5.1. Genetic and Physical Maps

The convention is to divide genome mapping methods into two categories.

- Genetic mapping is based on the use of genetic techniques to construct maps showing the positions of genes and other sequence features on a genome. Genetic techniques include cross-breeding experiments or, in the case of humans, the examination of family histories (pedigrees). Genetic mapping is described in Section 5.2.
- Physical mapping uses molecular biology techniques to examine DNA molecules directly in order to construct maps showing the positions of sequence features, including genes. Physical mapping is described in Section 5.3.

5.2. Genetic Mapping

As with any type of map, a genetic map must show the positions of distinctive features. In a geographic map these markers are recognizable components of the landscape, such as rivers, roads and buildings. What markers can we use in a genetic landscape?

5.2.1. Genes were the first markers to be used

The first genetic maps, constructed in the early decades of the 20th century for organisms such as the fruit fly, used genes as markers. This was many years before it was understood that genes are segments of DNA molecules. Instead, genes were looked upon as abstract entities responsible for the transmission of heritable characteristics from parent to offspring. To be useful in genetic analysis, a heritable characteristic has to exist in at least two alternative forms or phenotypes, an example being tall or short stems in the pea plants originally studied by Mendel. Each phenotype is specified by a different allele of the corresponding gene. To begin with, the only genes that could be studied were those specifying phenotypes that were distinguishable by visual examination. So, for example, the first fruit-fly maps showed the positions of genes for body color, eye color, wing shape and suchlike, all of these phenotypes being visible simply by looking at the flies with a low-power microscope or the naked eye. This approach was fine in the early days but geneticists soon realized that there were only a limited number of visual phenotypes whose inheritance could be studied, and in many cases their analysis was complicated because a single phenotype could be affected by more than one gene. For example, by 1922 over 50 genes had been mapped onto the four fruit-fly chromosomes, but nine of these were for eye color; in later research, geneticists studying fruit flies had to learn to distinguish between fly eyes that were colored red, light red, vermilion, garnet, carnation, cinnabar, ruby, sepia, scarlet, pink, cardinal, claret, purple or brown. To make gene maps more comprehensive it would be necessary to find characteristics that were more distinctive and less complex than visual ones.

The answer was to use biochemistry to distinguish phenotypes. This has been particularly important with two types of organisms - microbes and humans. Microbes, such as bacteria and yeast, have very few visual characteristics so gene mapping with these organisms has to rely on biochemical phenotypes such as those listed in Table 5.1. With humans it is possible to use

visual characteristics, but since the 1920s studies of human genetic variation have been based largely on biochemical phenotypes that can be scored by blood typing. These phenotypes include not only the standard blood groups such as the ABO series (Yamamoto *et al.*, 1990), but also variants of blood serum proteins and of immunological proteins such as the human leukocyte antigens (the HLA system). A big advantage of these markers is that many of the relevant genes have multiple alleles. For example, the gene called *HLA-DRB1* has at least 290 alleles and *HLA-B* has over 400. This is relevant because of the way in which gene mapping is carried out with humans (Section 5.2.4). Rather than setting up many breeding experiments, which is the procedure with experimental organisms such as fruit flies or mice, data on inheritance of human genes have to be gleaned by examining the phenotypes displayed by members of a single family. If all the family members have the same allele for the gene being studied then no useful information can be obtained. It is therefore necessary for the relevant marriages to have occurred, by chance, between individuals with different alleles. This is much more likely if the gene being studied has 290 rather than two alleles.

5.2.2. DNA markers for genetic mapping

Genes are very useful markers but they are by no means ideal. One problem, especially with larger genomes such as those of vertebrates and flowering plants, is that a map based entirely on genes is not very detailed. This would be true even if every gene could be mapped because, as we saw in Chapter 2, in most eukaryotic genomes the genes are widely spaced out with large gaps between them (see Figure 2.2). The problem is made worse by the fact that only a fraction of the total number of genes exist in allelic forms that can be distinguished conveniently. Gene maps are therefore not very comprehensive. We need other types of marker.

Mapped features that are not genes are called DNA markers. As with gene markers, a DNA marker must have at least two alleles to be useful. There are three types of DNA sequence feature that satisfy this requirement: restriction fragment length polymorphisms (RFLPs), simple sequence length polymorphisms (SSLPs), and single nucleotide polymorphisms (SNPs).

Restriction fragment length polymorphisms (RFLPs)

RFLPs were the first type of DNA marker to be studied. Recall that restriction enzymes cut DNA molecules at specific recognition sequences (Section 4.1.2). This sequence specificity means that treatment of a DNA molecule with a restriction enzyme should always produce the same set of fragments. This is not always the case with genomic DNA molecules because some restriction sites are polymorphic, existing as two alleles, one allele displaying the correct sequence for the restriction site and therefore being cut when the DNA is treated with the enzyme, and the second allele having a sequence alteration so the restriction site is no longer recognized. The result of the sequence alteration is that the two adjacent restriction fragments remain linked together after treatment with the enzyme, leading to a length polymorphism (Figure 5.4). This is an RFLP and its position on a genome map can be worked out by following the inheritance of its alleles, just as is done when genes are used as markers. There are thought to be about 10^5 RFLPs in the human genome, but of course for each RFLP there can only be two alleles (with and without the site). The value of RFLPs in human gene mapping is therefore limited by the high possibility that the RFLP being studied shows no variability among the members of an interesting family.

In order to score an RFLP, it is necessary to determine the size of just one or two individual restriction fragments against a background of many irrelevant fragments. This is not a trivial problem: an enzyme such as EcoRI, with a 6-bp recognition sequence, should cut approximately once every 46 = 4096 bp and so would give almost 800 000 fragments when used with human DNA. After separation by agarose gel electrophoresis (see Technical Note 2.1), these 800 000 fragments produce a smear and the RFLP cannot be distinguished. Southern hybridization, using a probe that spans the polymorphic restriction site, provides one way of visualizing the RFLP (Figure 5.5A), but nowadays PCR is more frequently used. The primers for the PCR are designed so that they anneal either side of the polymorphic site, and the RFLP is typed by treating the amplified fragment with the restriction enzyme and then running a sample in an agarose gel (Figure 5.5B).

Simple sequence length polymorphisms (SSLPs)

SSLPs are arrays of repeat sequences that display length variations, different alleles containing different numbers of repeat units (Figure 5.6A). Unlike RFLPs, SSLPs can be multi-allelic as each SSLP can have a number of different length variants. There are two types of SSLP, both of which were described

- Minisatellites, also known as variable number of tandem repeats (**VNTRs**), in which the repeat unit is up to 25 bp in length;
- Microsatellites or **simple tandem repeats** (**STRs**), whose repeats are shorter, usually dinucleotide or tetranucleotide units.

Microsatellites are more popular than minisatellites as DNA markers, for two reasons. First, minisatellites are not spread evenly around the genome but tend to be found more frequently in the telomeric regions at the ends of chromosomes. In geographic terms, this is equivalent to trying to use a map of lighthouses to find one's way around the middle of an island. Microsatellites are more conveniently spaced throughout the genome. Second, the quickest way to type a length polymorphism is by PCR (Figure 5.6B), but PCR typing is much quicker and more accurate with sequences less than 300 bp in length. Most minisatellite alleles are longer than this because the repeat units are relatively large and there tend to be many of them in a single array, so PCR products of several kb are needed to type them. Typical microsatellites consist of 10–30 copies of a repeat that is usually no longer than 4 bp in length, and so are much more amenable to analysis by PCR. There are 6.5×10^5 microsatellites in the human genome (see Table 1.3).

Single nucleotide polymorphisms (SNPs)

These are positions in a genome where some individuals have one nucleotide (e.g. a G) and others have a different nucleotide (e.g. a C) (Figure 5.7). There are vast numbers of SNPs in every genome, some of which also give rise to RFLPs, but many of which do not because the sequence in which they lie is not recognized by any restriction enzyme. In the human genome there are at least 1.42 million SNPs, only 100 000 of which result in an RFLP

A single nucleotide polymorphism (SNP).

Although each SNP could, potentially, have four alleles (because there are four nucleotides), most exist in just two forms, so these markers suffer from the same drawback as RFLPs with regard to human genetic mapping: there is a high possibility that a SNP does not display any variability in the family that is being studied. The advantages of SNPs are their abundant numbers and the fact that they can be typed by methods that do not involve gel electrophoresis. This is important because gel electrophoresis has proved difficult to automate so any detection method that uses it will be relatively slow and labor-intensive. SNP detection is more rapid because it is based on oligonucleotide hybridization analysis. An oligonucleotide is a short single-stranded DNA molecule, usually less than 50 nucleotides in length, that is synthesized in the test tube. If the conditions are just right, then an oligonucleotide will hybridize with another DNA molecule only if the oligonucleotide forms a completely base-paired structure with the second molecule. If there is a single mismatch - a single position within the oligonucleotide that does not form a base pair - then hybridization does not occur (Figure 5.8). Oligonucleotide hybridization can therefore discriminate between the two alleles of an SNP. Various screening strategies have been devised (Mir and Southern, 2000), including DNA chip technology (Technical Note 5.1) and **solution hybridization techniques.**

DNA microarrays and chips. High-density arrays of DNA molecules for parallel hybridization analyses. DNA microarrays and chips are designed to allow many hybridization experiments to be performed in parallel. Their main applications have been in the screening (more...)

- A DNA chip is a wafer of glass or silicon, 2.0 cm^2 or less in area, carrying many different oligonucleotides in a high-density array. The DNA to be tested is labeled with a fluorescent marker and pipetted onto the surface of the chip. Hybridization is detected by examining the chip with a fluorescence microscope, the positions at which the fluorescent signal is emitted indicating which oligonucleotides have hybridized with the test DNA. Many SNPs can therefore be scored in a single experiment (Wang *et al*., 1998; Gerhold *et al*., 1999).
- **Solution hybridization techniques** are carried out in the wells of a microtiter tray, each well containing a different oligonucleotide, and use a detection system that can discriminate between unhybridized single-stranded DNA and the double-stranded product that results when an oligonucleotide hybridizes to the test DNA. Several systems have been developed, one of which makes use of a pair of labels comprising a fluorescent dye and a compound that quenches the fluorescent signal when brought into close proximity with the dye. The dye is attached to one end of an oligonucleotide and the quenching compound to the other end. Normally there is no fluorescence because the oligonucleotide is designed in such a way that the two ends base-pair to one another, placing the quencher next to the dye (Figure 5.9). Hybridization between oligonucleotide and test DNA disrupts this base pairing, moving the quencher away from the dye and enabling the fluorescent signal to be generated (Tyagi *et al*., 1998).

Figure 5.9

One way of detecting an SNP by solution hybridization. The oligonucleotide probe has two end-labels. One of these is a fluorescent dye and the other is a quenching compound. The two ends of the oligonucleotide base-pair to one another, so the fluorescent (more...)

5.2.3. Linkage analysis is the basis of genetic mapping

Now that we have assembled a set of markers with which to construct a genetic map we can move on to look at the mapping techniques themselves. These techniques are all based on genetic linkage, which in turn derives from the seminal discoveries in genetics made in the mid 19th century by Gregor Mendel.

The principles of inheritance and the discovery of linkage

Genetic mapping is based on the principles of inheritance as first described by Gregor Mendel in 1865 (Orel, 1995). From the results of his breeding experiments with peas, Mendel concluded that each pea plant possesses two alleles for each gene, but displays only one phenotype. This is easy to understand if the plant is pure-breeding, or homozygous, for a particular characteristic, as it then possesses two identical alleles and displays the appropriate phenotype (Figure 5.10A). However, Mendel showed that if two pure-breeding plants with different phenotypes are crossed then all the progeny (the F_1 generation) display the same phenotype. These F_1 plants must be heterozygous, meaning that they possess two different alleles, one for each phenotype, one allele inherited from the mother and one from the father. Mendel postulated that in this heterozygous condition one allele overrides the effects of the other allele; he therefore described the phenotype expressed in the F_1 plants as being dominant over the second, recessive phenotype (Figure 5.10B). This is the perfectly correct interpretation of the interaction between the pairs of alleles studied by Mendel, but we now appreciate that this simple dominant-recessive rule can be complicated by situations that he did not encounter. One of these is incomplete dominance, where the heterozygous phenotype is intermediate between the two homozygous forms. An example is when red carnations are crossed with white ones, the F_1 heterozygotes being pink. Another complication is codominance, when both alleles are detectable in the heterozygote. Codominance is the typical situation for DNA markers.

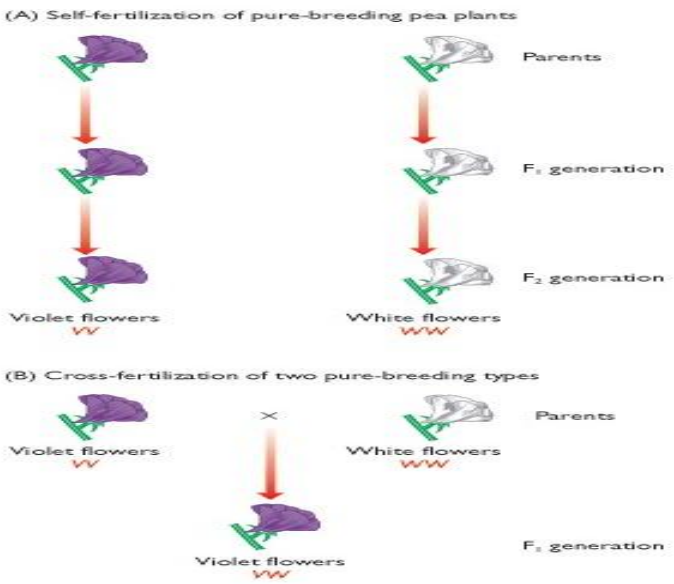

Figure 5.10 Homozygosity and heterozygosity

As well as discovering dominance and recessiveness, Mendel carried out additional crosses that enabled him to establish two Laws of Genetics. The First Law states that ***alleles segregate randomly***. In other words, if the parent's alleles are *A* and *a*, then a member of the F_1 generation has the same chance of inheriting *A* as it has of inheriting *a* (Figure 5.11A). The Second Law is that ***pairs of alleles segregate independently***, so that inheritance of the alleles of gene A is independent of inheritance of the alleles of gene B (Figure 5.11B). Because of these laws, the outcomes of genetic crosses are predictable (Figure 5.11C).

Figure 5.11

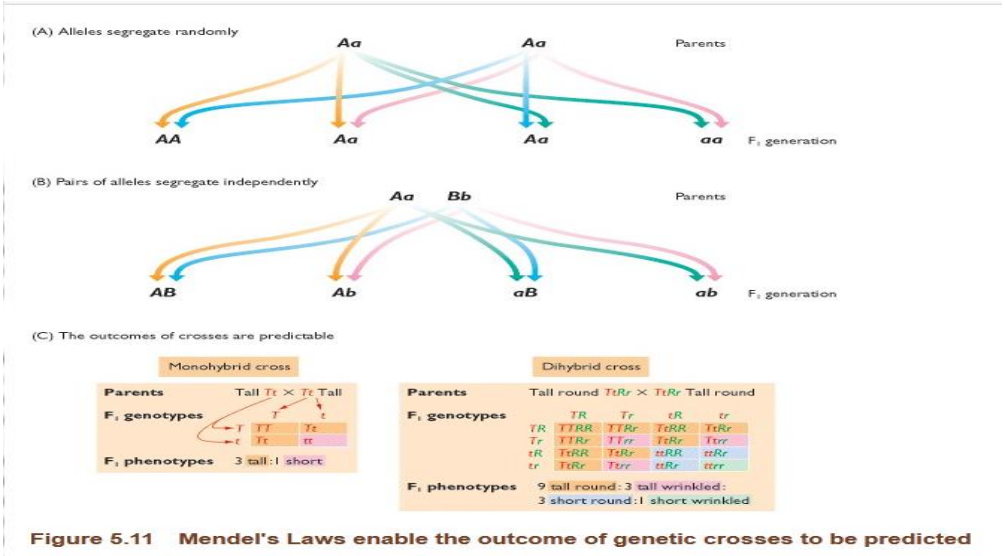

Figure 5.11 Mendel's Laws enable the outcome of genetic crosses to be predicted

Mendel's Laws enable the outcome of genetic crosses to be predictedWhen Mendel's work was rediscovered in 1900, his Second Law worried the early geneticists because it was soon established that genes reside on chromosomes, and it was realized that all organisms have many more genes than chromosomes. Chromosomes are inherited as intact units, so it was reasoned that the alleles of some pairs of genes will be inherited together because they are on the same chromosome (Figure 5.12). This is the principle of genetic linkage, and it was quickly shown to be correct, although the results did not turn out exactly as expected. The complete linkage that had been anticipated between many pairs of genes failed to materialize. Pairs of genes were either inherited independently, as expected for genes in different chromosomes, or, if they showed linkage, then it was only partial linkage: sometimes they were inherited together and sometimes they were not (Figure 5.13). The resolution of this contradiction between theory and observation was the critical step in the development of genetic mapping techniques.

Figure 5.12

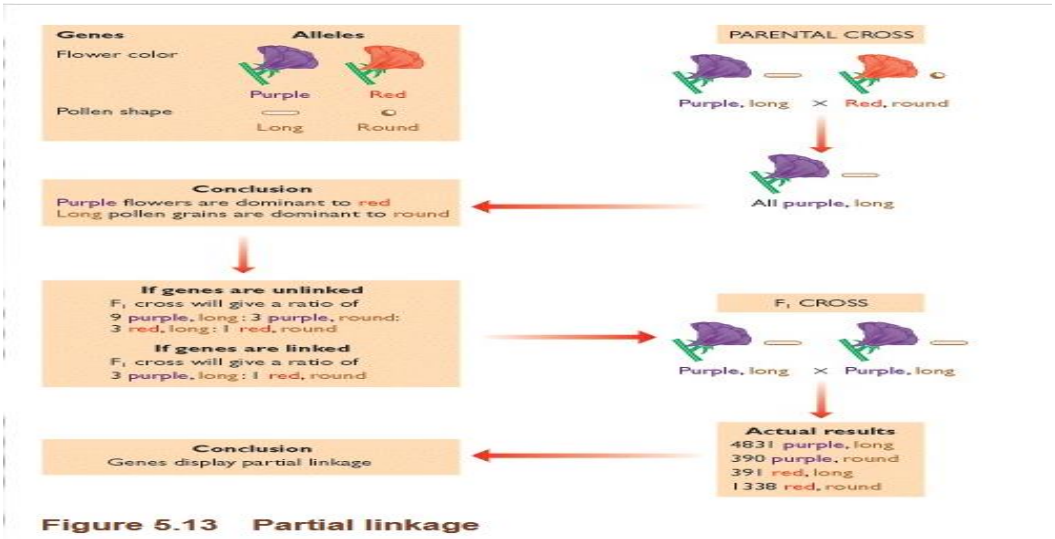

Figure 5.13 Partial linkage

Figure 5.13

Partial linkage.

Partial linkage is explained by the behavior of chromosomes during meiosis

The critical breakthrough was achieved by Thomas Hunt Morgan, who made the conceptual leap between partial linkage and the behavior of chromosomes when the nucleus of a cell divides. Cytologists in the late 19th century had distinguished two types of nuclear division: mitosis and meiosis. Mitosis is more common, being the process by which the diploid nucleus of a somatic

cell divides to produce two daughter nuclei, both of which are diploid (Figure 5.14). Approximately 10^{17} mitoses are needed to produce all the cells required during a human lifetime. Before mitosis begins, each chromosome in the nucleus is replicated, but the resulting daughter chromosomes do not immediately break away from one another. To begin with they remain attached at their centromeres and by cohesin proteins which act as 'molecular glue' holding together the arms of the replicated chromosomes (see Figure 13.23). The daughters do not separate until later in mitosis when the chromosomes are distributed between the two new nuclei. Obviously it is important that each of the new nuclei receives a complete set of chromosomes, and most of the intricacies of mitosis appear to be devoted to achieving this end.

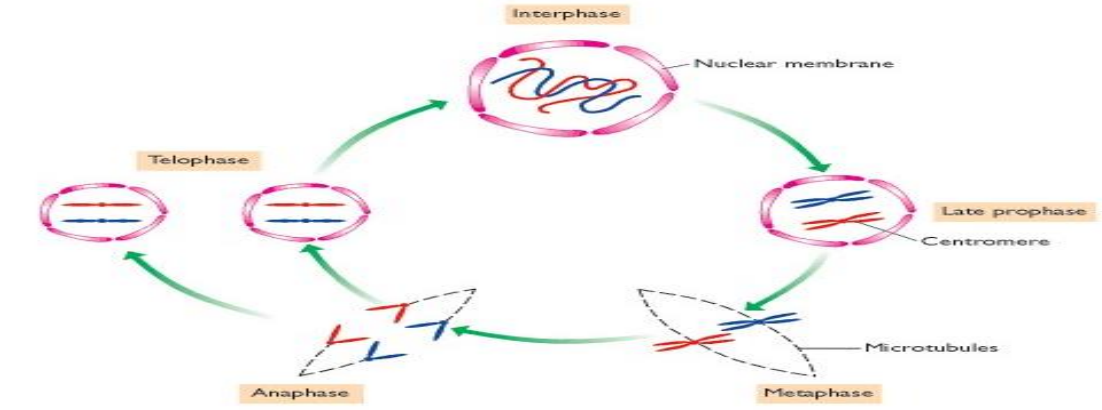

Figure 5.14 Mitosis

Figure 5.14

Mitosis

Mitosis illustrates the basic events occurring during nuclear division but is not directly relevant to genetic mapping. Instead, it is the distinctive features of meiosis that interest us. Meiosis occurs only in reproductive cells, and results in a diploid cell giving rise to four haploid gametes, each of which can subsequently fuse with a gamete of the opposite sex during sexual reproduction. The fact that meiosis results in four haploid cells whereas mitosis gives rise to two diploid cells is easy to explain: meiosis involves two nuclear divisions, one after the other, whereas mitosis is just a single nuclear division. This is an important distinction, but the critical difference between mitosis and meiosis is more subtle. Recall that in a diploid cell there are two separate copies of each chromosome (Chapter 1). We refer to these as pairs of homologous chromosomes. During mitosis, homologous chromosomes remain separate from one another, each member of the pair replicating and being passed to a daughter nucleus independently of its homolog. In meiosis, however, the pairs of homologous chromosomes are by no means independent. During meiosis I, each chromosome lines up with its homolog to form a bivalent (Figure 5.15). This occurs after each chromosome has replicated, but before the replicated structures split, so the bivalent in fact contains four chromosome copies, each of which is destined to find its way into one of the four gametes that will be produced at the end of the meiosis. Within the bivalent, the chromosome arms (the chromatids) can undergo physical

breakage and exchange of segments of DNA. The process is called crossing-over or recombination and was discovered by the Belgian cytologist Janssens in 1909. This was just 2 years before Morgan started to think about partial linkage.

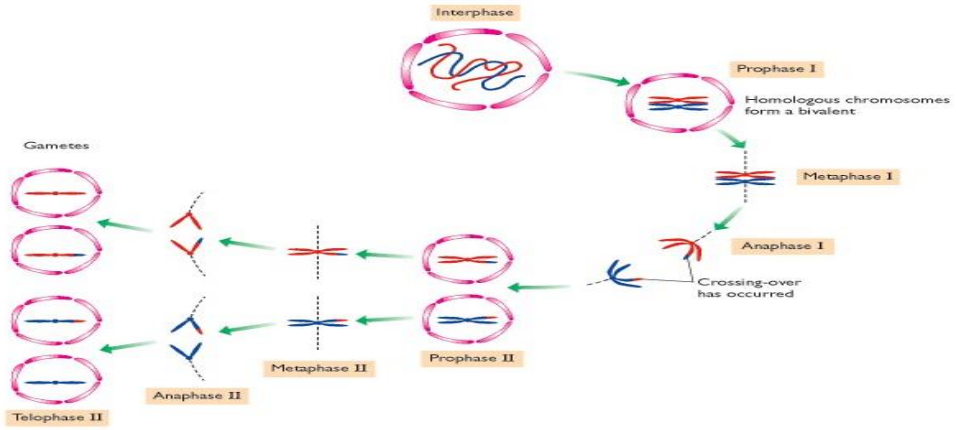

Figure 5.15 Meiosis

Meiosis.

How did the discovery of crossing-over help Morgan explain partial linkage? To understand this we need to think about the effect that crossing-over can have on the inheritance of genes. Let us consider two genes, each of which has two alleles. We will call the first gene A and its alleles *A* and *a*, and the second gene B with alleles *B* and *b*. Imagine that the two genes are located on chromosome number 2 of *Drosophila melanogaster*, the species of fruit fly studied by Morgan. We are going to follow the meiosis of a diploid nucleus in which one copy of chromosome 2 has alleles *A* and *B*, and the second has *a* and *b*. This situation is illustrated in Figure 5.16. Consider the two alternative scenarios:

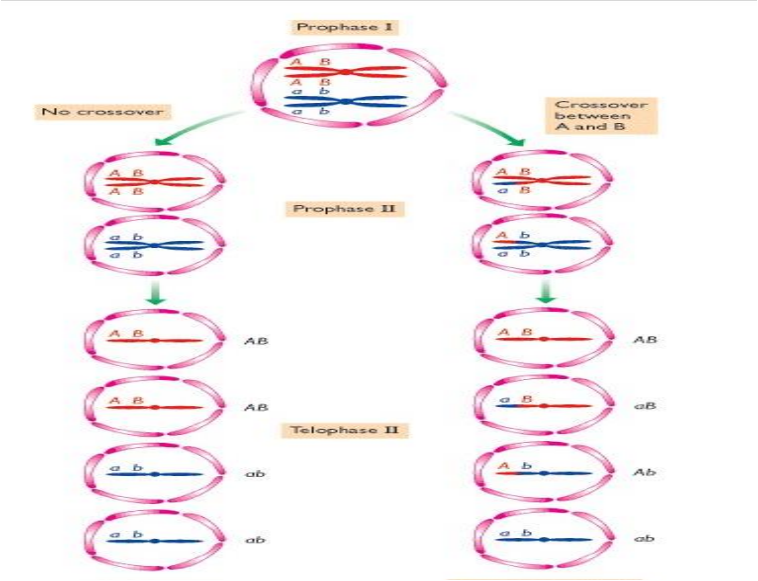

Figure 5.16

The effect of a crossover on linked genes. The drawing shows a pair of homologous chromosomes, one red and the other blue. A and B are linked genes with alleles *A*, *a*, *B* and *b*. On the left is a meiosis with no crossover between A and B: two of the resulting (more...)

1.***A crossover does not occur between genes A and B.*** If this is what happens then two of the resulting gametes will contain chromosome copies with alleles *A* and *B*, and the other two will contain *a* and *b*. In other words, two of the gametes have the genotype *AB* and two have the genotype *ab*.

2.***A crossover does occur between genes A and B.*** This leads to segments of DNA containing gene B being exchanged between homologous chromosomes. The eventual result is that each gamete has a different genotype: 1 *AB*, 1 *aB*, 1 *Ab*, 1 *ab*.

Now think about what would happen if we looked at the results of meiosis in a hundred identical cells. If crossovers never occur then the resulting gametes will have the following genotypes:

This is complete linkage: genes A and B behave as a single unit during meiosis. But if (as is more likely) crossovers occur between A and B in some of the nuclei, then the allele pairs will not be inherited as single units. Let us say that crossovers occur during 40 of the 100 meioses. The following gametes will result:

The linkage is not complete, it is only partial. As well as the two **parental** genotypes (*AB*, *ab*) we see gametes with recombinant genotypes (*Ab*, *aB*).

From partial linkage to genetic mapping

Once Morgan had understood how partial linkage could be explained by crossing-over during meiosis he was able to devise a way of mapping the relative positions of genes on a chromosome. In fact the most important work was done not by Morgan himself, but by an undergraduate in his laboratory, Arthur Sturtevant (Sturtevant, 1913). Sturtevant assumed that crossing-over was a random event, there being an equal chance of it occurring at any position along a pair of lined-up chromatids. If this assumption is correct then two genes that are close together will be separated by crossovers less frequently than two genes that are more distant from one another. Furthermore, the frequency with which the genes are unlinked by crossovers will be directly proportional to how far apart they are on their chromosome. The recombination frequency is therefore a measure of the distance between two genes. If you work out the recombination frequencies for different pairs of genes, you can construct a map of their relative positions on the chromosome (Figure 5.17).

Figure 5.17

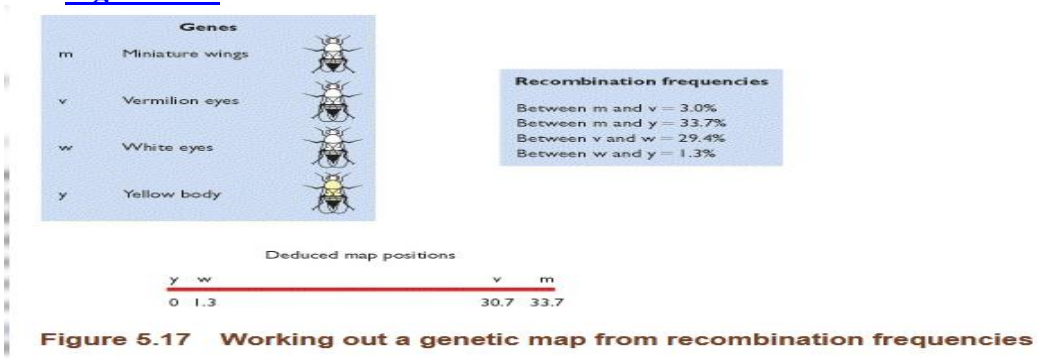

Figure 5.17 Working out a genetic map from recombination frequencies

Working out a genetic map from recombination frequencies.

It turns out that Sturtevant's assumption about the randomness of crossovers was not entirely justified. Comparisons between genetic maps and the actual positions of genes on DNA molecules, as revealed by physical mapping and DNA sequencing, have shown that some regions of chromosomes, called recombination hotspots, are more likely to be involved in crossovers than others. This means that a genetic map distance does not necessarily indicate the physical distance between two markers (see Figure 5.22). Also, we now realize that a single chromatid can participate in more than one crossover at the same time, but that there are limitations on how close together these crossovers can be, leading to more inaccuracies in the mapping procedure. Despite these qualifications, linkage analysis usually makes correct deductions about gene order, and distance estimates are sufficiently accurate to generate genetic maps that are of value as frameworks for genome sequencing projects.

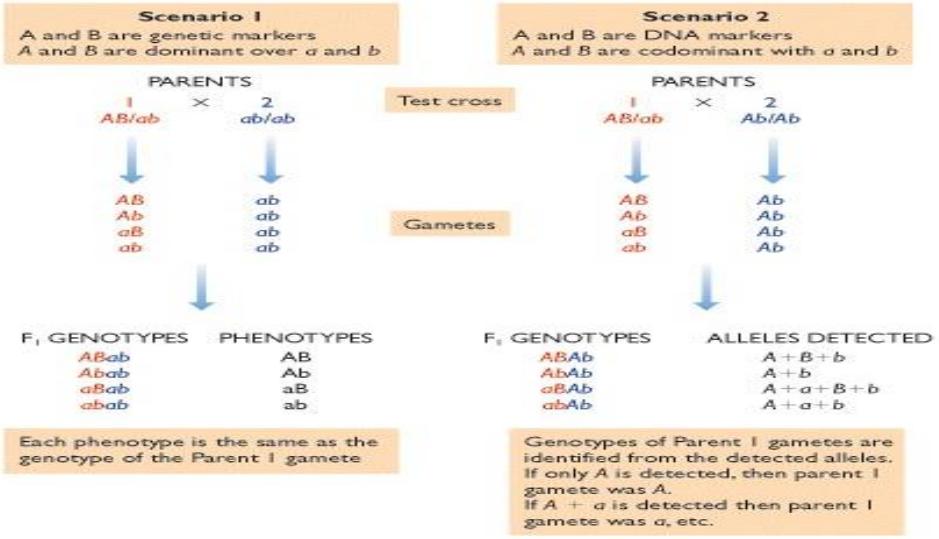

Figure 5.18 Two examples of the test cross

Linkage analysis with different types of organism

To see how linkage analysis is actually carried out, we need to consider three quite different situations:

- Linkage analysis with species such as fruit flies and mice, with which we can carry out planned breeding experiments;
- Linkage analysis with humans, with whom we cannot carry out planned experiments but instead make use of family pedigrees;
- Linkage analysis with bacteria, which do not undergo meiosis.

Linkage analysis when planned breeding experiments are possible

The first type of linkage analysis is the modern counterpart of the method developed by Morgan and his colleagues. The method is based on analysis of the progeny of experimental crosses set up between parents of known genotypes and is, at least in theory, applicable to all eukaryotes. Ethical considerations preclude this approach in humans, and practical problems such as the length of the gestation period and the time taken for the newborn to reach maturity (and hence to participate in subsequent crosses) limit the effectiveness of the method with some animals and plants.

If we return to Figure 5.16 we see that the key to gene mapping is being able to determine the genotypes of the gametes resulting from meiosis. In a few situations this is possible by directly examining the gametes. For example, the gametes produced by some microbial eukaryotes, including the yeast *Saccharomyces cerevisiae*, can be grown into colonies of haploid cells, whose genotypes can be determined by biochemical tests. Direct genotyping of gametes is also

possible with higher eukaryotes if DNA markers are used, as PCR can be carried out with the DNA from individual spermatozoa, enabling RFLPs, SSLPs and SNPs to be typed. Unfortunately, sperm typing is laborious. Routine linkage analysis with higher eukaryotes is therefore carried out not by examining the gametes directly but by determining the genotypes of the diploid progeny that result from fusion of two gametes, one from each of a pair of parents. In other words, a genetic cross is performed.

The complication with a genetic cross is that the resulting diploid progeny are the product not of one meiosis but of two (one in each parent), and in most organisms crossover events are equally likely to occur during production of the male and female gametes. Somehow we have to be able to disentangle from the genotypes of the diploid progeny the crossover events that occurred in each of these two meioses. This means that the cross has to be set up with care. The standard procedure is to use a test cross. This is illustrated in Figure 5.18, Scenario 1, where we have set up a test cross to map the two genes we met earlier: gene A (alleles *A* and *a*) and gene B (alleles *B* and *b*), both on chromosome 2 of the fruit fly. The critical feature of a test cross is the genotypes of the two parents:

Gene mapping by human pedigree analysis

With humans it is of course impossible to pre-select the genotypes of parents and set up crosses designed specifically for mapping purposes. Instead, data for the calculation of recombination frequencies have to be obtained by examining the genotypes of the members of successive generations of existing families. This means that only limited data are available, and their interpretation is often difficult because a human marriage rarely results in a convenient test cross, and often the genotypes of one or more family members are unobtainable because those individuals are dead or unwilling to cooperate.

The problems are illustrated by Figure 5.19. In this example we are studying a genetic disease present in a family of two parents and six children. Genetic diseases are frequently used as gene markers in humans, the disease state being one allele and the healthy state being a second allele. The pedigree in Figure 5.19A shows us that the mother is affected by the disease, as are four of her children. We know from family accounts that the maternal grandmother also suffered from this disease, but both she and her husband - the maternal grandfather - are now dead. We can include them in the pedigree, with slashes indicating that they are dead, but we cannot obtain any further information on their genotypes. Our aim is to map the position of the gene for the genetic disease. For this purpose we are studying its linkage to a microsatellite marker M, four alleles of which - M_1, M_2, M_3 and M_4 - are present in the living family members. The question is, how many of the children are recombinants?

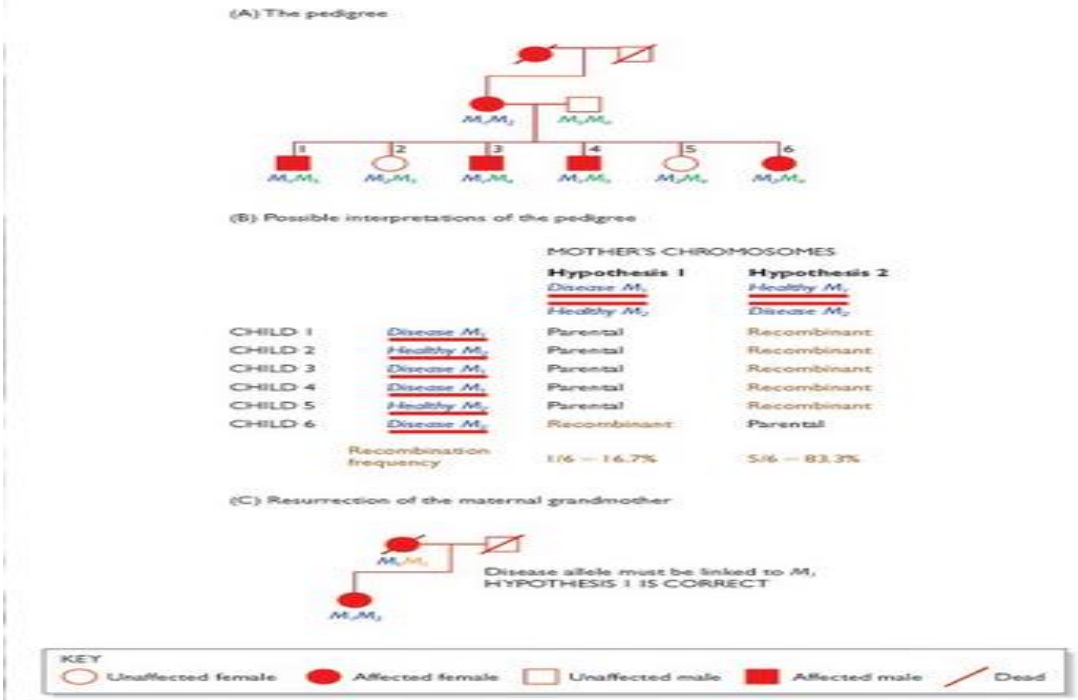

Figure 5.19 An example of human pedigree analysis