15 How to use linkage to map a disease gene

The aim is to discover the chromosomal location of the gene responsible for a mendelian condition. This is the first step in the 'positional cloning' process that was used to identify the genes underlying most of the more frequent mendelian conditions during the 1980s and 1990s. As we show below, it depends on being able to recruit large multi-case families for DNA analysis. The near-disappearance of such family studies (apart from autozygosity mapping, see *Section 8.2*) from current human genetic research is not so much because they have been rendered obsolete by advances in sequencing, but more because virtually every case where family studies were possible had already been studied and the disease gene identified. What remained were the excessively rare or sporadic conditions where family studies were impossible.

Linkage analysis remains a basic technique for gene mapping; here we show how it works, using the example of dyschromatosis symmetrica hereditaria (DSH), an autosomal dominant condition (OMIM 127400, see *Figure 15.1*). Although this condition is rare and is not particularly troublesome clinically, we chose it because all the steps in this classical route for gene identification are described in a single paper in a source that is likely to be accessible to most readers (Miyamura *et al.*, 2003). For more detail on linkage analysis see Ott (1999), Strachan and Read (2019), or any genetics textbook.

Figure 15.1 – Dyschromatosis symmetrica hereditaria.
Patients with DSH have small hyperpigmented and hypopigmented macules on the backs of their hands and the dorsal surface of their feet. The abnormalities are otherwise asymptomatic and do not affect the general health of the person.
Photos courtesy of Drs Tamio Suzuki and Yasushi Tomita, Nagoya University.

First, we must discuss the concepts of recombinants and non-recombinants.

Recombinants and non-recombinants

This section reinforces material that was covered rather more briefly in *Chapter 8*.

Consider how alleles at two loci might segregate through a family (*Figure 15.2*).

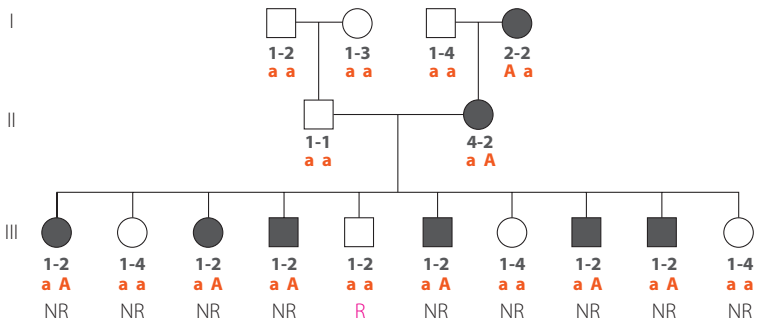

Figure 15.2 – A family in which alleles at two loci are segregating.
There is an autosomal dominant condition with alleles A, a, where affected individuals are shaded, and a 4-allele microsatellite (see *Box 8.1*) for which genotypes are shown. Where parental origin of each allele can be determined, the paternal alleles are shown on the left below each symbol and the maternal alleles on the right. The text below explains how individuals in generation III can be classified as non-recombinant (NR) or recombinant (R).

Consider the *combination* of alleles that individual II–2 received from her parents. From her father she inherited allele 4 of the microsatellite and the non-affected allele a of the condition. From her mother she inherited allele 2 of the marker and the affected allele (A) for the condition. Now look at her 10 children. Their father (II–1) is homozygous for allele 1 of the marker and the a (non-affected) allele of the condition, so each of his children inherited those two alleles from him. Knowing this, we can work out what each inherited from their mother. Individuals 1, 3, 4, 6, 8 and 9 inherited microsatellite allele 2 and A. This is the same combination that their mother inherited from her own mother. Individuals 2, 7 and 10 inherited 4 and a, which is the combination their mother inherited from her father. All of these individuals inherited one or other of the grandparental combinations, and are **non-recombinant**. But individual 5 inherited from his mother 2 and a, which is not a grandparental combination. This individual (or, strictly speaking, the maternal meiosis that produced the egg that gave rise to him) is **recombinant**.

All genetic mapping depends on the behavior of chromosomes at meiosis (see *Figure 2.7*). Genetic mapping works by studying two (or more) loci in a set of independent meioses and estimating the proportion of meioses where there was recombination between them. Non-homologous chromosomes assort independently. This means that alleles at loci on different chromosomes have a 50% chance of ending up in the same gamete (non-recombinant) and a 50% chance of ending up in different gametes (recombinant). If the recombination fraction is significantly below 50%, the loci are said to be linked, and should be on the same chromosome. Loci that are on the same chromosome will travel together unless a crossover between the paired homologous chromosomes separates

them (*Figure 15.3*). They will be separated if a crossover occurs at a position between the two loci. This will happen often to widely separated loci, but only rarely to those close together. Thus the chance of recombination between two loci is a measure of the distance between them. Unlinked loci may be on different chromosomes or may be on the same chromosome but distant from each other. Studying a series of linked loci and estimating the pairwise recombination fractions allows the loci to be arranged in order along a genetic map.

Loci that are separated by recombination in 1% of meioses are defined as being 1 centiMorgan (cM) apart (the unit is named after TH Morgan). Genetic distances, defined in this way and measured in cM, are not the same as physical distances, measured in bp, kb or Mb of DNA. The two would correspond exactly only if the chance of a crossover were identical in every stretch of a chromosome. In fact, some regions have a higher frequency of crossovers than others. The order of loci should still be the same on genetic and physical maps, but the spacing may be different. As a rule of thumb, 1 cM corresponds to 1 Mb – but there are considerable local variations.

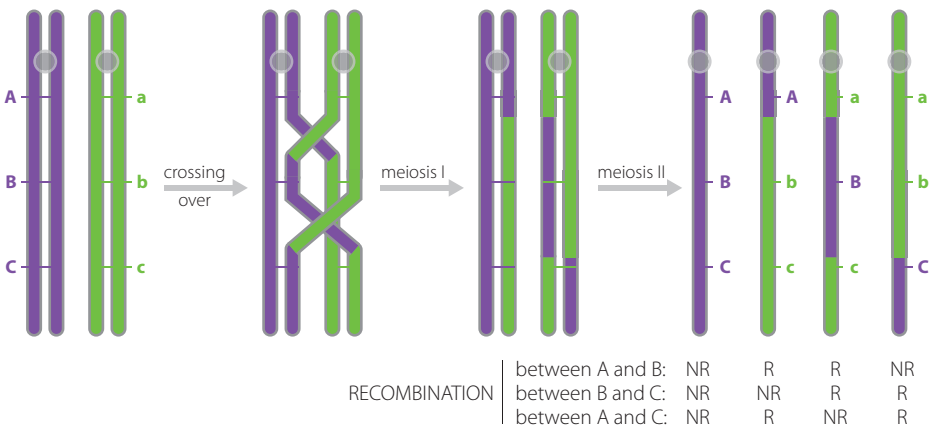

Figure 15.3 – In meiosis homologous chromosomes pair together, and then exchange random matching segments (crossing over). Here we show the effect of two crossovers on the segregation of alleles at three loci for which the individual is heterozygous.

Before we consider how these principles can be applied to human genetic mapping, we describe a breeding experiment in *Drosophila* fruit flies, the organism in which the concepts of genetic mapping were developed by TH Morgan and his group in his famous 'fly room' at Columbia University around 1910–1920. This illustrates the concepts without the statistical complexities of linkage analysis in humans.

Genetic mapping: an example from *Drosophila*

Before we get into the complexities of genetic mapping in humans, here is an example (from Griffiths *et al.*, 1999) from fruit flies, where one can set up crosses in any way desired and score thousands of offspring as recombinant or non-recombinant.

Three recessive characters were studied:

- vermilion (v) vs. normal (red, v^+) eyes

- crossveinless (cv) vs. normal (cv^+) wings
- cut (ct) vs. normal (ct^+) wing border

The available fly stocks were:

- red eyes, crossveinless and cut wings (v^+/v^+, cv/cv, ct/ct)
- vermilion eyes, normal wings (v/v, cv^+cv^+, ct^+/ct^+)

(a) Two lines were produced for the mapping experiment:
- triply heterozygous, v/v^+, cv/cv^+, ct/ct^+ with an all normal phenotype (progeny of crossing flies of the two original stocks)
- triply recessive v/v, cv/cv, ct/ct with vermilion eyes, crossveinless and cut wings (established by selecting progeny of the triple heterozygous flies that showed all three recessive characters; although these must be recombinant flies, we are not interested in counting recombinants here, we just want to set up a line of flies with the triple recessive phenotype).

(b) Triple heterozygote females are crossed with male flies homozygous for all three recessive characters, and the offspring phenotyped:
- The male parent contributed the three recessive alleles, so the phenotype of each progeny fly gives a direct readout of the genotype of the maternal gamete that produced it.

(c) For each pair of loci, recombinants are identified and the recombination fraction calculated. Gametes are recombinant if they contain a combination of alleles different from the parental combinations (v^+, cv, ct) and (v, cv^+, ct^+) – see the way the triply heterozygous females were obtained.
Results from 1448 flies were:

- v^+, cv, ct (580) and v, cv^+, ct^+ (592): non-recombinant ($n = 1172$)
- v, cv, ct^+ (45) and v^+, cv^+, ct (40): recombinant between (v, ct) and cv loci ($n = 85$)
- v, cv, ct (89) and v^+, cv^+, ct^+ (94): recombinant between v and (cv, ct) loci ($n = 183$)
- v, cv^+, ct (3) and v^+, cv, ct^+ (5): recombinant between (v, cv) and ct loci ($n = 8$)

Note that only 276 of the 1448 gametes (19%) from the mothers are recombinant. Clearly the three loci are linked.

(d) Combining the data establishes the order and genetic distances of the loci. The rarest class must be those that require a double recombination to produce them. This establishes that the gene order must be v – ct – cv.
Next, we count recombinants in each interval and calculate the recombination fraction:

- Between v and ct: 183 + 8 = 191/1448 = 13.2%
- Between ct and cv: 85 + 8 = 93/1448 = 6.4%
- Between v and cv: 183 + 85 + (2 x 8) = 284/1448 = 19.6%

The map established by this experiment is: v – 13.2 cM – ct – 6.4 cM – cv.

Genetic mapping in humans

The principle is exactly the same as in the *Drosophila* example above: we wish to calculate the recombination fraction between the *DSH* locus and each of a large number of microsatellite markers spread across the entire genome. The complications arise because we cannot set up ideal crosses as in *Drosophila*, and score hundreds of children; we have

to take families as we find them, often the grandparents are not available or parental genotypes are such that recombinants cannot be counted directly, and families are small.

The first step in the investigation was to ascertain several large families where the condition was segregating, and to obtain DNA from as many family members, both affected and unaffected, as possible (*Figure 15.4*). The DNA samples were typed for 343 microsatellite markers spaced across the genome.

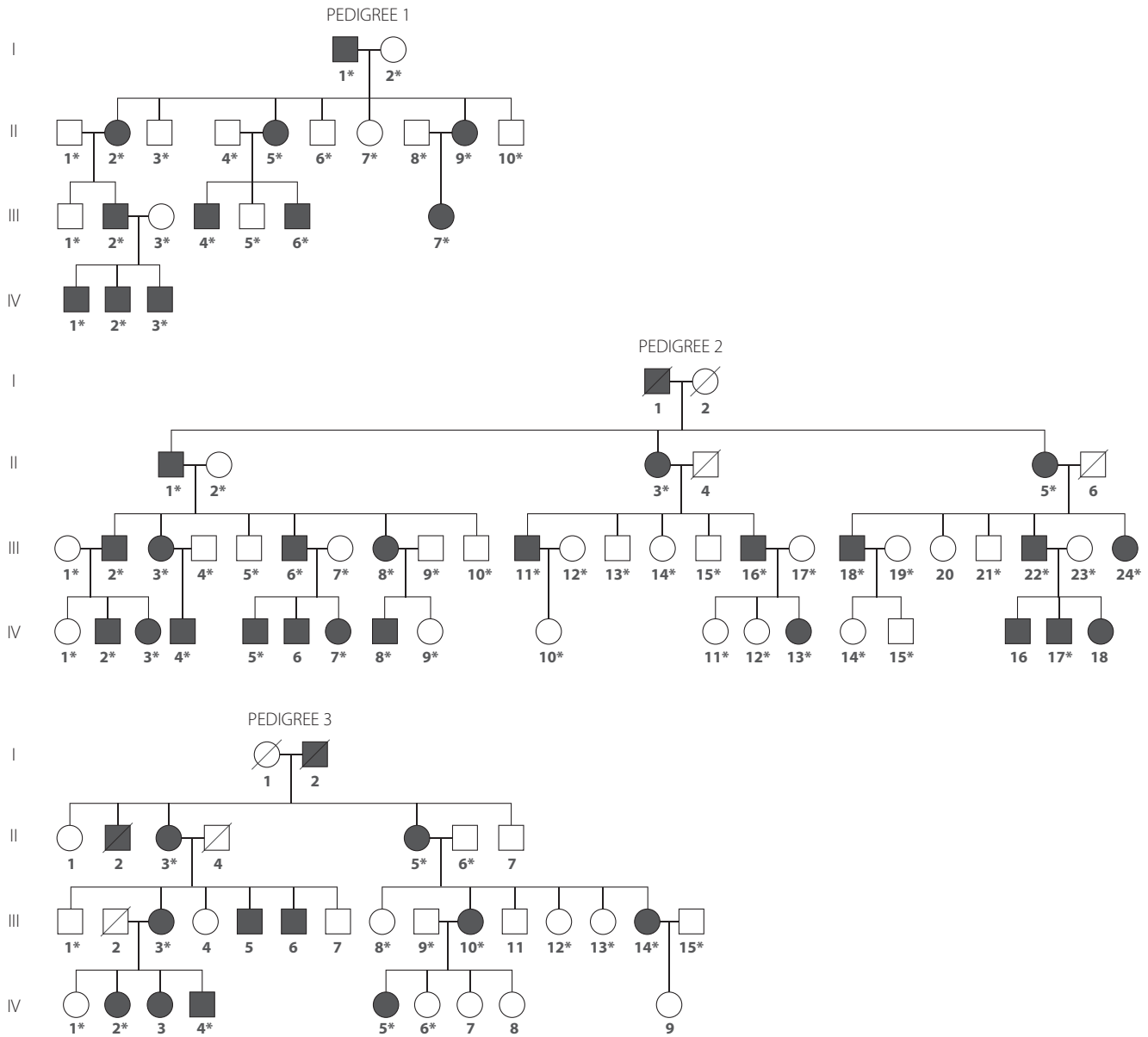

Figure 15.4 – Pedigree of three families in which DSH was segregating were studied to map the locus responsible.

Individuals who provided DNA are marked by asterisks. Adapted from Miyamura et al. (2003) with permission from Elsevier.

The problem of identifying recombinants in human pedigrees

In the pedigree in *Figure 15.2* we could immediately identify that individual III–5 was recombinant and all his sibs were non-recombinant. Often, however, it is not so easy. Suppose DNA from the grandparents was not available – maybe they were deceased, lived far away or were unwilling to provide samples (*Figure 15.5*).

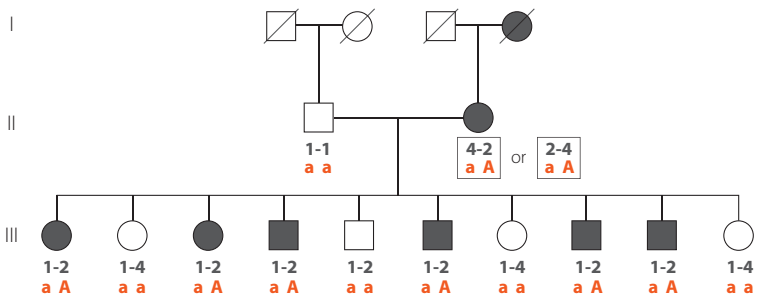

Figure 15.5 – The same family as in Figure 15.2, but this time grandparental DNA is not available.
Now we cannot tell which of the two alternative boxed combinations of alleles ('phases') is correct for individual II–2. We can still identify the combination she passed on to each of her children in generation III, but we can no longer tell whether there are 9 non-recombinants and 1 recombinant, or 9 recombinants and 1 non-recombinant.

Now we can no longer unambiguously identify recombinants, although clearly this pedigree contains useful linkage information. It could easily get worse. Further enquiries have located the estranged sister of individual II–2. It turns out she also had the family condition, but died some years ago from an unrelated cause; her partner abandoned her many years ago and is not traceable. However, she had three children who have now been traced and are willing to give samples for the study. Two of them have the family condition, the third is unaffected. The two affected people type 2-1 for the microsatellite marker, their unaffected sister types 4-1. This is surely a valuable addition to the pedigree, but since neither of their parents can be genotyped it does not provide definite non-recombinant meioses, it just intuitively strengthens the hypothesis of linkage.

Researchers seeking to map a rare condition could not ignore pedigrees like this one, but they badly needed some method of extracting the linkage information that they contain which does not depend on simply counting recombinant and non-recombinant meioses. Lod scores are the method they need.

The solution to the problem: lod scores

The word 'lod' is a contraction of 'logarithm of the odds'. The odds in question are the odds, given all the pedigree data, that the two loci are linked with a certain specified recombination fraction, compared to the chance they are not linked. Back in 1955 Newton Morton showed, in a blockbuster paper, that the lod score was the most efficient way of extracting linkage information from the sort of non-ideal pedigrees we have just sketched, and provided formulae to calculate it for various fairly simple pedigree structures. Positive lod scores are evidence in favor of linkage, negative scores are evidence against. The

threshold for significance is a lod score of +3.0. The reasoning behind this is explained below. We can illustrate the process using the family in *Figure 15.2*.

In this family we can score 9 meioses as non-recombinant and 1 as recombinant. Suppose the loci are linked with recombination fraction θ. The chance any meiosis would be recombinant is θ, and the chance it would be non-recombinant is 1–θ. The overall likelihood is therefore $(1-\theta)^9 \times \theta$. If, on the other hand, the loci are unlinked, the chance of any meiosis being either recombinant or non-recombinant is ½. The lod score is therefore

$$\log [((1-\theta)^9 \times \theta) / (\tfrac{1}{2})^{10}]$$

We can make a table of the lod scores for different values of θ:

Recombination fraction, θ	0	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
Lod score, Z	–infinity	1.51	1.60	1.55	1.44	1.28	1.09	0.62	0

The maximum lod score is seen if the recombination fraction is 0.1 – as expected intuitively since we have 1 recombinant out of 10 meioses. Nevertheless, a score of 1.60 falls below the threshold for statistical significance, which is a score of 3.0. Thus this pedigree supports the hypothesis that the two loci are linked, but only suggestively. Note that if you used a chi-squared test you would get a significant value – but this is not the appropriate test to use, as explained below.

We can also calculate the lod score for the modified pedigree in *Figure 15.5*. *A priori*, either phase for the mother II–2 is equally likely, so the likelihood, given linkage, is

$$\tfrac{1}{2} [(1-\theta)^9 \times \theta] + \tfrac{1}{2} [(\theta^9 \times (1-\theta))]$$

The table of lod scores is now:

Recombination fraction, θ	0	0.05	0.1	0.15	0.2	0.25	0.3	0.4	0.5
Lod score, Z	–infinity	1.21	1.30	1.25	1.14	0.98	0.79	0.33	0

Because of the uncertainty, the lod scores are all lower.

The threshold of significance

A lod score of +3.0 is normally taken as the threshold of statistical significance. This may seem surprising: a logarithm of odds of 3.0 implies 1000:1 odds – why use such a stringent threshold rather than the conventional *P*=0.05? The answer lies in the low prior probability that two randomly chosen loci should be linked. To start with, they are far more likely to be on different chromosomes than coincidentally on the same one. And even loci on the same chromosome do not show linkage unless they are fairly close together – there is normally at least one crossover on each chromosome arm during meiosis. There has been debate about the appropriate figure to use for the prior probability of linkage, but a general consensus is 1:50. This low prior probability must be taken into account when deciding the threshold of significance.

The usual way of incorporating the prior probability into a calculation is to use a Bayesian calculation. See *Box 14.7* for an explanation. In the present case the calculation is as follows:

	Loci are linked	Loci are not linked
Prior probability	1/50	49/50
Conditional likelihood (lod score of +3.0)	1000	1
Joint likelihood	20	~ 1

In other words, with a lod score of 3.0, the overall odds of linkage are 20:1, which is the conventional *P*=0.05 threshold. This also shows why it is wrong to just use a chi-squared calculation of the significance.

For more complicated pedigrees a hand calculation becomes impossibly complex. Fortunately a number of computer programs, such as MLINK (Lathrop and Lalouel, 1984), are available that can calculate a lod score for any pedigree on any set of assumptions about allele frequencies or mode of inheritance. In the case of the DSH study, lod scores were calculated for DSH versus each of the 343 microsatellite markers used. Most of the markers gave negative lod scores, but a series of adjacent markers on chromosome 1 gave the lod scores shown in *Table 15.1*.

Table 15.1 – Linkage analysis in the families shown in *Figure 15.4.*

Marker	LOD at Θ =:						
	0	0.05	0.1	0.15	0.2	0.3	0.4
D1S424	–inf.	0.17	1.15	1.51	1.59	1.30	0.66
D1S206	–inf.	–1.21	0.56	1.29	1.59	1.51	0.90
D1S502	–inf.	3.81	4.88	5.02	4.75	3.51	1.68
D1S252	–inf.	5.46	5.96	5.78	5.29	3.77	1.79
D1S498	–inf.	4.49	4.43	4.09	3.62	2.42	1.06
D1S484	–inf.	1.80	2.99	3.34	3.31	2.62	1.42
D1S196	–inf.	–0.76	0.46	0.99	1.21	1.17	0.72
D1S218	–inf.	–7.17	–4.03	–2.40	–1.41	–0.36	0.02

The table shows lod scores calculated by computer for a series of microsatellite markers from the long arm of chromosome 1. Θ symbolizes the recombination fraction. See text for discussion.

The microsatellites are named according to a standard scheme (D = DNA segment, 1 = on chromosome 1; S = single copy; the number just records the order in which each marker was first described).

Adapted from Miyamura *et al.* (2003) with permission from Elsevier.

Things to note about this table are:

- Some of the lod scores are well above 3, showing strong evidence for linkage.
- The fact that the lod score at zero recombination is minus infinity for every marker means that with every marker there are some recombinants between the marker and the disease. In other words, none of the markers is extremely close to the disease locus.

- The markers are shown in chromosomal order. The lod scores are much stronger for markers in the center of the table than those at the top and bottom. This implies that the disease locus is somewhere in the middle part of the region covered by these markers.
- Don't try to read too much into the exact value of a strongly positive lod score. The lod score for one marker may be lower than that for an adjacent marker because one or two individuals happened to be homozygous for one marker and not for the other. Offspring of homozygous individuals contribute no information to the mapping process, so the lod score is lower. But the overall pattern shows strong support for linkage to markers in the central part of the region covered.

The markers shown in *Table 15.1* span a genetic distance of just over 60 cM on the long arm of chromosome 1. Since there are recombinants with each marker, there is scope to narrow down the candidate region by using a fresh set of more closely spaced markers. *Figure 15.6* shows data for *D1S498* and 11 new microsatellites that map in the 13 cM gap between *D1S498* and *D1S484*. The different alleles for each marker are given arbitrary numbers, and the genotypes of seven individuals are shown, each as two haplotypes. The haplotypes were deduced by typing the parents. Rather than calculate lod scores (irrelevant since at this stage they knew that this region was linked to the disease locus), the researchers used the haplotypes to pinpoint individual recombinations.

In a given family, most affected people had inherited the same complete haplotype of markers with the disease gene – that is, in the meiosis in the parent who transmitted the disease gene to that individual, that region of chromosome had not been disrupted by any crossover. Most affected people in Family 1 had the haplotype shown in green in individual 1–I1. Similarly, in Family 2 most had the blue haplotype, and in Family 3 most had the yellow haplotype on their disease-bearing chromosome. However, several individuals had recombinant haplotypes, indicated in *Figure 15.6* by the part-colored

	family 1				family 2						family 3			
INDIVIDUAL	1-I1		1-IV3		2-II1		2-IV4		2-IV5		3-III3		3-IV4	
D1S498	5	2	5	4	2	3	2	2	2	5	3	2	2	4
D1S2347	1	4	1	1	4	1	4	1	4	1	1	3	3	1
D1S2345	4	4	4	8	7	2	7	1	2	1	2	2	2	1
D1S2858	2	1	2	2	1	1	1	2	1	2	2	2	2	2
D1S305	6	6	6	7	6	5	6	1	6	6	6	8	8	2
D1S2715	5	6	5	5	1	1	1	5	1	3	6	3	3	1
D1S2777	1	4	4	3	3	3	3	3	3	3	5	3	5	3
D1S2624	3	3	3	3	5	1	4	3	5	1	4	2	4	4
D1S506	6	6	5	4	5	5	5	4	5	5	5	2	5	5
D1S2635	7	8	8	8	7	8	8	7	7	7	7	7	7	8
D1S2771	3	1	1	1	1	1	1	3	1	1	1	1	1	3
D1S2707	6	2	2	6	4	6	2	6	4	6	5	6	5	2

Figure 15.6 – Haplotypes of seven affected individuals from the three families in Figure 15.4.

See text for discussion. Reproduced from Miyamura et al. (2003) with permission from Elsevier.

boxes. Because these individuals were affected, they must have inherited the disease allele. This therefore had to be located within the colored segment of their recombinant haplotype, or in the immediately adjacent gap. Thus individual IV–3 in Family 1 tells us that the disease locus must lie above marker *D1S2777*. (Although *D1S2715* is the lowermost marker in the colored part of the haplotype, the disease gene might lie between this and the next marker down – but not at or below *D1S2777*). Individual IV–4 in Family 3 tells us it must lie below marker *D1S2715* (in the yellow segment of the chromosome). Therefore it must actually lie between these two markers. The other recombinant individuals are consistent with this localization.

This analysis narrowed down the candidate region to just 500 kb at chromosomal location 1q21.3. Only seven genes were listed in that region in the human reference genome. DNA from affected individuals in each pedigree (and another unrelated individual with the same disease) was then scanned for mutations in each gene. Changes were found in each sample in one of the genes, *DSRAD* (double-stranded RNA-specific adenosine deaminase). Within a family each affected person had the same change, but they were different between families, and not all in the same exon (exons 2, 10, 10 and 15 in the four unrelated cases).

It remained to prove that the changes detected were the true pathogenic mutation. Pointers to this were as follows.

- Two of the four mutations were nonsense mutations, which are unlikely to be common polymorphic variants.
- The other two mutations were both mis-sense mutations involving amino acids that were completely conserved in the corresponding protein in 11 different animal species surveyed. Such conservation implies that these amino acids have important functions, so the mutations are unlikely to be common polymorphisms.
- DNA from 116 unrelated normally pigmented Japanese adults was tested for each mutation; none was found. Again, this argues against their being non-pathogenic polymorphism.

Surprisingly, when the *DSRAD* gene was knocked out in mice, the result was an embryonic lethal, even in heterozygotes. Although unexpected, this does not disprove the identification of *DSRAD* as the dyschromatosis gene. Although the similarities between mice and humans greatly outweigh the differences, there are differences. These are especially likely to be seen where an effect is due to haploinsufficiency (where a 50% level of gene activity is not enough for normal function), as here. Note also that none of the research described shows why a 50% reduction in the activity of this gene should cause pigmentary abnormalities of the skin of just the hands and feet. Often identifying the gene is just the start of research into what it does.

References

Griffiths AJF, Gelbart WH, Lewontin RC and Miller JH (1999) *Modern Genetic Analysis*. WH Freeman.

Lathrop GM and Lalouel JM (1984) Easy calculations of lod scores and genetic risks on small computers. *Am. J. Hum. Genet.* **36**: 460–465.

Miyamura Y, Suzuki T, Kono M, *et al.* (2003) Mutations of the RNA-specific adenosine deaminase (*DSRAD*) gene are involved in dyschromatosis symmetrica hereditaria. *Am. J. Hum. Genet.* **73**: 693–699.

Morton NE (1955) Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**: 277–318.

Ott J (1999) *Analysis of Human Genetic Linkage*, 3rd edn. Johns Hopkins University Press.

Strachan T and Read AP (2019) *Human Molecular Genetics,* 5th edn. CRC Press.