
D	41.0	5.2

Answer 8.19

The ratio of the values for the percentage concordances in MZ twins and DZ twins is a key indicator of heritability. Disease D has the highest ratio (41.0/5.2 = 7.88) and the highest heritability. Disease B has the lowest ration (15.2/11.3 = 1.35) and the lowest heritability.

Question 8.20

Strategies to identify the genes that underlie single gene disorders have often relied on first obtaining a subchromosomal location for the disease gene. List two approaches that have been taken to identify subchromosomal locations for these disorders.

Answer 8.20

1. Linkage analyses. Testing markers from across all the chromosomes to see if alleles at any of the marker loci show a tendency to co-segregate with the disease in families.
2. Cytogenetic analysis to look for disease-associated chromosomal rearrangements. The most profitable have been translocations and inversions (the hope is that one of the breakpoints has resulted in a change of expression by cleaving within or close to the disease gene), and interstitial or terminal deletions.

Question 8.21

The first genome-wide human genetic map was created by taking a completely different approach to the approaches used to create genetic maps in model organisms. What was the essential difference?

Answer 8.21

In model organisms, genetic mapping relied on *phenotypes.* That is the genetic maps were obtained by mapping gene mutations that cause readily identifiable phenotypes. In *Drosophila*, for example, crosses can easily be set up to breed mutant white-eyed flies with flies that have abnormal curly wings; the progeny are then examined to see if the two mutant phenotypes segregate together or not. In humans, however, that kind of approach could never be applied—a different strategy was needed.

Instead of having a genetic map based on gene mutations, the solution to making a human genetic map was to identify *general DNA variants* (that mostly do not map within coding sequences, and usually have no known effects on the phenotype). Different types of variants were identified and mapped to specific locations in our genome, beginning with restriction fragment length polymorphisms (RFLPs) that created or destroyed a restriction site followed by microsatellite polymorphisms that varied in the copy number of short tandem repeats. These provided anonymous genetic *markers* for each chromosome and the segregation of alleles at these marker loci were then followed through multigeneration families to identify neighboring markers (because their alleles would tend to segregate together).

Question 8.22

To carry out exome sequencing the desired exome is first captured from a sample of genomic DNA. How is that achieved?

Answer 8.22

Exome capture is achieved by a DNA hybridization procedure in which probes that have been designed to represent a reference human exome are used as bait to capture complementary sequences from the genomic DNA sample. The probes typically have a biotin group covalently bonded at one end. The biotinylated probes are mixed with a test sample of genomic DNA that has been randomly sheared to produce small DNA fragments, and the DNA mixture is heated to produce single-stranded fragments.

After cooling, single-stranded biotinylated probes from the reference exome are able to hybridize to complementary exon sequences from the test sample. The mix is then exposed to magnetic beads that have been coated with streptavidin, a protein that has an extraordinarily high affinity for biotin. Biotinylated strands with complementary exon sequences from the test sample will bind to the streptavidin groups on the magnetic beads and they can be purified by using a magnet to remove the beads that are then washed to remove any non-specifically bound sequences. Subsequently, the captured exon sequences from the test sample can be eluted after raising the

temperature to break the hydrogen bonds holding them to the complementary reference exome sequences.

Explanation 8.22
Figure 8.7 gives an illustration that shows oligonucleotide probes, but related methods use RNA or DNA probes.

Question 8.23

The risk ratio, λ, is commonly used in complex disease. What does it mean and how do risk ratios compare in complex disease and monogenic disorders?

Answer 8.23
It means the ratio of the disease risk for the relative of an affected person divided by the disease risk for an unrelated person. For example, λ_s means the disease risk for a sib (brother or sister) of an affected person divided by the disease risk for an unrelated person. The risk ratio is a measure of how important genetic factors are in the etiology of a disease. Inevitably, the highest λ values are in monogenic disorders – for example, if the general lifetime risk of cystic fibrosis in a population of European origin were 1 in 2000 and the risk to a fetus is 1 in 4 where the prospective parents have had a previously affected child then λ_s = ¼ divided by 1/2000 = 500. For most complex diseases, however, λ_s is below 30.

Question 8.24

Illustrate how the heritability of a disease can change over time using an example of a) a monogenic disorder and b) a complex disease.

Answer 8.24

a) *Phenylketonuria*. The deficiency of phenylalanine hydroxylase in phenylketonuria produces elevated phenylalanine and toxic by-products that can result in cognitive disability. In the recent past the disease was almost wholly due to genetic factors, and so the heritability was extremely high. In modern times, neonatal screening programs in many countries allow early detection and treatment using low-phenylalanine diets. Now, in societies with advanced health care, phenylketonuria results mostly from environmental factors that lead to failure to deliver the treatment (inefficiency in health care systems, reluctance of families to seek out treatment, non-compliance with the diet, and so on).
b) *Type 2 diabetes*. There has been a huge recent increase in type 2 diabetes in many populations, mostly as a result of increasingly unhealthy diets and lack of exercise. As a result of the major changes in environmental factors (diet, exercise) the heritability of this

disorder in many populations is now much reduced when compared with just a few decades ago.

Question 8.25

What is the difference between parametric and non-parametric linkage analyses and under what circumstances are they applied to studying human genetic disease.

Answer 8.25

Parametric linkage analyses require a specific genetic model that gives details of certain key parameters: the mode of inheritance, disease gene frequency, and penetrance of disease genotypes. For Mendelian disorders, the parameters can generally be provided without great difficulty (although penetrance can sometimes be difficult to gauge). Parametric linkage analyses have been very successful in mapping genes for Mendelian disorders, and Mendelian subsets of a complex disease, such as in large autosomal dominant pedigrees with early-onset Alzheimer disease. But they have more limited applicability in complex disease (because of the difficulty in providing the required parameters).

Non-parametric linkage analyses do not require any genetic model to be stipulated and have been deployed to analyze segregation patterns in complex disease. Affected sib-pair analysis, for example, relies simply on analyzing affected sibs only in multiple families to see if there are regions of the genome where there is greater haplotype sharing in affected sib pairs across multiple different families than would be expected by chance.

Question 8.26

What is a haplotype block, and how are they organized in the human genome?

Answer 8.26

A haplotype is a combination of alleles at linked loci on a single chromosome, and a haplotype block is a segment of DNA that shows limited haplotype diversity. About 85% of our genomic DNA is a mosaic structure composed of haplotype blocks. As an illustration of the limited degree of haplotype diversity, in some cases haplotype blocks that encompass as many as 8 SNP loci (each with two alleles) show only two haplotypes out of the 256 (2^8) possible haplotypes.

At some time in our recent history, the human population was reduced to a very small number – perhaps just 10,000 or so individuals – and only a proportion of these individuals were able to transmit their genomic DNA through many generations to people living today. Because the DNA in the genomes of current humans originated from a limited number of ancestors, and because of recombination over the generations, very small segments of our genomic DNA were contributed by a tiny number of our ancestors. As a result, there is limited haplotype diversity across most of the genome.

The average size and diversity of human haplotype blocks does show some variation between different populations. In populations of European and Asiatic ancestry, the average haplotype block is 5.9 kb long with an average of about 3.6 different haplotypes per block, but in the Yoruban population of Nigeria the blocks average 4.8 kb in size and there are an average of 5.1 different haplotypes per haplotype block. (All human populations ultimately originated in Africa, and current African populations have greater genetic diversity than other human populations).

Question 8.27

What is the basis of HLA associations with autoimmune disorders?

Answer 8.27

In autoimmune disorders there is a breakdown in the ability to distinguish self from nonself. As a result, cells in the body can be attacked by certain antibodies (autoantibodies), and by autoreactive T cells that inappropriately recognize certain host antigens (autoantigens); the activated T cells kill certain populations of host cells, such as insulin-producing pancreatic beta cells in type 1 diabetes. In auto-reactive T-cell responses, host peptides serve as autoantigens. HLA proteins can present peptide antigens to T-cells (to do that an HLA protein binds a peptide autoantigen and then transports it to the cell surface where the HLA-peptide combination can be specifically recognized by T cells carrying a particular T cell receptor). HLA proteins differ in their ability to bind the peptide autoantigen. As a result, specific HLA proteins are associated with autoimmune disease.

Question 8.28

Before genomewide association (GWA) studies became successful, association studies used to rely on candidate gene approaches. How successful were the candidate gene approaches?

Answer 8.28

The success of candidate gene association studies was limited in the sense that only a few types of variant were identified to be disease susceptibility factors or protective factors, whereas very large numbers of disease susceptibility and protective factors were subsequently identified by GWA scans. On the other hand, however, candidate gene association studies have identified extremely important disease-associated variants, notably various HLA proteins that make the largest genetic contribution to diverse autoimmune diseases, the *APOE*ε4* allele that makes the largest known genetic contribution to Alzheimer disease, and *CTLA4* variants that are important in Graves diseases and type I diabetes mellitus.

Question 8.29

In the years when association studies were limited to candidate gene approaches what were the technological drawbacks that prevented genomewide association (GWA) studies and what developments made GWA studies possible?

Answer 8.29

Unlike linkage analysis, which can work over quite long distances across chromosomes, association analysis works only over very short distances. Thus, whereas human genomewide linkage analyses can easily be conducted with just a few hundred polymorphic DNA markers, carrying out GWA studies requires a very high density of markers – hundreds of thousands of polymorphic DNA markers are required so that the spacing of neighboring markers across the genome gets down to kilobases (instead of the several megabases of DNA that separates neighboring markers in genomewide linkage analyses). And even if hundreds of thousands of markers could be available how could they be assayed quickly?

For a long time, we simply didn't have enough polymorphic DNA markers, but that changed after strenuous attempts were made to identify human single nucleotide polymorphisms (SNPs). SNPs have the drawback that they are not very polymorphic, compared to microsatellites, but they have the big advantage that they occur extremely frequent in the human genome, and millions of human SNP markers have been developed. The second technological development was a system of assaying hundreds of thousands of SNPs. That became possible when microarray technology was developed to allow many hundreds of thousands of different SNPs to be typed in parallel in a single experiment.

Question 8.30

How successful have genomewide association studies been in identifying genetic susceptibility to complex disease? What has been the main value of these studies?

Answer 8.30

In one sense genomewide association studies have not been as successful as hoped. For many diseases, notably the inflammatory bowel diseases, GWA studies have implicated very many novel loci as disease susceptibility but for others, including highly heterogeneous neuropsychiatric disorders, such as autism spectrum disorder, GWA studies have had limited success. Even where GWA studies have been very successful in identifying novel disease-susceptibility variants, however, the variants have almost always been of very weak effect, often with an odds ratio of 1.2 or less. Exceptions include some novel factors that strongly predispose to age-related macular degeneration (a leading cause of vision loss in older adults that results from a progressive deterioration of the macula, a central region of the retina). But many of the variants with high odds ratios were identified in the pre-GWA era (such as apoE4 in Alzheimer disease, the common *NOD2* alleles in Crohn's disease, and especially HLA alleles that remain,

by some distance, the strongest known genetic variants in autoimmune disorders). That means GWA studies are often not going to provide that much help in predicting disease risk.
In another sense, however, GWA studies have been outstandingly successful. They have provided large numbers of disease susceptibility loci that allow remarkable new insights into the molecular pathogenesis. Sometimes, whole new pathways of disease have been suggested. In acute macular degeneration, GWA studies have demonstrated the very important contribution made by variants of several different complement factor genes, stressing the vital role of innate immune system response in the pathogenesis. These findings will provide new biomarkers of disease and the hope is that greater insights into the pathogenesis of complex disease will also lead to new and important therapeutic strategies. In addition to disease susceptibility factors, GWA studies have had a measure of success in identifying new protective factors that might be valuable in future approaches to provide resistance to infectious diseases.

Question 8.31

List three possible explanations for the general failure of GWA studies to identify genetic factors that collectively might explain the heritability of complex diseases.

Answer 8.31

1. *Large numbers of common variants with very weak effect*. GWA studies with a few thousand cases and controls are well suited to detecting susceptibility factors with odds ratios of 1.5 or more, but many genuine susceptibility factors might be missed if they have weaker effects (odds ratios of less than 1.2). To have a high chance of detecting these variants, GWA studies need to use much larger numbers of cases and controls, either directly or in meta-analyses using aggregate data from multiple individual studies.
2. *Rare variants of large effect*. A major limitation of GWA studies is that they are restricted to identifying associations with common (frequent) variants. Much of the disease susceptibility might conceivably be due to a heterogeneous set of rare variants with individually strong effects (high odds ratios).
3. *Gene–gene and gene–environment interactions*. The concept of heritability is flawed. It assumes additive effects by different loci, and the proportion of heritability explained by known GWA variants does not take into account genetic interactions between loci. Heritability is also traditionally separated into genetic and environmental components, but this is simplistic: genes interact with the environment.

Question 8.32

Certain common alleles are known to be associated with specific complex diseases. Why has purifying selection not led to these alleles being eliminated from the population?

Answer 8.32

Many complex diseases are of late onset, and because short human lifespans used to be common until quite recently, susceptibility alleles for aging-related disorders might have had very little effect on reproductive rates over large numbers of generations. Alleles causing diseases that manifest only later in life might therefore have been protected to a very considerable degree from natural selection.

Many common disease alleles also seem to have some advantages. Certain HLA alleles are susceptibility factors for specific autoimmune disorders but are also very important in allowing immune responses against certain viruses and other intracellular pathogens. The common *NOD2* alleles that confer susceptibility to Crohn's disease in many populations of European origin also seem to have been maintained by natural selection. Balancing selection is most probably involved—the deleterious haplotypes might confer some *heterozygote advantage*. Several different susceptibility factors conferring increased risk for type 1 diabetes are also known to be simultaneously protective factors for Crohn's disease.

Common disease variants may also have conferred some advantages in the recent past. The '*thrifty gene' hypothesis* proposes that certain genetic variants confer a selective advantage in populations exposed to famine, and that they were advantageous in the past when food supplies were limited. But in modern societies in which food is plentiful, the same variants have become susceptibility factors for type 2 diabetes. The *APOE*e4* allele, which predisposes to cardiovascular disease as well as to Alzheimer disease (and so will have an effect on reproductive rates), seems to have been the ancestral *APOE* allele. It has been imagined to have been selectively advantageous to early humans who had a low-calorie, low-fat diet. Over time, however, it has increasingly been replaced by the *APOE*e3* allele, which offers the advantage of decreased cholesterol metabolism (reducing the risk of cardiovascular disease).

Question 8.33

Balancing selection might explain why certain common alleles known to be associated with specific complex diseases have not been eliminated by natural selection. List some examples that might suggest this to be the case.

Answer 8.33

- Certain HLA alleles are susceptibility factors for specific autoimmune disorders but are also very important in allowing immune responses against certain viruses and other intracellular pathogens.
- Several different susceptibility factors conferring increased risk for type 1 diabetes are also known to be simultaneously protective factors for Crohn's disease.

Question 8.34

Certain common alleles known to be associated with specific complex diseases may have been maintained in the human population because of beneficial advantage in the past. List some examples that might suggest this to be the case.

Answer 8.34

- The '*thrifty gene' hypothesis* proposes that certain genetic variants confer a selective advantage in populations exposed to famine, and that they were advantageous in the past when food supplies were limited. But in modern societies in which food is plentiful, the same variants have become susceptibility factors for type 2 diabetes.
- The *APOE*e4* allele, which predisposes to cardiovascular disease as well as to Alzheimer disease seems to have been the ancestral *APOE* allele. It has been imagined to have been selectively advantageous to early humans who had a low-calorie, low-fat diet. Over time, however, it has increasingly been replaced by the *APOE*e3* allele, which offers the advantage of decreased cholesterol metabolism (reducing the risk of cardiovascular disease).

Question 8.35

List three observations that support the common disease-common variant hypothesis.

Answer 8.35

1. In this hypothesis, different combinations of common variants at multiple loci are believed to aggregate in specific individuals to increase disease risk and so it offers an explanation for the observation that there is such a steep falling away of disease risk in relatives of probands with a common disease (common diseases often appear as sporadic cases).
2. Long-known common variants have been implicated in specific complex diseases, including many individual HLA alleles that are associated with specific autoimmune disorders, *APOE*e4* and Alzheimer disease, various *NOD2* alleles and Crohn's disease.
3. Some individual common disease alleles are known to be subject to balancing selection (being simultaneously a susceptibility factor for one disease and a protective factor for another disease, as in the case of various individual variants that confer increased risk of type I diabetes also confer reduced risk of Crohn's disease). If balancing selection were a frequent occurrence common variants would not be eliminated by purifying selection and might be expected to make a large contribution in complex disease.

Question 8.36

List three justifications for the common disease-rare variant hypothesis

Answer 8.36

1. The design of GWA studies means that the variants that can be identified as disease risk factors are all common variants. However, GWA studies have found only a very few variants that have even moderately strong effect as a disease susceptibility factor: the great majority have odds ratios of less than just 1.2. Even when summed together the variants generally account for a small proportion of the genetic variance that one might expect, and so there is a gap to fill that might be filled, at least in part, by rare variants.
2. One rationale for the common disease–rare variant hypothesis is that, given the great mutational heterogeneity of individual single-gene disorders, it is not clear why complex diseases should be any different. At one extreme, many complex diseases are known to have Mendelian subsets in which the pathogenesis is due to very rare mutations of extremely strong effect (so that the phenotype is highly penetrant).
3. Recent explosive growth in human populations means that most coding sequence variants are rare variants. (In a recent study, 73% of all protein-coding single nucleotide variants (SNVs) and 83% of the deleterious SNVs, were estimated to have originated in the past 5000 to 10,000 years and so are rare variants).

Question 8.37

To what extent do late-onset common Alzheimer disease and rare autosomal dominant Alzheimer disease show common pathologies at the physiological level and what is the evidence for common biological pathways in the Mendelian and complex disease versions of this disease?

Answer 8.37

In rare dominantly inherited early-onset Alzheimer disease, the disease usually presents between ages 30 and 60 years; disease onset in the common non- Mendelian form normally occurs after age 65 years. The early-onset and late-onset forms have the same post-mortem brain pathology—abundant extracellular plaques, largely composed of amyloid-β peptides of slightly different sizes, and intracellular neurofibrillary tangles mostly made of tau protein.

The evidence for common biological pathways in these two disease subsets revolves around amyloid-β (Aβ) pathways. The first gene to be implicated in complex late-onset Alzheimer disease was the *APOE* gene which works in various pathways involving Aβ. They include: clearance (removing Aβ from the brain by receptor-mediated transport), the pathway that converts Aβ monomers to Aβ oligomers. and a pathway where cerebrovascular events lead to the production of Aβ.

Genomewide linkage studies of autosomal dominant early-onset Alzheimer disease have identified three causative genes: the *APP* gene (which produces amyloid precursor protein), and *PSEN1* and *PSEN2*, which are both involved in processing APP to make Aβ. *APOE* has also been shown to be a modifier gene that affects the phenotype in autosomal dominant early-onset forms, for example in families with presenilin mutations.

Question 8.38

HLA associations with autoimmune disorders are among the very strongest of associations between genetic variants and disease, but have rarely been mapped to the amino acid level. An exception is the HLA association with rheumatoid arthritis. What is known about this association, and how did detailed knowledge of how HLA proteins function help to identify the amino acids implicated in the association?

Answer 8.38

Detailed fine-scale association mapping can be carried out to home in on causal variants, and detailed knowledge of how HLA proteins work was helpful in focusing attention on key amino acids involved in the HLA associations with rheumatoid arthritis. The genetic determinants were expected to be variant amino acids involved in presenting autoantigens and that meant the amino acids involved in peptide binding were already known. The HLA peptide-binding sites form well-established grooves in the protein and that meant that the identities of the key amino acids that are important in peptide-binding. Fine-scale association mapping has shown that most of the HLA association with rheumatoid arthritis is due to five amino acids, all located in the peptide-binding grooves. Three occur in the HLA-DRβ1 chain at amino acid positions 11, 71 and 74, one occurs in HLA-B at amino acid position 9, and one at amino acid position 9 in HLA-DPβ1.

Question 8.39

Naturally occurring genetic variants are known to act as protective factors to confer reduced risk of infectious disease. List some examples.

Answer 8.39

- Possibly the most celebrated example, because of the therapeutic potential is the 32 bp deletion in the *CCR5* gene. This variant has been identified as a protective factor against developing HIV-AIDS, and is carried by about 10% of individuals of European descent. *CCR5* makes a chemokine receptor that works as a co-receptor with the CD4 receptor on helper T cells but the 32 bp deletion results in inactivation, and rare homozygotes for the 32 bp deletion are resistant to HIV-AIDS. (Box 9.7 on page 365 outlines the therapeutic potential)
- Certain genetic variants of the *ATP2B4* gene confer reduced risk of malaria. Red blood cells are the host cells for the pathogenic stage of the malarial parasite *Plasmodium falciparum* and the ATP2B4 gene is known to make a protein that works as the main calcium pump of red blood cells.
- The β^S-globin allele associated with HbS and sickle cell disease also confers resistance to *Plasmodium falciparum* malaria by making the red blood cell an inhospitable environment for the *P. falciparum* parasite.

- Genetic variants that inactivate the gene encoding the Duffy blood group antigen act as a protective factors to reduce the risk of *Plasmodium vivax* malaria. The Duffy blood group gene makes a chemokine receptor that is needed for *Plasmodium vivax* to gain access to red blood cells.

Question 8.40

Genomewide association studies have been very successful in identifying susceptibility factors in inflammatory bowel diseases, and have provided valuable insights into the pathogenesis of these diseases. What types of insights have they offered?

Answer 8.40

The genomewide (GWA) association studies on Cohn's disease and ulcerative caused a substantial rethink about the pathogenesis, and the importance of some of the implicated pathways came as a surprise. The GWA risk variants implicated, for example, as many as five genes with a role in autophagy in Crohn's disease (a lysosomal degradation pathway that naturally disposes of worn-out intracellular organelles and very large protein aggregates). The autophagy machinery is now known to interact with many different stress response pathways in cells, including those involved in controlling immune responses and inflammation.
Another striking—and unexpected—finding was the important role of interleukin-23 (IL-23) pathways in both types of inflammatory bowel disease. Tissue injury in these conditions had once been thought to be primarily mediated by classical helper T-cell populations, but GWA studies clearly implicated IL-23 and the activation of Th17 (a recently discovered subpopulation of helper T cells) with the resulting production of IL-17 and chronic inflammation.

Question 8.41

Outside of cancers and infectious diseases, various other complex diseases are known to be strongly influenced by environmental factors. Give three examples of environmental factors that are known to increase the risk of specific complex diseases other than cancers and infectious diseases.

Answer 8.41

1. Over-consumption and excess of fatty foods as a risk factor for type 2 diabetes
2. Smoking as a risk factor for coronary artery disease, Crohn's disease and age- related macular disease
3. Gut microorganisms in inflammatory bowel disease and type 1 diabetes.