Part IV

Applications of Gene Manipulation and Genomics

CHAPTER 25

Applications of genomics: understanding the basis of polygenic disorders and identifying quantitative trait loci

Introduction

Geneticists use the term "complex trait" to describe any phenotype that does not exhibit classical Mendelian recessive or dominant inheritance attributable to a single gene locus. Most, but not all, complex traits can be explained by polygenic inheritance, i.e. these traits require the simultaneous presence of mutations in multiple genes. Polygenic traits may be classified (Lander & Schork 1994) as discrete traits, measured by a specific outcome (e.g. development of diabetes or cleft palate), or quantitative traits measured by a continuous variable (e.g. grain yield, body weight). In general, discrete traits are of particular interest to human geneticists and quantitative traits are of particular interest to plant and animal breeders, although there are significant exceptions. There is another difference. Human populations are outbred whereas plant and animal breeders use inbred populations and the methods used for gene identification reflect these differences.

Despite the differences cited above, the general methodology used to identify genes associated with discrete traits and quantitative traits is the same (Glazier *et al.* 2002). This methodology involves four steps:

1. Establish significant genome-wide evidence for linkage or association of the trait with a particular chromosomal region. Typically, the trait will be localized to a 10–30 cM region of the genome. In humans this equates to 10–30 Mb of DNA with a coding potential of 100–300 genes.
2. Fine mapping is undertaken to reduce the size of the critical region to one that permits sequencing to be undertaken.
3. DNA sequence analysis is undertaken to identify any candidate nucleotide variants.
4. Attempts are made to demonstrate that replacement of the variant nucleotide(s) results in the swapping of one phenotype for another.

This basic methodology forms a recurring theme in the examples described in the sections that follow.

Investigating discrete traits in outbreeding populations (genetic diseases of humans)

Mendelian traits are relatively easy to study but account for only a small proportion of human disease. Most human diseases are polygenic in nature and these include cardiovascular disease, asthma, cancer, diabetes, rheumatoid arthritis, obesity, alcoholism, and schizophrenia. Such complex diseases involve multiple genes, environmental effects, and their interactions. Rather than being caused by specific and relatively rare mutations, complex diseases and traits may result principally from genetic variation that is relatively common in the population. The fact that a large number of genes, many with small effects, are involved in many complex diseases greatly complicates efforts to identify genetic regions involved in the disease process and makes replication of results difficult. The distinction in terminology between Mendelian and complex traits is not meant to imply that complex diseases do not follow the rules of Mendelian inheritance; rather, it is an indication that the inheritance pattern of complex traits is difficult to discern.

There are three main approaches to mapping the genetic variants involved in a disease: functional cloning, the candidate gene strategy, and positional cloning (see Chapter 6). In functional cloning, knowledge of the underlying protein defect leads to localization of the responsible gene. In the candidate gene approach, genes with known or proposed function with the potential to influence the disease phenotype are investigated for a direct role in disease. Positional gene cloning is used when the biochemical nature of the disease is unknown (the norm!). The responsible gene is mapped to the correct location on the chromosome and successive narrowing of the candidate interval eventually results in the identification of the correct gene.

The gene-finding methods described above are used in conjunction with two other analytical methods. These are model-free (or nonparametric)

linkage analysis and association (or linkage disequilibrium) mapping. Model-free methods make no assumption about the inheritance pattern, the number of loci involved, or the role of environment. Rather, they depend solely on the principle that two affected relatives will have disease-predisposing alleles in common. In linkage disequilibrium (LD) mapping one looks at co-inheritance in populations of unrelated individuals. There is another difference. Linkage analysis can be used only for coarse mapping (e.g. only 10% recombination will be observed in a region of 10 Mb), whereas LD can be used for fine mapping as resolution is limited only by the spacing of the markers used. Consequently, most effort is being devoted to developing physical maps with a high marker density.

The reader who is not familiar with the methods used in the study of human genetics will have great difficulty in understanding the primary literature. The reason for this is the widespread use of specialist terminology. To assist readers a glossary is provided in Box 25.1.

Box 25.1 Glossary of terms used in human genetics

Ascertainment bias

This is the difference in the likelihood that affected relatives of the cases will be reported to the geneticist as compared with the affected relatives of controls.

Concordance

If two related individuals in a family have the same disease they are said to be concordant for the disorder (cf. discordance).

Discordance

If only one member of a pair of relatives is affected with a disorder then the two relatives are said to be discordant for the disease (cf. concordance).

Familial aggregation

Because relatives share a greater proportion of their genes with one another than with unrelated individuals in the population, a primary characteristic of diseases with complex inheritance is that affected individuals tend to cluster in families. However, familial aggregation of a disease does not necessarily mean that a disease has a genetic basis as other factors could be at work.

Founder effect

If one of the founders of a new population happens to carry a relatively rare allele, that allele will have a far higher frequency than it had in the larger group from which the new population was derived. The founder effect is well illustrated by the Amish in Pennsylvania, the Afrikaners in South Africa, and the French-Canadians in Quebec. An early Afrikaner brought the gene for variegate porphyria and the incidence of this gene in South Africa is 1 in 300 compared with 1 in 100,000 elsewhere.

Genome scan

This is a method whereby DNA of affected individuals is systematically analyzed using hundreds of polymorphic markers in a search for regions that are shared by the two sibs (cf.) more frequently than on a purely random basis. When elevated levels of allele sharing are found at a polymorphic marker it suggests that a locus involved in the disease is located close to the marker. However, the more polymorphic the loci studied the more likely it is that elevated allele sharing occurs by chance alone and hence one looks for high LOD scores.

Index case

See "proband".

Multiplex family

A family with two or more affected members.

continued

Box 25.1 *continued*

Nonparametric (model-free) analysis

This method makes no assumption concerning the number of loci or the role of environment and chance in causing lack of penetrance (q.v.). Instead, it depends solely on the assumption that two affected relatives will have disease-predisposing alleles in common.

Parametric (model-based) linkage analysis

This method of analysis assumes that there is a particular mode of inheritance (autosomal dominant, X-linked, etc.) that explains the inheritance pattern. Therefore one looks for evidence of a genetic locus that recombines with a frequency that is less than the 50% expected with unlinked loci.

Penetrance

In clinical experience, some disorders are not expressed at all even though the individuals in question carry the mutant alleles. Penetrance is the probability that such mutant alleles are phenotypically expressed.

Proband

The member through whom a family with a genetic disorder is first brought to attention (ascertained) is the proband or index case if he or she is affected.

Relative risk

The familial aggregation (q.v.) of a disease can be measured by comparing the frequency of the disease in the relatives of an affected individual with its frequency in the general population. The relative risk ratio is designated by the symbol λ. In practice, one measures λ for a particular class of relative, e.g. sibs, parents.

Sibs

Brothers and sisters are sibs.

Simplex family

A family in which just one member has been diagnosed with a particular disease.

Transmission disequilibrium test

This tests whether any particular alleles at a marker are transmitted more often than they are not transmitted from heterozygous parents to affected offspring. The benefit of this test is that it only requires trios (q.v.).

Trio

An affected child plus both parents.

Model-free (nonparametric) linkage analysis looks at the inheritance of disease genes and selected markers in several generations of the same family

Any kind of genetic marker can be used in linkage mapping and in classical genetics these markers are other phenotypic traits. In practice, it is difficult to detect linkages for loci more than 25 cM apart. Thus, to be useful, markers need to be distributed throughout the genome at a frequency of at least one marker every 10 cM. In humans there are not enough phenotypic traits that have been mapped to give anything like the desired marker density. For this reason, physical markers are very attractive. The first such markers to be described were restriction fragment length polymorphisms (RFLPs, see p. 346) but the ones favored today are the single nucleotide polymorphisms (SNPs, see p. 349) because they occur once every 1000 bp. When large multigeneration pedigrees are available (e.g. the Centre d'Etude du Polymorphisme Humain (CEPH) families, p. 362) linkage analysis is a powerful technique for locating disease genes and has been applied to a number of simple Mendelian traits. The probability of linkage is calculated and expressed as a logarithm_{10} of odds (LOD) score, with a value above 3 being significant (Box 25.2). If linkage to a marker is observed then the chromosomal location of that marker is also the location of the disease gene.

Box 25.2 Logarithm of odds scores
(Adapted from Connor & Ferguson-Smith 1997)

Figure B25.1 shows pedigrees for two families affected by an autosomal dominant disorder. In family A the affected man in the second generation has received the disease allele together with RFLP allele 1 from his father. Similarly, he has received the normal allele and RFLP allele 2 from his mother. If these two loci are on the same chromosome then it follows that he must have one chromosome that carries the disease allele together with RFLP allele 1 and the other carries the normal allele and RFLP allele 2. Consequently, the arrangement of the disease and marker alleles, also known as the *phase*, can be deduced with certainty in this individual. If the loci are linked it will be apparent in the next generation as a tendency for the disease allele to segregate with RFLP allele 1 and the normal allele to segregate with RFLP allele 2. This is indeed the case in family A, where four affected offspring carry RFLP allele 1 and the five unaffected children only carry RFLP allele 2.

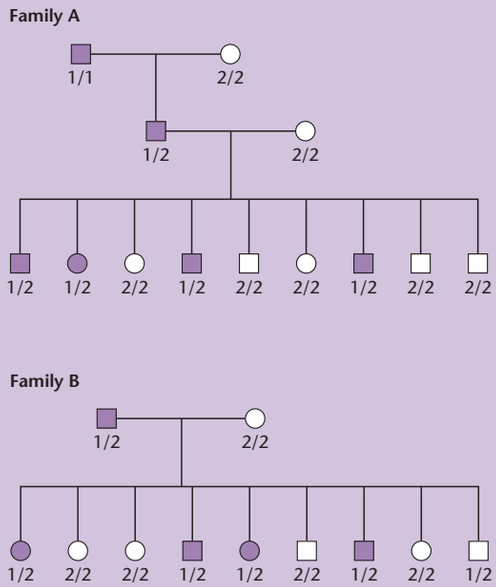


Fig. B25.1 Two families with an autosomal dominant trait showing results of DNA analysis for a marker restriction fragment length polymorphism (RFLP) with alleles 1 and 2.

If the loci described above are not linked, the probability of such a striking departure from independent assortment occurring by chance in nine offspring is the probability of correctly calling heads or tails for nine consecutive tosses of a coin. That is:

$$(0.5)^9 = 0.002.$$

However, if these two loci are linked such that there is only a 10% chance of crossing over (i.e. a recombination fraction, or θ, of 0.1), the probability of the disease segregating with RFLP allele 1 or the normal allele with RFLP allele 2 is:

$$(0.9)^9 = 0.4.$$

It follows that linkage at 10% recombination is 200 times (0.4/0.002) more likely than no linkage. Similarly, if the disease allele and the RFLP allele are identical, then no recombination could occur and the recombination fraction would be zero. For this family, this is 500 times (1/0.002) more likely than no linkage.

The usual way of representing these probability ratios is as logarithms, referred to as logarithm of odds (LOD) or Z scores. For family A at a recombination fraction of 10% the LOD score is $\log_{10} 200 = 2.3$ and at a recombination fraction of zero it is $\log_{10} 500 = 2.5$.

The figure also shows a two-generation pedigree for family B. In this case, all four affected siblings carry RFLP allele 2 and another four healthy siblings do not. As in family A, this signifies a marked disturbance of independent assortment and suggests linkage between the disease and RFLP allele 1. If this is the case, the youngest child must represent a recombinant because he has inherited RFLP allele 1 from his father but not the disease. However, it could be that the youngest child is non-recombinant and all the other children represent crossovers between the two loci.

continued

Box 25.2 *continued*

Although this is much less likely, it cannot be excluded in the absence of phase information from the grandparents. Calculation of the LOD scores for such a family is more complicated because the two possible phases must be taken into account. Whichever phase is considered, at least one recombination event must have taken place and hence the recombination fraction cannot be zero (see Table B25.1).

Analysis of the combined data from the two families shows that the maximum LOD score is 3.3 and this occurs at 10% recombination.

Table B25.1 Logarithm of odds (LOD) scores at values of the recombination fraction from 0 to 40% for the two families shown in Fig. B25.1.

	Recombination fraction (%)				
	0	**10**	**20**	**30**	**40**
Family A	2.7	2.3	1.8	1.3	0.7
Family B	$-\infty$	1.0	0.9	0.6	0.3
Total	$-\infty$	3.3	2.7	1.9	1.0

The ease with which data from phase-known and phase-unknown families can be combined in this way is the reason why the use of LOD scores has become universal for the analysis of linkage data. The maximum value of the LOD score gives a measure of the statistical significance of the result. A value greater than 3 is usually accepted as demonstrating that linkage is present and in most situations it corresponds to the 5% level of significance used in conventional statistical tests. Conversely, if LOD scores below – 2 are obtained, this indicates that linkage has been excluded at the corresponding values of the recombination fraction.

The relationship between the recombination fraction and the actual physical distance between the loci depends on several factors. A recombination fraction of 0.1 (10% recombination) corresponds to a map distance of 10 cM. However, with increasing distance between the loci the recombination fraction falls as a result of the occurrence of double crossovers. In humans, 1 cM is equivalent to 1 Mb of DNA on average.

Conventional linkage analysis seldom works for complex diseases. The involvement of many genes and the strong influence of environmental factors mean that large multigeneration pedigrees are seen only rarely. Consequently, analysis is undertaken of families in which both parents and at least two children (sib pairs) have the disease in question. These are known as *nuclear families*. The way this analysis is undertaken is shown in Fig. 25.1. Suppose that we

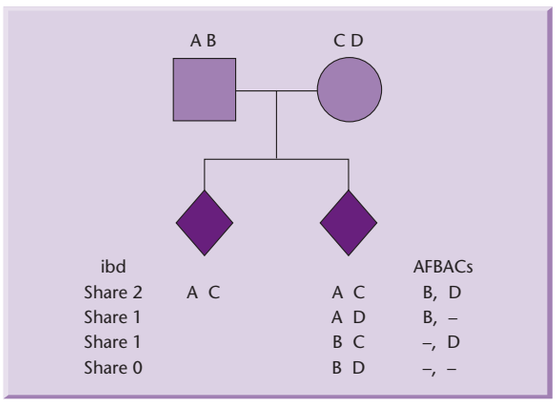


Fig. 25.1 (*left*) Affected sib pair families. A nuclear family pedigree is shown with the father (■) and mother (●) in the first row and the two affected children of either sex (◆) in the second row. Assume for simplicity that we can distinguish all four parental alleles, denoted A, B, C, and D in the genetic region under study, with the parental alleles ordered such that A and C are transmitted from the father and mother, respectively, to the first affected child. Four possible configurations among the two offspring with respect to the alleles inherited from the parents are possible: they can share both parental alleles (AC); they can share an allele from the father (A) but differ in the alleles received from the mother (C and D); they can share an allele from the mother (C) but differ in the alleles received from the father (A and B); or they can share no parental alleles in common. These four configurations are equally likely if there is no influence of the genetic region under consideration on the disease. The parental alleles that are never transmitted to the affected sib pair in each family type are used as a control population in association studies using nuclear family data, the so-called affected family-based control (AFBAC) sample. (Redrawn with permission from Thomson 2001.)

believe that a certain region of the genome is implicated in a disease state and that we can distinguish the four parental chromosomes (A, B, C, D). If the region under test does not carry a gene predisposing to disease then the chances of two affected sibs having two, one, or no parental chromosome regions in common are 25, 50, and 25%, respectively. On the other hand, deviation from this Mendelian random expectation indicates that the affected sibs have chromosome regions that are *identical by descent* (ibd) suggesting the presence of genes predisposing to the disease in question. Physical markers, particularly microsatellites, are ideal for distinguishing the chromosome regions derived from each parent. Not only are microsatellites highly polymorphic, but a sufficient number of them have been placed throughout the genome.

The first complex disease to be analyzed using genome-wide linkage scans was type 1 diabetes (Field *et al.* 1994, Hashimoto *et al.* 1994) and this demonstrated linkage to the major histocompatibility complex (see Box 25.3). Since then, a number of other complex diseases have been mapped including bipolar mood disorder (McInnes *et al.* 1996) and Crohn's disease (Hugot *et al.* 1996, Rioux *et al.* 2000). A similar methodology has been used to identify quantitative trait loci (QTLs) controlling adult height (Hirschhorn *et al.* 2001) and human longevity (Geesaman *et al.* 2003).

Box 25.3 The major histocompatibility complex

Higher animals, including humans, are able to distinguish between "self" and "non-self" and to mount a reaction against a very broad spectrum of foreign antigens. This reaction is mediated by the immune response. Genetic factors play a key role in the generation of the normal immune response and, as a result of mutation, in aberrant immune reactions including immunodeficiency and autoimmune disease. A large number of genes play a role in the development and functioning of the immune system but only those of the major histocompatibility complex (MHC) are considered here.

The MHC is composed of a large cluster of genes located on the short arm of chromosome 6. On the basis of structural and functional differences these genes are divided into three classes and each class is highly complex and polymorphic. Two of the three classes correspond to the genes for human leukocyte antigens (HLA) that are cell surface proteins. These antigens are very important for the normal functioning of the immune system and were first discovered following attempts to transplant tissue between unrelated individuals. A class I antigen consists of two polypeptide units, a polymorphic peptide encoded by the MHC, and an invariant polypeptide encoded by a gene outside the MHC. Class two molecules are heterodimers of α and β subunits, both of which are encoded by the MHC. The class III genes are not HLA genes but include genes for polymorphic serum proteins and membrane receptors.

The HLA system comprises many genes and is highly polymorphic with many antigenic variants having been recognized at the various loci (Table B25.2). Because the HLA alleles are so closely linked they are transmitted together as haplotypes. Each individual has two haplotypes, one on each copy of chromosome 6, and the alleles are co-dominant. Each child receives one

HLA locus	Antigenic variants (no.)	DNA variants (no.)
HLA-A	25	83
HLA-B	53	186
HLA-C	11	42
HLA-DR (β chain only)	20	221
HLA-DQ (α and β chains)	9	49
HLA-DP (α and β chains)	6	88

Table B25.2 Protein and DNA variation at HLA loci. Because of the redundancy of the genetic code it is possible to have more DNA sequence variants than protein variants.

continued

Box 25.3 *continued*

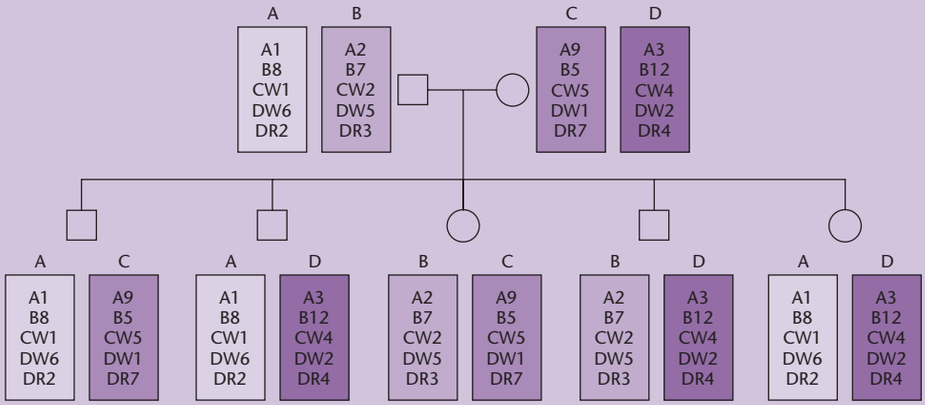


Fig. B25.2 The inheritance of HLA haplotypes. Usually a haplotype is transmitted, as shown in this figure, as a unit. In extrememly rare instances, a parent will transmit a recombinant haplotype to the child.

haplotype from each parent (Fig. B25.2) and there is a 25% chance that two children with the same parents inherit matching HLA haplotypes. Because the success of tissue transplantation is closely linked to the degree of similarity between HLA haplotypes, the favored donor for bone marrow or organ transplantation is a brother or sister who has an identical HLA haplotype.

As more and more information has accumulated about the HLA genes it has become clear that there is an association between specific HLA genes or haplotypes and certain diseases. For example, in one national study, only 9% of the population had the HLA-B27 allele but it was present in 95% of those with the chronic inflammatory disease ankylosing spondylitis. Similarly, 28% of the population carried the HLA-DQ2 allele but it was present in 99% of the population with celiac disease. It is unlikely that HLA genes alone are responsible for specific diseases. Rather, they may contribute to disease predisposition along with other genetic and environmental factors. For example, they probably influence the susceptibility of different individuals to particular infectious agents. They also can play a role in complex diseases as exemplified by type 1 diabetes.

There are two major types of diabetes mellitus: juvenile-onset or insulin dependent (type 1) and adult-onset or insulin-independent (type 2). Type 1 diabetes has a frequency of 0.5% in the Caucasian population and results from an autoimmune destruction of the insulin-producing cells in the pancreas. Genetic factors alone do not cause type 1 diabetes because if one twin of an identical pair develops the disease there is only a 40% chance that the matching twin also will become diabetic. Nevertheless, there is strong evidence for genetic factors and, as noted on p. 490, the first study on model-free analysis of a complex disease linked type 1 diabetes with the MHC locus. Individuals heterozygous for HLA-DR3 or HLA-DR4 are particularly susceptible to diabetes. This fits with the concept of type 1 diabetes being an autoimmune disease, since DR3 and DR4 are found in a locus known to regulate the immune response.

Further insight into the mechanism responsible for type 1 diabetes has come from a molecular analysis of the HLA-DQ genes. The presence of aspartic acid at position 57 of the DQβ chain is closely associated with resistance to type 1 diabetes, whereas other amino acids at this position confer susceptibility. About 95% of patients with type 1 diabetes are homozygous for DQβ genes that do not encode aspartate at position 57. Since position 57 of the β chain is critical for antigen binding and presentation to T cells, changes in this amino acid could play a role in the autoimmune response that destroys the insulin-producing cells.

Linkage disequilibrium (association) studies look at the co-inheritance of markers and the disease at the population level

Whereas linkage analysis is undertaken in families, association studies are undertaken on unrelated cases and controls. If there is a significant association of a marker and a disease state then this may implicate a candidate gene in the etiology of a disease. Alternatively, an association can be caused by LD of marker allele(s) with the gene predisposing to disease. LD implies close physical linkage of the marker and the disease gene. As might be expected, LD is not stable over long time periods because of the effects of meiotic recombination. Thus, the extent of LD decreases in proportion to the number of generations since the LD-generating event. In general, the closer the linkage of two SNPs then the longer the LD will persist in the population but other factors do have an influence, e.g. extent of inbreeding, presence of recombination hotspots, etc. Reich *et al.* (2001) have shown that, in a US population of northern European descent, LD typically extends for about 60 kb. By contrast, LD in a Nigerian population extends for a much shorter distance (5 kb) reflecting the fact that northern Europeans are of more recent evolutionary origin.

So, it should be apparent that the ideal population for LD studies will be one that is isolated, has a narrow population base, and can be sampled not too many generations from the event causing the disease mutation. The Finnish and Costa Rican populations are considered ideal because they are relatively homogenous and show LD over a much wider distance than US populations. This is particularly important because it influences the number of markers that need to be used. In a typical *linkage* analysis one uses markers every 10 Mb (10 cM) but for LD studies one needs many more markers. For a US population of northern European descent the markers would need to be every 20–50 kb on average but this could be extended to every 200–500 kb for Finnish or Costa Rican populations.

Genome-wide LD scans have been undertaken to locate simple Mendelian traits. Lee *et al.* (2001) were able to use such a methodology to localize the critical region for a rare genetic disease (SLSJ cytochrome oxidase deficiency) in a close-knit isolated community. More typically, LD analysis is used to fine-map traits following initial localization to chromosomal regions by linkage analysis as described in the previous section. SNPs are ideal for this purpose because over 10 million of them from different ethnic groups have been mapped (www.ncbi.nlm.nih.gov/SNP/). For studies on complex diseases, evidence for LD is sought in nuclear families or trios because this avoids possible ethnic mismatching between patients and randomly ascertained controls. In such cases, the parental alleles that are never transmitted to the affected offspring are used as the controls, the so-called affected family-based control (AFBAC) sample.

Two studies on Crohn's disease illustrate how LD can be used in fine mapping (Fig. 25.2). In the first study (Rioux *et al.* 2001), linkage analysis had shown that susceptibility to Crohn's disease mapped to an 18-Mb region of chromosome 5 with a maximal LOD score at marker D5S1984. Using 56 microsatellites, LD was detected between Crohn's disease and two other markers, IRF1p1 and D5S1984, which are 250 kb apart. All the known genes in this region were examined for allelic variants that could confer increased susceptibility to Crohn's disease but no candidate genes were identified. This was a little surprising because this genomic region encodes the cytokine gene cluster that includes many plausible candidate genes for inflammatory disease. Because no obvious candidates had emerged, a detailed SNP map was prepared with the markers spaced every 500 bp. Many of these SNPs showed LD with susceptibility to Crohn's disease, confirming the presence of a gene predisposing to Crohn's disease in the area under study.

In the second study (Hugot *et al.* 2001), linkage analysis had mapped a susceptibility locus for Crohn's disease to chromosome 16. With the aid of 26 microsatellites the locus was mapped to a 5-Mb region between markers D16S541 and D16S2623. LD analysis showed a weak association of Crohn's disease with marker D16S3136, which lies between the other two markers. A 260-kb region around marker D16S3136 was sequenced but only one characterized gene was identified and this did not appear to be a likely Crohn's disease candidate. Sequencing also identified 11 SNPs and three of these showed strong LD with Crohn's disease in 235 affected families, indicating that the susceptibility locus was nearby. By using the GRAIL program and an expressed sequence tag (EST) homology search (p. 171) a number of putatively transcribed regions were identified and one of these (*NOD2*, but now known as *CARD15*) was identified as the susceptibility locus. Further analysis showed that some of the SNPs used in the LD study were the causative mutations.

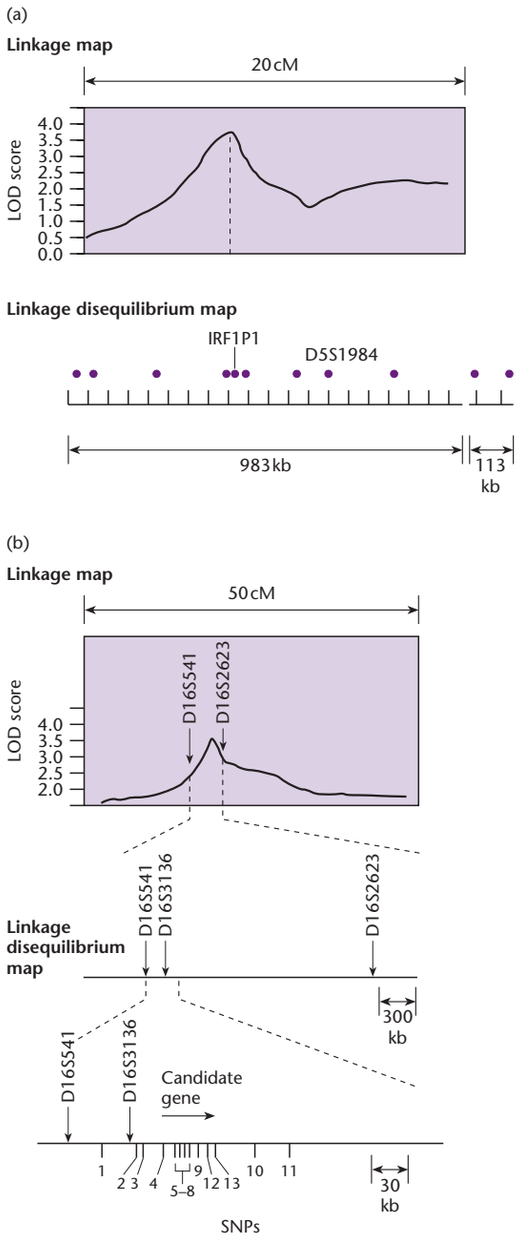


Fig. 25.2 Details of the mapping of two loci associated with Crohn's disease. (a) Mapping of a locus on chromosome 5 by Rioux *et al.* (2001). (b) Mapping of a locus on chromosome 16 by Hugot *et al.* (2001). The numbers along the bottom line correspond to the SNPs used in fine mapping. All the SNPs except 10 and 11 showed tight linkage (see text for further details).

Once a disease locus is identified, all the 'omics can be used to analyze it in detail

In the study of Crohn's disease (CD) described above, Hugot *et al.* (2001) mapped three polymorphisms to the *NOD2/CARD15* locus. These polymorphisms were R702W (an arg/trp replacement at position 702), G908R (a gly/arg replacement at position 908), and a frameshift mutation (1007fs). Together these mutations represent 81% of the polymorphisms seen at the *NOD2/CARD15* locus and a search for other mutations at this locus identified a further 27 variants (Lesage *et al.* 2002). Once the different mutations had been identified population studies became possible and these have shown different allele frequencies in different CD populations throughout the world. For example, the *NOD2/CARD15* mutations are absent in Japanese CD populations but frequent in European populations. Also, Jewish CD populations have a much higher prevalence of a particular allele than the three most common European mutations combined. More important, mutations at the *NOD2/CARD15* locus account for only 25% of cases of CD and determine only ileal disease (Ahmad *et al.* 2002) whereas mutations in the HLA genes (see Box 25.3) determine overall susceptibility to CD.

The *NOD2/CARD15* locus encodes a protein that is a member of a family of intracellular cytosolic proteins that have a role in response to bacterial antigens. Expression studies have shown that it is synthesized in epithelial cells in the small and large intestine. The highest levels of expression are in the specialized epithelial Paneth cells, which are located in the crypts of the small intestine. Although the function of the Paneth cells is unknown they have been shown to secrete antibacerial substances in response to bacterial cell-wall components.

The structure of the *NOD2/CARD15* protein is shown in Fig. 25.3. The C-terminal portion of the molecule consists of a leucine-rich region (LRR) that is involved in bacterial binding. Approximately

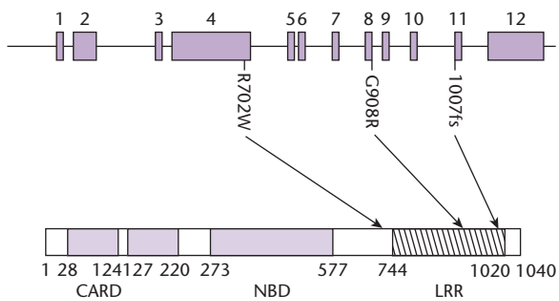


Fig. 25.3 The intron/exon structure of the *NOD2/CARD15* gene and the two-dimensional NOD2/CARD15 protein structure. CARD, caspase activating recruitment domain; NBD, nucleotide binding domain; LRR, leucine-rich region. Reproduced from Russell *et al.*, with permission from Elsevier.

93% of the mutations in *NOD2/CARD15* have been located in the LRR region and they have a diminished ability to activate nuclear factor-κβ (NF-κβ). This suggests that *NOD2/CARD15* acts to protect the intestinal population from bacterial invasion and in CD this protective mechanism malfunctions (Bonen *et al.* 2003).

The murine *NOD2/CARD15* locus has been identified and shown to function in the same way as its human equivalent, e.g. expression is induced by bacterial cell-wall components. A knockout mouse model has been developed but the mice do not develop the intestinal pathology characteristic of CD (Pauleau & Murray 2003). Rather, the knockout mice are more likely to survive bacterial challenge than wildtype mice. This unexpected result shows that the phenotype of complex traits is dependent on the total genetic background of the host in which a mutated allele sits.

The integration of global information about DNA, mRNA, and protein can be used to facilitate disease-gene identification

In the study on Crohn's disease cited in the previous section, the 'omics techniques were used to better understand the biochemical basis of the disease after the gene had been identified. However, in a completely different approach, Mootha *et al.* (2003) used the 'omics to identify the gene that is defective in patients with a particular disease (French-Canadian type Leigh syndrome, LSFC). At the outset it was known that the disease is caused by a deficiency in cytochrome oxidase even though patients do not have mutations in genes for the structural subunits or assembly factors. Genome-wide association studies had shown that the gene maps to chromosome 2p16-21. Using the tools of bioinformatics, 30 genes were identified in the candidate region and there was strong experimental evidence for 15 of them (Fig. 25.4). However, no connection was known between any of these 15 genes and mitochondrial biology.

Functionally related genes tend to be transcriptionally coregulated and this certainly is true in yeast for those genes involved in oxidative phosphorylation. Therefore, Mootha *et al.* (2003) decided to systematically identify genes that exhibited expression patterns resembling those of known mitochondrial genes. The rationale was that any gene that is coregulated with mitochondrial genes might encode polypeptides targeted to this organelle. An examination of the data in publicly available microarray databases showed that just one of the 15 genes (*LRPPRC*) had an expression pattern similar to other mitochondrial proteins.

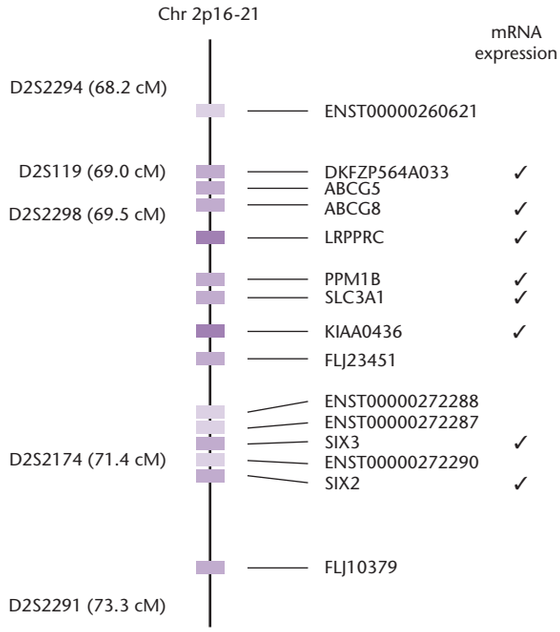


Fig. 25.4 Physical map of the LSFC candidate region. Microsatellite markers and genetic distances are shown to the left of the chromosome map. Genes with varying levels of annotation support are shown with different colors (RefSeq gene, blue; Ensembl gene, green; human mRNA, orange). An additional 15 computationally predicted genes lie within this region but are not shown. Genes represented in mRNA expression sets are indicated with a check to the right of the gene names. Reproduced from Mootha *et al.* (2003). Copyright (2003) National Academy of Sciences, USA.

If the *LRPPRC* gene product is a mitochondrial protein then it should be found in intact mitochondria. Since nothing was known about the gene product, Mootha *et al.* (2003) adopted a novel proteomics approach. They purified human mitochondria, extracted the proteins from them, digested them with trypsin, and analyzed them by tandem mass spectrometry (MS/MS). All the peptides that were identified then were checked to see if they could have been encoded by any of the 15 genes in the candidate region. A total of 12 peptides matched with sequences in the *LRPPRC* gene and no matches were found with any other genes in the LSFC candidate region.

The above result strongly suggested that defects in the *LRPPRC* gene are responsible for LSFC. To confirm that this is indeed the case, all the *LRPPRC* exons were sequenced in DNA from LSFC patients and normal controls. With one exception, DNA from all the patients was found to have a single base change

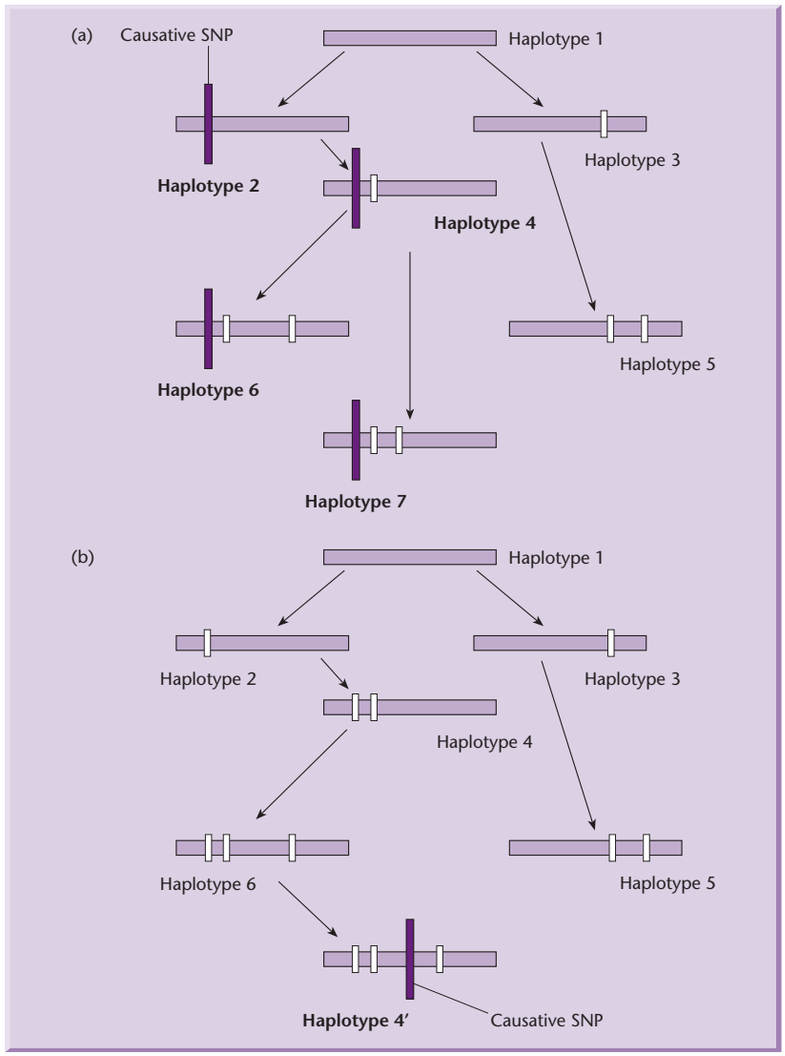


Fig. 25.5 The two ways in which a causative single nucleotide polymorphism (SNP) can become associated with a particular haplotype. In (a) the causative SNP arises early, whereas in (b) it arises late. (Redrawn with permission from Judson *et al.* 2000.)

in exon 9 that would cause a missense change in an amino acid that is conserved in humans, mice, rats, and *Fugu*. One patient was found to be heterozygous for the exon 9 mutation but also was heterozygous for a deletion in exon 35. These results provide definitive genetic proof that *LRPPRC* is the gene that is defective in LSFC and clearly show how the 'omics can be used in gene identification.

The existence of haplotype blocks should simplify linkage disequilibrium analysis

The pattern of SNPs in a stretch of DNA is known as the haplotype. Figure 25.5 shows the evolution of a number of theoretical haplotypes, some of which include an SNP causing disease. From this figure, it is clear that not all SNPs would be predictive of the disease. Also, not all haplotypes are informative and their detection in an LD association study would complicate interpretation of the data. This is exactly the situation encountered in the two studies on Crohn's disease described above. To understand haplotype structure better, Daly *et al.* (2001) undertook a detailed analysis of 103 SNPs within the 500-kb region on chromosome 5q31 associated with Crohn's disease. Their results showed a picture of discrete haplotype blocks of tens to hundreds of kilobases, each with limited diversity punctuated by apparent sites of recombination (Fig. 25.6). In a corresponding study, Johnson *et al.* (2001) genotyped 122 SNPs in nine genes from 384 individuals and found a limited number of haplotypes (Fig. 25.7).

The existence of haplotype blocks should greatly simplify LD analysis. Rather than using all the SNPs

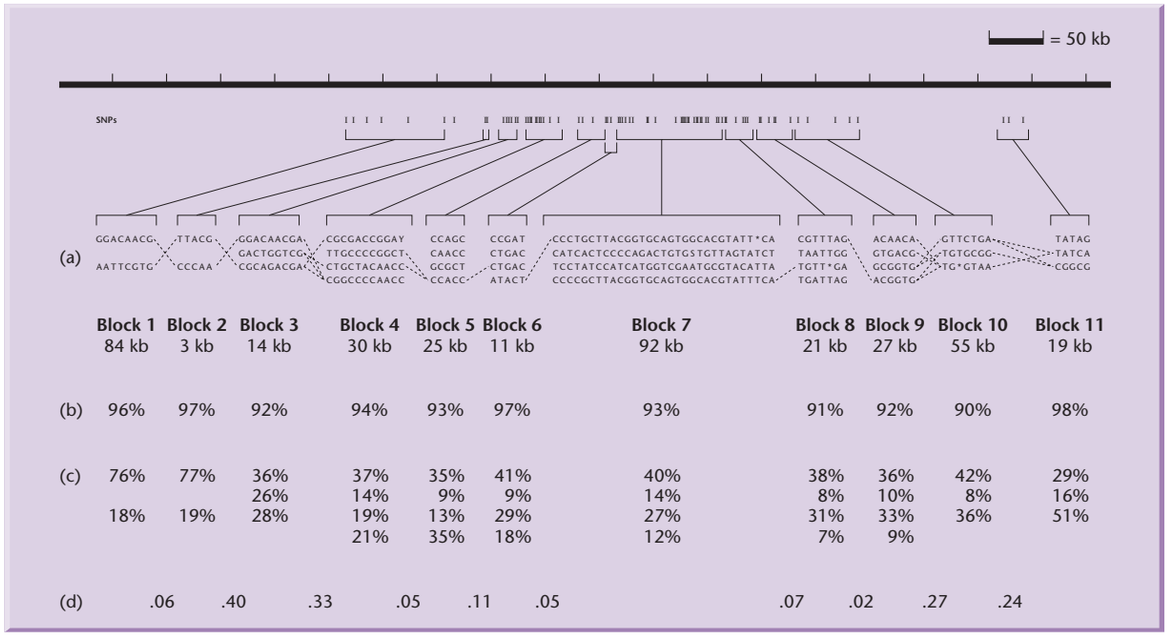
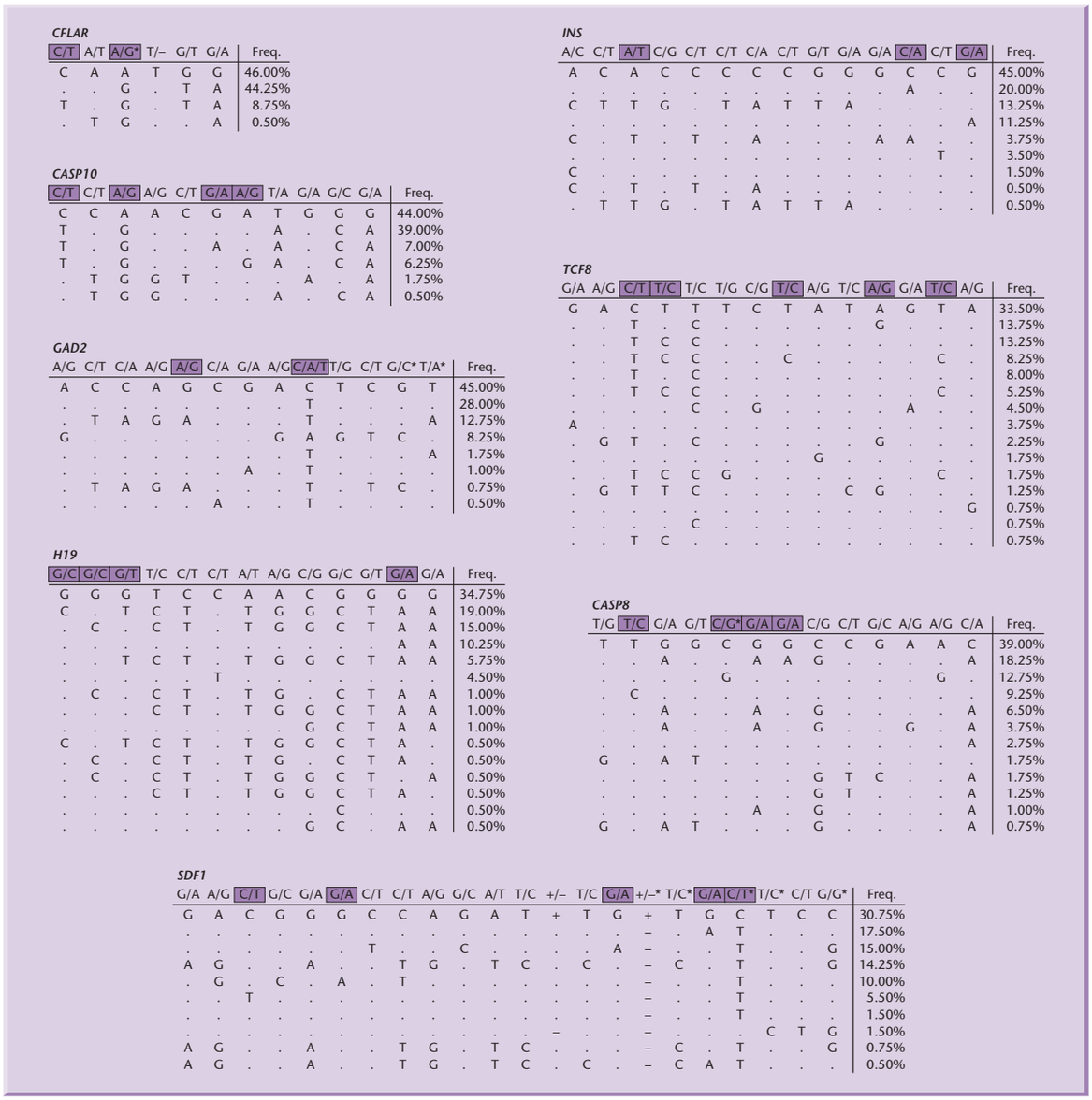


Fig. 25.6 Block-like haplotype diversity at 5q31. (a) Common haplotype patterns in each block of low diversity. Dashed lines indicate locations where more than 2% of all chromosomes are observed to transition from one common haplotype to a different one. (b) Percentage of observed chromosomes that match one of the common patterns exactly. (c) Percentage of each of the common patterns among untransmitted chromosomes. (d) Estimated rate of haplotype exchange between the block. (Reprinted from Daly *et al.* 2001 by permission of Nature Publishing Group, New York.)

in a region, we can identify exactly which SNPs will be redundant and which will be informative in association studies. The latter are referred to as haplotype tag SNPs (htSNPs) and they are markers that capture the haplotype of a genomic region. Thus, once the haplotype blocks in any given region are identified they can be treated as alleles and tested for LD. This not only simplifies the analysis but it reduces the number of SNPs that need to be genotyped. With this in mind, the US National Institutes of Health has funded a haplotype mapping ("HapMap") project whose aim is to catalog the common haplotype blocks in multiple human populations (Couzin 2002).

Although the concept of the HapMap project is very simple, reducing it to practice may be more complex (Cardon & Abecasis 2003). In the first detailed genome-wide analysis of haplotype blocks, Gabriel *et al.* (2002) focused on SNPs with minor allele frequencies >10% in 51 genomic regions. Their samples included individuals from a number of distinct ethnic groups. Two significant observations were made. First, haplotype blocks can be detected with relatively few markers and that within each block, three to five haplotypes can account for 90% of all chromosomes in the population. Second, the haplotype blocks are shorter in populations of African ancestry (average 11 kb) than in the other ethnic groups (average 22 kb). Based on these results, Gabriel *et al.* (2002) proposed that through careful SNP selection all the common haplotypes in the genome could be identified with no more than 300,000–1,000,000 SNPs.

The above proposal is dependent on block boundaries and haplotype diversity remaining relatively stable as more markers are examined. However, a study of chromosome 19 indicated that only one-third of the chromosome exhibited a block-like structure (Phillips *et al.* 2003). Also, most investigations of haplotype blocks have used a small set of SNPs that are not representative of the variants in the population. Rather, common alleles are over-represented and rare alleles are under-represented. This frequency bias is unlikely to be a problem if common diseases are caused by common variants (Pritchard & Cox 2002) but it will limit the utility of the method for diseases caused by rarer alleles. Finally, the size of the haplotype blocks is influenced by marker density: denser marker panels have identified more short blocks whereas sparser marker panels have identified a smaller number of blocks of greater length. Despite these limitaions, the existence of haplotypes undoubtedly will facilitate the identification of disease genes.



CFLAR

C/T	A/T	A/G*	T/–	G/T	G/A	Freq.
C	A	A	T	G	G	46.00%
.	.	G	.	T	A	44.25%
T	.	G	.	T	A	8.75%
.	T	G	.	.	A	0.50%

CASP10

C/T	C/T	A/G	A/G	C/T	G/A	A/G	T/A	G/A	G/C	G/A	Freq.
C	C	A	A	C	G	A	T	G	G	G	44.00%
T	.	G	.	.	.	.	A	.	C	A	39.00%
T	.	G	.	.	A	.	A	.	C	A	7.00%
T	.	G	.	.	.	G	A	.	C	A	6.25%
.	T	G	G	T	.	.	.	A	.	A	1.75%
.	T	G	G	.	.	.	A	.	C	A	0.50%

GAD2

A/G	C/T	C/A	A/G	A/G	C/A	G/A	A/G	C/A/T	T/G	C/T	G/C*	T/A*	Freq.
A	C	C	A	G	C	G	A	C	T	C	G	T	45.00%
.	.	.	.	.	.	.	.	T	.	.	.	.	28.00%
.	T	A	G	A	.	.	.	T	.	.	.	A	12.75%
G	.	.	.	.	.	.	G	A	G	T	C	.	8.25%
.	.	.	.	.	.	.	.	T	.	.	.	A	1.75%
.	.	.	.	.	.	A	.	T	.	.	.	.	1.00%
.	T	A	G	A	.	.	.	T	.	T	C	.	0.75%
.	.	.	.	.	A	.	.	T	.	.	.	.	0.50%

H19

G/C	G/C	G/T	T/C	C/T	C/T	A/T	A/G	C/G	G/C	G/T	G/A	G/A	Freq.
G	G	G	T	C	C	A	A	C	G	G	G	G	34.75%
C	.	T	C	T	.	T	G	G	C	T	A	A	19.00%
.	C	.	C	T	.	T	G	G	C	T	A	A	15.00%
.	.	.	.	.	.	.	.	.	.	.	A	A	10.25%
.	.	T	C	T	.	T	G	G	C	T	A	A	5.75%
.	.	.	.	.	T	.	.	.	.	.	.	.	4.50%
.	C	.	C	T	.	T	G	.	C	T	A	A	1.00%
.	.	.	C	T	.	T	G	G	C	T	A	A	1.00%
.	.	.	.	.	.	.	.	G	C	T	A	A	1.00%
C	.	T	C	T	.	T	G	G	C	T	A	.	0.50%
.	C	.	C	T	.	T	G	.	C	T	A	.	0.50%
.	C	.	C	T	.	T	G	G	C	T	.	A	0.50%
.	.	.	C	T	.	T	G	G	C	T	A	.	0.50%
.	.	.	.	.	.	.	.	.	C	.	.	.	0.50%
.	.	.	.	.	.	.	.	G	C	.	A	A	0.50%

INS

A/C	C/T	A/T	C/G	C/T	C/T	C/A	C/T	G/T	G/A	G/A	C/A	C/T	G/A	Freq.
A	C	A	C	C	C	C	C	G	G	G	C	C	G	45.00%
.	.	.	.	.	.	.	.	.	.	.	A	.	.	20.00%
C	T	T	G	.	T	A	T	T	A	.	.	.	.	13.25%
.	.	.	.	.	.	.	.	.	.	.	.	.	A	11.25%
C	.	T	.	T	.	A	.	.	.	A	A	.	.	3.75%
.	.	.	.	.	.	.	.	.	.	.	.	T	.	3.50%
C	.	.	.	.	.	.	.	.	.	.	.	.	.	1.50%
C	.	T	.	T	.	A	.	.	.	.	.	.	.	0.50%
.	T	T	G	.	T	A	T	T	A	.	.	.	.	0.50%

TCF8

G/A	A/G	C/T	T/C	T/C	T/G	C/G	T/C	A/G	T/C	A/G	G/A	T/C	A/G	Freq.
G	A	C	T	T	T	C	T	A	T	A	G	T	A	33.50%
.	.	T	.	C	.	.	.	.	.	G	.	.	.	13.75%
.	.	T	C	C	.	.	.	.	.	.	.	.	.	13.25%
.	.	T	C	C	.	.	C	.	.	.	.	C	.	8.25%
.	.	T	.	C	.	.	.	.	.	.	.	.	.	8.00%
.	.	T	C	C	.	.	.	.	.	.	.	C	.	5.25%
.	.	.	.	C	.	G	.	.	.	.	A	.	.	4.50%
A	.	.	.	.	.	.	.	.	.	.	.	.	.	3.75%
.	G	T	.	C	.	.	.	.	.	G	.	.	.	2.25%
.	.	.	.	.	.	.	.	G	.	.	.	.	.	1.75%
.	.	T	C	C	G	.	.	.	.	.	.	C	.	1.75%
.	G	T	T	C	.	.	.	.	C	G	.	.	.	1.25%
.	.	.	.	.	.	.	.	.	.	.	.	.	G	0.75%
.	.	.	.	C	.	.	.	.	.	.	.	.	.	0.75%
.	.	T	C	.	.	.	.	.	.	.	.	.	.	0.75%

CASP8

T/G	T/C	G/A	G/T	C/G*	G/A	G/A	C/G	C/T	G/C	A/G	A/G	C/A	Freq.
T	T	G	G	C	G	G	C	C	G	A	A	C	39.00%
.	.	A	.	.	A	A	G	.	.	.	.	A	18.25%
.	.	.	.	G	.	.	.	.	.	.	G	.	12.75%
.	C	.	.	.	.	.	.	.	.	.	.	.	9.25%
.	.	A	.	.	A	.	G	.	.	.	.	A	6.50%
.	.	A	.	.	A	.	G	.	.	G	.	A	3.75%
.	.	.	.	.	.	.	.	.	.	.	.	A	2.75%
G	.	A	T	.	.	.	.	.	.	.	.	.	1.75%
.	.	.	.	.	.	.	G	T	C	.	.	A	1.75%
.	.	.	.	.	.	.	G	T	.	.	.	A	1.25%
.	.	.	.	.	A	.	G	.	.	.	.	A	1.00%
G	.	A	T	.	.	.	G	.	.	.	.	A	0.75%

SDF1

G/A	A/G	C/T	G/C	G/A	G/A	C/T	C/T	A/G	G/C	A/T	T/C	+/–	T/C	G/A	+/–*	T/C*	G/A	C/T*	T/C*	C/T	G/G*	Freq.
G	A	C	G	G	G	C	C	A	G	A	T	+	T	G	+	T	G	C	T	C	C	30.75%
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	–	.	A	T	.	.	.	17.50%
.	.	.	.	.	.	T	.	.	C	.	.	.	.	A	–	.	.	T	.	.	G	15.00%
A	G	.	.	A	.	.	T	G	.	T	C	.	C	.	–	C	.	T	.	.	G	14.25%
.	G	.	C	.	A	.	T	.	.	.	.	.	.	.	–	.	.	T	.	.	.	10.00%
.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	–	.	.	T	.	.	.	5.50%
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	–	.	.	T	.	.	.	1.50%
.	.	.	.	.	.	.	.	.	.	.	.	–	.	.	–	.	.	.	C	T	G	1.50%
A	G	.	.	A	.	.	T	G	.	T	C	.	.	.	–	C	.	T	.	.	G	0.75%
A	G	.	.	A	.	.	T	G	.	T	C	.	C	.	–	C	A	T	.	.	.	0.50%

Fig. 25.7 Common European haplotypes and their haplotype tag single nucleotide polymorphisms (htSNPs) observed at nine genes. Boxed SNPs represent the htSNPs that can capture the common haplotypes that are segregating in European populations. Dots represent the allele that is found on the most common haplotype. Asterisks indicate SNPs described in dbSNP. (Reprinted from Johnson *et al*. 2001 by permission of Nature Publishing Group, New York.)

Investigating quantitative trait loci (QTLs) in inbred populations

Particular kinds of genetic cross are necessary if QTLs are to be mapped

The basis of all QTL detection, regardless of the species to which it is applied, is the identification of association between genetically determined phenotypes and specific genetic markers. However, there is a special problem in mapping QTLs and other complex trait genes and that is penetrance; i.e. the degree to which the transmission of a gene results in the expression of a trait. For a single gene trait, biological or environmental limitation accounts for penetrance, but in a multigenic trait the genetic context is important. Hence, the consequences of inheriting one gene rely heavily on the co-inheritance of others. For this reason careful thought needs to be given to the analysis methodology.

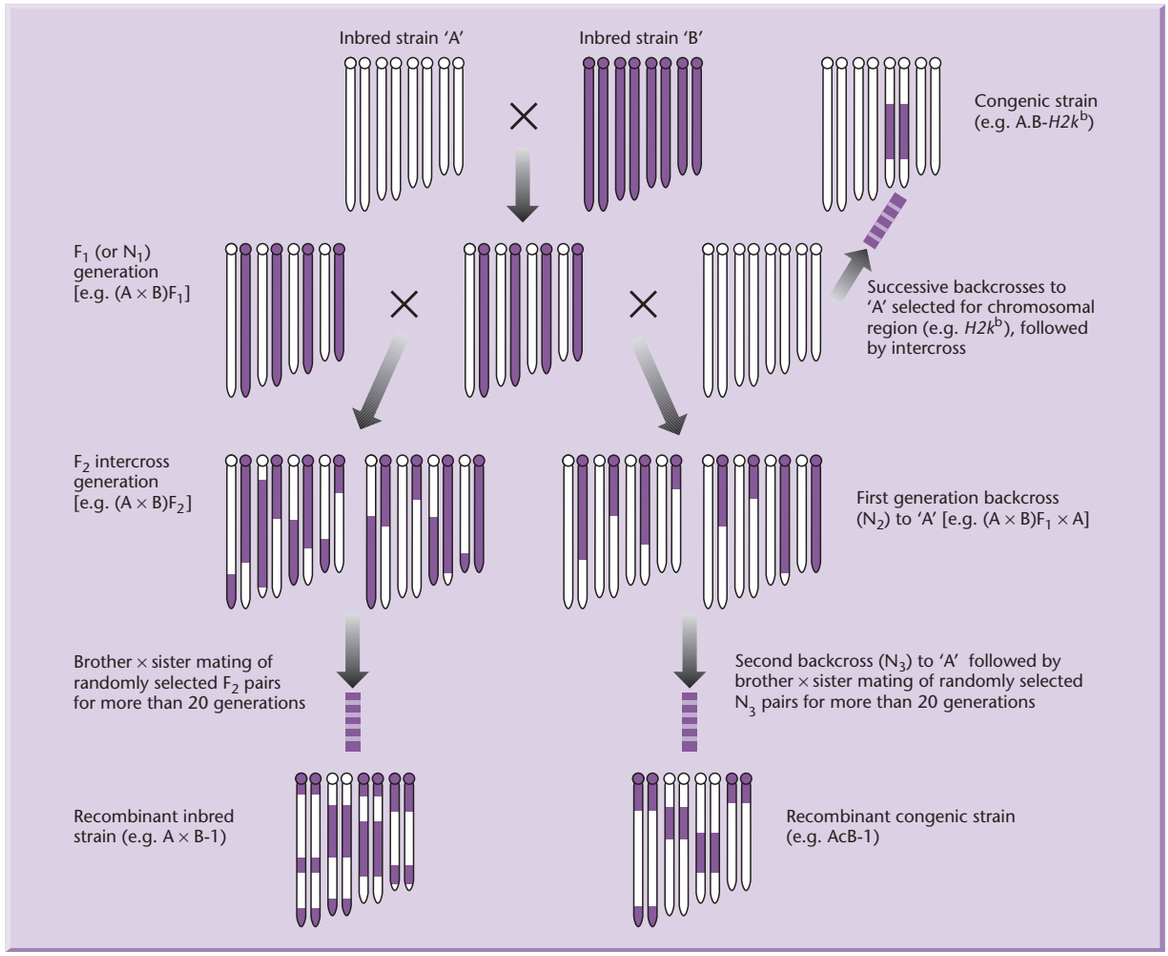


Fig. 25.8 Crosses used in the analysis of complex traits. (Adapted from Frankel 1995 by permission of Elsevier Science.)

In inbreeding populations, i.e. crops and laboratory and domestic animals, the kinds of cross used to dissect QTLs are shown in Fig. 25.8. Populations are typically generated by crosses between inbred strains, usually the first generation backcross (N_2) or intercross (F_2). Higher generation crosses (N_3), panels of recombinant inbred lines (RIL), recombinant congenic strains (RC), or inbred strains themselves may also be used. The chromosomal content in these panels is the heart of the study. They define which alleles are inherited in individuals, so that chromosomal associations can be made, and provide genetic recombination information so that location within a chromosome can be deduced. F_1 hybrids are genetically identical to each other but individuals in subsequent generations are not. Backcross progeny reveal recombination events on only one homolog, the one inherited from the F_1 parent, but intercross progeny reveal such events on both homologs. RILs and RC strains also harbor recombinations but, unlike backcross and intercross progeny, these are homozygous at all loci as a result of inbreeding. A congenic strain of animal or near isogenic line (NIL) of a plant has only one chromosomal region that distinguishes it from one of its parents. Because they have an unchanging genotype, RILs and congenic strains/NILs offer an elegant way of discriminating between the role of the environment and of genetic factors in the expression of phenotype.

Identifying QTLs involves two challenging steps

The traditional approach for QTL analysis occurs in two stages and both are very resource intensive. The first step consists of a large cross between at least two strains in which hundreds or thousands of progeny are assayed for relevant phenotypes (Fig. 25.9) and

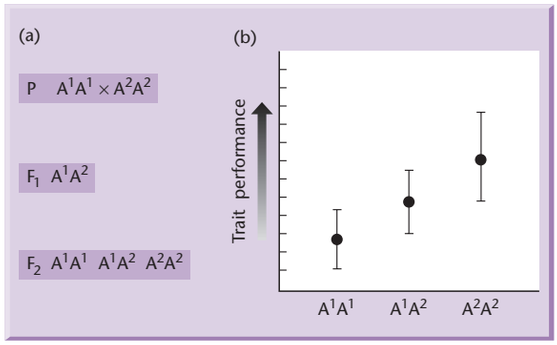


Fig. 25.9 Relationship between the performance for a quantitative trait and genotypes at a marker locus for an F_2 interbreeding population. (a) Genetic composition of the F_2 population sampled. (b) Quantitative performance of different types of F_2 genotypes.

genotyped for polymorphic markers spanning the genome. Because such crosses involve the simultaneous segregation of multiple QTLs only those of large effect are likely to be detected. The second step involves molecular identification of the genetic variants that are responsible for each QTL. This step typically involves studying individual QTLs in isolation by performing 5–10 generations of backcrosses. The objective is to construct congenic strains/NILs with chromosomal segments carrying alternative alleles of the QTL on an otherwise isogenic background. The congenic strains/NILs then are interbred to facilitate fine-structure mapping. A classical example of this approach is provided by studies on sugar content and fruit shape of tomatoes.

One of the major objectives of tomato breeding is to increase the content of total soluble solids (TSS or brix; mainly sugars and amino acids) in fruits to improve taste and processing qualities. TSS in fruits of wild tomatoes (*Lycopersicon pennelli*) can reach up to 15% of the fruit's weight, three times higher than is seen in cultivated tomatoes (*L. esculentum*). To resolve the genetic basis for this variation, a set of 50 NILs was developed from a cross between *L. pennelli* and *L. esculentum*. Each of the NILs contained a single RFLP-defined *L. pennelli* chromosome segment and collectively they covered the genome. Using these NILs it was possible to identify 23 QTLs that increase brix (i.e. polygenic inheritance controls brix). One of these QTLs (Brix9-2-5) was located on a particular NIL that was defined by a 9 cM segment on chromosome 9. When plants of *L. esculentum*, the NIL homozygous for the *L. pennelli* allele of Brix9-2-5 and their F_1 hybrids were compared over a three-year period, the only trait associated with the NIL was brix (Fridman *et al.* 2000).

To map the Brix9-2-5 QTL, 7000 F_2 progeny of the NIL hybrid (Fig. 25.10) were subjected to RFLP analysis with two markers (CP44 and TG225) and this identified 145 recombinants. Of these recombinants, 28 were localized to the region between the two ends of BAC91A4 and they could be subdivided into six recombination groups. When the brix content of the recombination groups was tested the Brix9-2-5 QTL was found to be associated with an 18 kb segment of the chromosome. This 18-kb region was sequenced and various primer pairs used to amplify different regions. One of these amplicons, about 1 kb in length, was found to cosegregate with the QTL. Further mapping reduced the QTL to a 484 bp region of chromosome 9 and sequence analysis identified this as part of an invertase expressed exclusively in flowers and fruits of tomatoes. Initially the biochemical changes resulting from the Brix9-2-5 allele were not clear but an extension of the QTL analysis to five different tomato species localized the polymorphism to an amino acid change near the catalytic site (Fridman *et al.* 2004).

Two key morphological changes that accompanied tomato domestication were fruit size and shape. Wild tomatoes (*L. pennelli*) have small (~2 g), round berries whereas commercial cultivars have fruit weighing 50–1000 g that comes in a variety of shapes (round, oval, pear-shaped, etc). Genetic crosses between wild and cultivated tomatoes, like those described above, have shown that most of the variation in size and shape is due to fewer than 30 QTLs. One of the key QTLs, fw2.2, accounts for about 30% of the variance in fruit weight and was the first plant QTL to be mapped and cloned (Alpert & Tanksley 1996). The fw2.2 polymorphism has been localized to a gene encoding a 22 kD protein (ORFX) with similarities to a human oncogene (Frary *et al.* 2000). Comparative sequencing of fw2.2 alleles suggested that the modulation of fruit size was attributable to 5′ regulatory variation among the alleles rather than to differences in the structural protein. Further support for this idea was provided by an analysis of transcription in NILs (Cong *et al.* 2002). Large- and small-fruited tomatoes differed in peak transcript levels by one week and this was associated with changes in mitotic activity in the early stages of fruit development. Finally, an artificial gene-dosage series was constructed by generating transgenic plants with 0, 1, 2, 3, or 4 copies of the small-fruited allele of fw2.2 (Liu *et al.* 2003). Analysis of a variety of

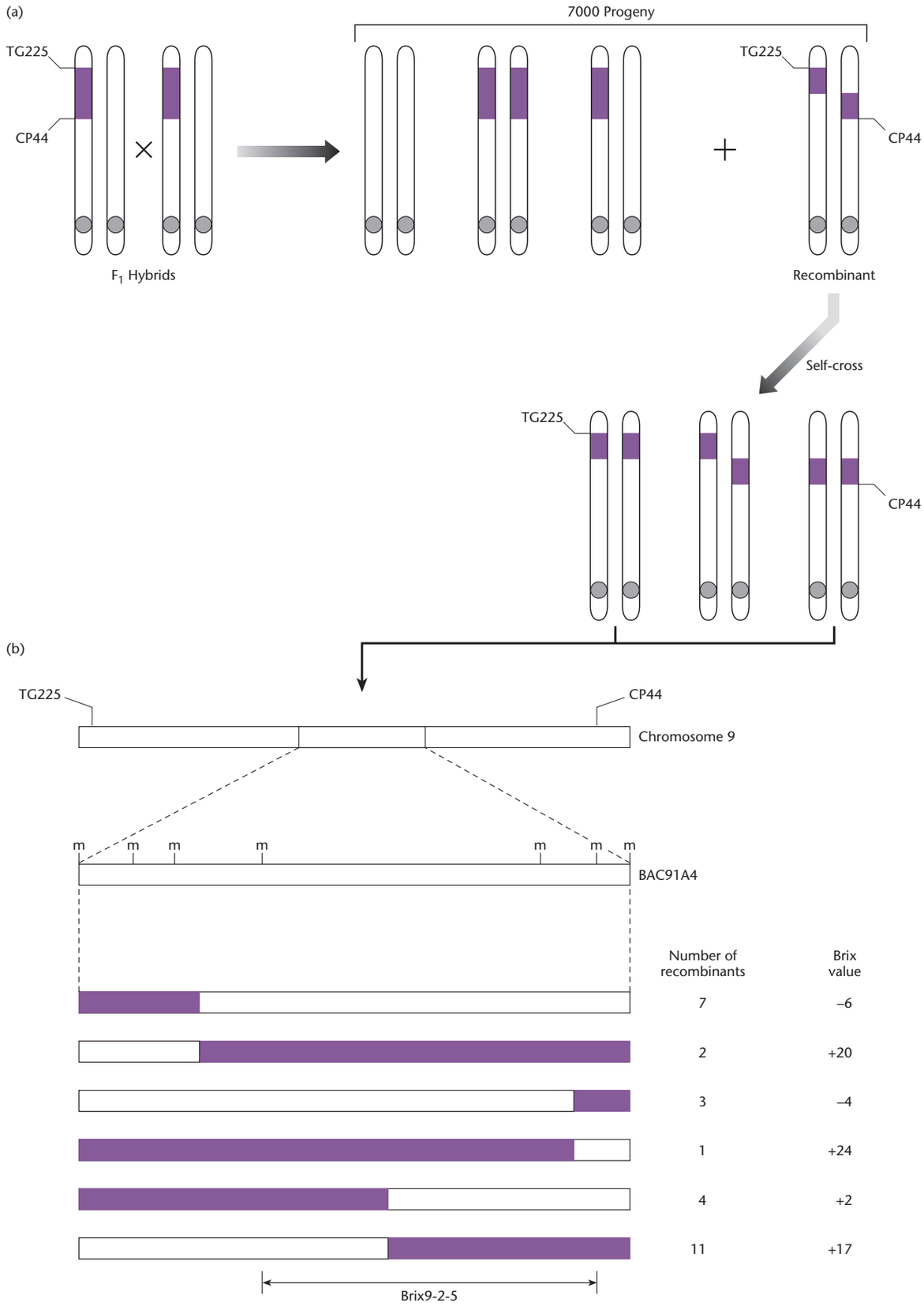


Fig. 25.10 Mapping of the Brix9-2-5 locus. (a) F_1 hybrids are self-crossed and the 7000 progeny tested with RFLP markers CP44 and TG225 to detect those in which recombination has occurred in the region derived from *L. pennelli* (shown in blue). The recombinants are crossed and selection made for homozygous recombinants. (b) The 28 recombinants where crossing over occurred within the region present in BAC91A4 were subjected to fine-scale mapping to identify the crossover more precisely and were tested for brix content.

Table 25.1 Cloned QTLs in plants (adapted from Paran & Zamir 2003).

Plant	QTL	Phenotype	Gene	Variation
Arabidopsis	*EDI*	Flowering time	Cryptochrome photoreceptor	Altered protein function
Arabidopsis	*PHYA*	Hypocotyl elongation	Phytochrome A	Altered protein function
Rice	*Hd1*	Flowering time	Transcription factor	Loss of function
Rice	*Hd6*	Flowering time	Protein kinase	Loss of function
Rice	*Hd3a*	Flowering time	FLOWERING LOCUS T	Unknown
Maize	*tb1*	Plant architecture	Transcription factor	Expression level
Maize	*Dwarf8*	Flowering time	Transcription factor	Unknown
Tomato	*Brix9-2-5*	Sugar content	Invertase	Altered protein function
Tomato	*fw2.2*	Fruit weight	Regulatory gene	Expression level
Tomato	*Ovate*	Fruit shape	Regulatory gene	Loss of function

agronomic factors showed a clear correlation between fruit size and gene dosage and that fw2.2 controls fruit growth in a tissue-specific manner.

Various factors influence the ability to isolate QTLs

At the time of going to press only a very limited number of QTLs had been mapped and sequenced, and Table 25.1 shows the list for plants. Examples from the animal world are given by Glazier *et al.* (2002). In the tomato examples described above, isolation was greatly facilitated by the fact that the selected QTL had a major effect. If the traits being studied were due to a number of combinations of several genes, each of much smaller effect, then analysis would have been much more difficult. Another factor that governs the success of searches for QTLs is the availability of physical markers. Fine-scale mapping was a key part of the identification of the tomato QTLs and such mapping would not be possible without markers such as SNPs, RFLPs, etc.

Fortunately, genome mapping and sequencing projects are in progress for a significant number of crop and domestic animal species (Box 25.4). Finally, from the details of the isolation of the Brix9-2-5 QTL presented above it should be clear that finding QTLs puts a heavy demand on resources. One way of minimizing this problem is to use chromosome substitution strains (Singer *et al.* 2004).

Chromosome substitution strains make the identification of QTLs easier

Chromosome substitution strains (CSSs) are produced as panels from crosses between a donor strain (A) and a host strain (B) such that strain CSS-*i* carries both copies of chromosome *i* from the donor strain but all other chromosomes from the host. Thus, for an organism with four chromosomes the panel would consist of: 1A1A2B2B3B3B4B4B, 1B1B2A2A3B3B4B4B, 1B1B2B2B3A3A4B4B, 1B1B2B2B3B3B4A4A. There is a significant difference between CSSs and congenic strains/NILs. The former have an entire chromosome substitution whereas the latter have only a portion of the chromosome substituted.

Construction of CSS panels is conceptually simple and occurs in two steps. In the first, AB F_1 hybrids are repeatedly backcrossed to the host strain (B) and at each cycle selection is made for progeny carrying a non-recombinant copy of the desired chromosome. Once strains are identified that are heterosomic (A/B) for the desired chromosome on an otherwise isogenic host (B/B) background they are intercrossed to produce progeny homosomic for A/A. In practice, constructing the panel is dependent on detailed physical maps and in the case of the mouse took seven years and involved analysis of 17,000 mice (Singer *et al.* 2004). However, once constructed, a CSS panel is a valuable resource for studying the genetic control of phenotypic variation.

The CSS mouse panel of Singer *et al.* (2004) consists of 22 strains, one for each of the 19 autosomes, the two sex chromosomes, and the mitochondria. This panel of strains was compared with the host and donor strains in terms of sterol levels, amino acid levels, diet-induced obesity, and anxiety. A difference in any of these parameters between the host and donor strains that is seen in one or more of the CSS strains is probably due to a QTL. Overall, evidence was found for 150 QTLs. More convincing proof of the existence of particular QTLs would be their

Box 25.4 Internet resources for domestic animal and crop plant genome projects

The following websites provide current information on the progress of mapping and sequencing the genomes of a variety of domesticated species. Many of the sites also provide links to further resources, including gene mutation and sequence databases and functional analysis tools.

Domestic animals

A useful overview is the document "Coordination of Programs on Domestic Animal Genomics: the Federal Framework" which can be found at www.csrees.usda.gov/nea/animals/pdfs/nstc_progress_rpt.pdf

Dog

http://mendel.berkeley.edu/dog.html Dog Genome Project (University of California, Berkeley, University of Oregon, Fred Hutchinson Cancer Research Center)

Cow

http://www.marc.usda.gov/genome/cattle/cattle.html Meat Animal Research Center

Pig

http://www.marc.usda.gov/genome/swine/swine.html Meat Animal Research Center
http://www.projects.roslin.ac.uk/pigmap/pigmap.html Pig genome resources at the Roslin Institute, including links to the PigMAP linkage map and PIGBASE pig genome database

Sheep

http://www.marc.usda.gov/genome/sheep/sheep.html Meat Animal Research Center
http://www.projects.roslin.ac.uk/sheepmap/front.html UK Sheep Genome Mapping Project, Roslin Institute

Horse

http://www.vgl.ucdavis.edu/~lvmillon/ Veterinary Genetics Laboratory, UC Davis

Chicken

http://www.ri.bbsrc.ac.uk/chickmap/ChickMapHomePage.html ChickMaP project, Roslin Institute
http://poultry.mph.msu.edu/ US Poultry Genome Project, Michigan State University

Crop plants

A useful overview of plant genome databases can be found at www.cbi.pku.edu.cn/mirror/GenomeWeb/plant-gen-db.html

Cotton

http://algodon.tamu.edu/htdocs-cotton/cottondb.html USDA Agricultural research center

Sorghum

http://algodon.tamu.edu/sorghumdb.html USDA Agricultural research center

Barley

http://ukcrop.net/barley.html Scottish Crop Research Institute

Maize

http://www.agron.missouri.edu/ Comprehensive maize genomics and functional genomics database

Wheat and oats

http://ars-genome.cornell.edu/cgi-bin/WebAce/webace?db=graingenes

mapping to specific locations on the implicated chromosome(s). When eight of the putative 150 QTLs were analyzed further it was possible to map them to the substituted chromosome and three of them were located close to known QTLs of similar phenotype. In an extension of this work, Krewson *et al.* (2004) noted that the onset of puberty in mice occurred earlier in the donor strain than in the host strain. Using the CSS panel, this phenotype was linked to chromosomes 6 and 13. When the CSS-*6* and CSS-*13* strains were crossed with the host strain, the F_1 generation had a timing of puberty onset that was intermediate between the two parents.

The attraction of the CSS mouse panel is that one can use it to identify different QTLs associated with a particular phenotype, e.g. anxiety, and then use the well-tried methods to identify the causative genes. Once these genes have been characterized in the mouse they can be used to probe the human sequence for orthologs. Once the human orthologs have been identified, patient pools can be screened for the presence of SNPs. That is, inbred mouse strains can be used to identify disease genes in outbreeding human populations. Similarly, other quantitative traits could be mapped in the mouse and used to identify orthologs in domestic animals.

The level of gene expression can influence the phenotype of a QTL

Reference to Table 25.1 shows that mutations in coding regions that result in alterations in protein function are responsible for some QTLs. However, other QTLs arise from changes in gene expression. Generally speaking, it is much more difficult to identify such changes because they occur in DNA regions outside the coding region and sometimes at some distance from it. Furthermore, little is known about how variation in DNA sequences might affect the baseline level of gene expression among individuals. The first attempt at quantifying this variation has been undertaken by Morley *et al.* (2004). Using microarrays, they ascertained the baseline expression level of ~8500 active genes in immortalized cells from 14 CEPH families (see p. 362). Then they focused on a subset of genes (~3500) with more variation in expression level between unrelated individuals than between replicate samples within individuals. Attempts then were made to link these genes to chromosomal locations and significant linkage was found for 984 of them. Both *cis*- and *trans*-acting loci were identified although most operate *in trans.*

The significance of the work of Morley *et al.* (2004) is that the level of gene expression is a trait like many others and now is amenable to genetic analysis. Analyzing typical quantitative traits in humans (e.g. blood pressure, levels of serum metabolites) has proved extremely difficult in the past but dissecting the genetic contribution now should be possible.

Understanding responses to drugs (pharmacogenomics)

The perfect therapeutic drug is one that effectively treats a disease and is free of unwanted side-effects. Over the past 25 years many important new classes of drugs have been launched. However, even the most successful and effective of these provide optimal therapy only to a subset of those treated. Some individuals with a particular disease may receive little or no benefit from a drug while others may experience drug-related adverse effects. Such individual variations in response to a drug are responsible for the high failure rates of new drug candidates at the clinical trials stage.

Pharmacogenomics is the study of the association between genomic, genetic, and proteomic data on the one hand and drug response patterns on the other. The objective is to explain inter-patient variability in drug response and to predict the likely response in individuals receiving a particular medicine. As such, pharmacogenomics has the potential to influence the way approved medicines are used as well as have an impact on how clinical trials are designed and interpreted during the drug development process. Relevant information may be derived during the clinical trial recruitment phase, following the treatment phase, or both.

Genetic variation accounts for the different responses of individuals to drugs

There are two fundamental causes of individual responses to drugs. The first of these is variation in the structure of the target molecule. If a drug acts by blocking a particular receptor then it may be that the receptor is not identical in all individuals. A good example is the variation observed in the response of patients with acute promyelocytic leukemia (APL) to all-*trans*-retinoic acid (ATRA). Some patients who contract APL do so because of a balanced translocation between chromosomes 15 and 17 that results in the formation of a chimeric PML-RARα receptor

gene. Other patients have a translocation between chromosomes 11 and 17 that results in the formation of a chimeric PLZF-RARα receptor. These chimeras are believed to cause APL by interference with RAR function. Clearly, the RAR receptor is different in the two types of APL and this is reflected in the fact that only the first type responds to ATRA (He *et al.* 1998).

Another good example of receptor variation is the polymorphism exhibited by the β_2-adrenergic receptor. Nine naturally occurring polymorphisms have been identified in the coding region of the receptor and one of these (Arg16Gly) has been well studied. Asthmatic patients who are homozygous for the Arg16 form of the receptor are 5.3 times more likely to respond to albuterol (salbutamol) than those who are homozygous for the Gly16 form. Heterozygotes give an intermediate response, being 2.3 times more likely to respond than Gly16 homozygotes (Martinez *et al.* 1997). Also of interest is the ethnic variation in frequency of this polymorphism: in Caucasians, Asians, and those of Afro-Caribbean origin it is 0.61, 0.40, and 0.50, respectively.

The second cause of variation in drug response is differences in pharmacokinetics: differences in the way that a particular drug is adsorbed, distributed, metabolized, and excreted (ADME) by the body. A drug that is not absorbed or that is metabolized too quickly will not be effective. On the other hand, a drug that is poorly metabolized could accumulate and cause adverse effects (Fig. 25.11). Obviously, variations in ADME can have multiple causes and the best studied are polymorphisms in drug transport and drug metabolism.

An important polymorphism associated with drug transport is a variant in the multidrug-resistance gene *MDR-1*. The product of this gene is an ATP-dependent membrane efflux pump whose function is the export of substances from the cell, presumably to prevent accumulation of toxic substances or metabolites. A mutation in exon 26 of the *MDR-1* gene (*C3435T*) correlated with alterations in plasma levels of certain drugs. For example, individuals homozygous for this variant exhibited fourfold higher plasma levels of digoxin after a single oral dose and an increased maximum concentration on chronic dosage (Hoffmeyer *et al.* 2000). Other substrates for this transporter are a number of important drugs that have a narrow therapeutic window; i.e. there is little difference between the drug concentration that gives the desired effect and that which is toxic. Therefore, this polymorphism could have a major impact on the requirement for individual dose adjustments for carriers of this mutation. In this context it should be noted that the mutation is probably very common, given that 40% of the German population are homozygous for it.

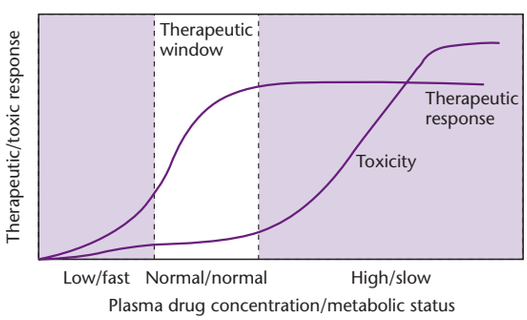


Fig. 25.11 Plasma drug concentrations in different patients after receiving the same doses of a drug metabolized by a polymorphic enzyme.

Most drugs are altered before excretion, either by modification of functional groups or by conjugation with molecules that enhance water solubility (e.g. sugars). Most functional group modifications are mediated by the cytochrome P450 group of enzymes and the genes that encode them are highly polymorphic (for review see Ingelman-Sundberg *et al.* 1999). Over 70 allelic variants of the *CYP2D6* locus have been described and, of these, at least 15 encode non-functional gene products. Phenotypically, four different types can be recognized (Table 25.2). The significance of these phenotypes is illustrated by the response of patients to the drug nortriptyline. Most patients require 75–150 mg per day in order to reach a steady-state plasma concentration of 50–150 μg per liter. However, poor metabolizers need only 10–20 mg per day, whereas ultrarapid metabolizers need 300–500 mg per day. If the genotype or phenotype of the patient is not known then there is a significant risk of overdosing and toxicity (poor metabolizers) or underdosing (ultrarapid metabolizers). The polymorphism of the *CYP2D6* locus is particularly important because it is implicated in the metabolism of over 100 drugs in addition to nortriptyline.

Pharmacogenomics is being used by the pharmaceutical industry

The pharmaceutical industry is using genomics in two ways (Roses 2004). First, it is using genomics at an early stage in clinical trials to differentiate those

Table 25.2 The four drug metabolism phenotypes.

Phenotype	Frequency (%)	Cause
Extensive metabolizers	75–85	
Intermediate metabolizers	10–15	Both gene copies encode an enzyme with lower activity than normal
Poor metabolizers	5–10	Homozygous for two low-activity or non-functional alleles
Ultrarapid metabolizers	2–7	Amplification of the *CYP2D6* locus

individuals who respond well to a drug from those who do not respond. This information then can be used to select responders for later clinical trials. This is known as "efficacy pharmacogenomics". The second application is known as "safety pharmacogenomics". This involves identifying unique markers associated with adverse effects and then managing subsequent risk.

Roses (2004) has provided a good example of a clinical trial where efficacy pharmacogenomics is being used for patient selection. The drug in question is an anti-obesity drug. In early clinical trials patients receiving the drug fell into three classes: non-responders, responders, and hyper-responders (Fig. 25.12). Analysis of the literature suggested 21 genes that might be responsible for the observed effect. Consequently, all the patients were screened for 112 SNPs associated with these 21 genes. One SNP was identified where non-responders were homozygous for one allele and hyper-responders were homozygous for the other allele. As expected, "normal" responders were heterozygous for the alternative alleles. A larger clinical trial is underway to determine if it is possible to exclude non-responders

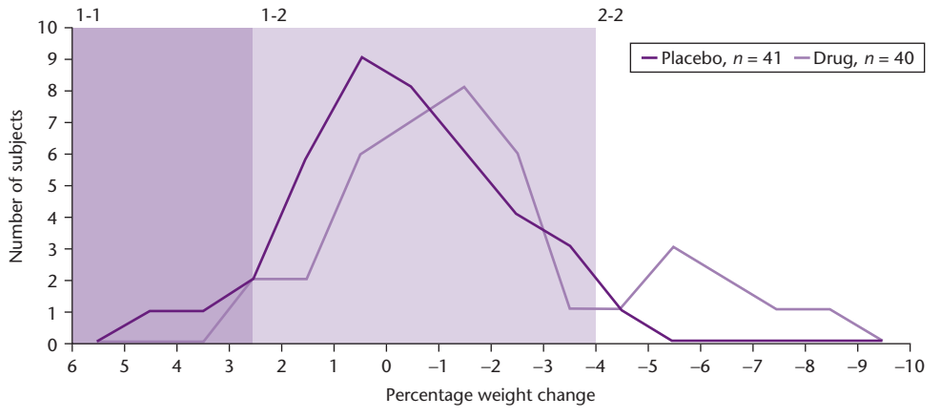


Fig. 25.12 Efficacy pharmacogenetics for an obesity drug. The graph illustrates the results of a small double-blind Phase-IIA efficacy clinical trial of a molecule intended for the treatment of obesity. Weight change in the 40 drug-treated patients (purple line) is compared with 41 placebo-treated patients (blue line) during the two-month double-blind trial. There is an obvious hyper-responder subgroup in the treated patients, with patients losing as much as 9% weight during the trial. Two SNPs from two candidate genes that are related to the proposed mechanism of action of this molecule, and one SNP from another candidate gene that is thought to be implicated in theories of obesity, segregated with the hyper-responders. Each hyper-responder was homozygous for a single allele (labeled 2-2; light-shaded section of the graph) with patients on the left side of the curves being more likely to be 1-1 homozygous (dark purple section). Heterozygous patients (1-2, light purple section) clustered in the middle. In this experiment, treated patients with the 2-2 genotype for any of the three SNPs on average lost ~3.3 kg, whereas treated patients with the 1-1 genotype gained an average of ~1.3 kg. This pattern reassures us that the molecule has efficacy and that subsequent Phase-IIB trials might be enriched by using only patients who are 1-2 heterozygous and 2-2 homozygous, with the exclusion of 1-1 homozygous patients. In addition, although the 1-1 subgroup might be less responsive to this specific treatment, this subgroup could be used in clinical trials of other obesity-drug candidates, which might subsequently allow a drug to be developed that is complementary to the first. *n*, total number of patients in study group. Reproduced from Roses *et al.*, with permission from *Nature*.

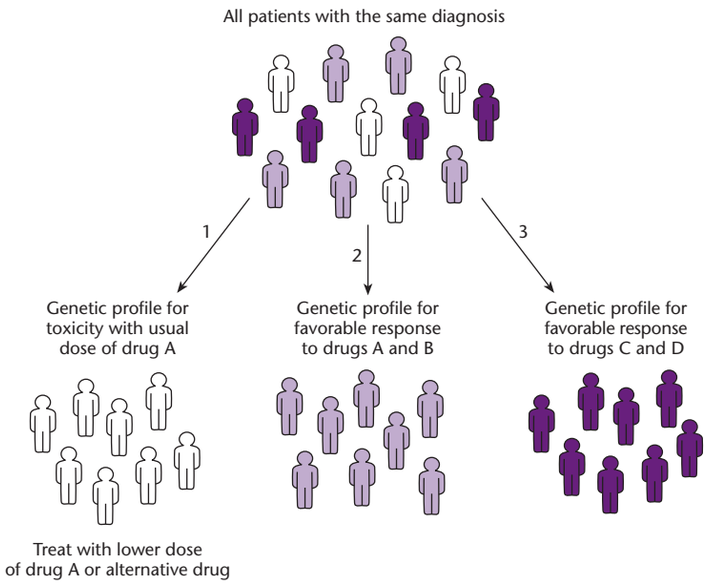


Fig. 25.13 Pharmacogenomics has the potential to subdivide a population of patients with the same empiric diagnosis (e.g. hypertension) into subgroups that have inherited differences in their metabolism of and/or sensitivity to particular drugs. One subset of the population might be at substantially greater risk of serious toxicity (1), whereas other subsets may have receptor polymorphisms or disease pathogenesis polymorphisms that make them more responsive to different treatment options (2 vs. 3).

from trials on the basis of their SNP profile. If so, it will reduce the cost of future clinical trials.

The occurrence of adverse effects to drugs is a major problem for pharmaceutical companies. If the frequency of adverse effects is high then they will be detected at an early stage in clinical development. More often, the frequency is low and significant adverse effects are not seen until post-launch, when tens of thousands of patients have been treated. Serious adverse effects might lead to product withdrawal and hence the accurate identification of individuals at risk would be invaluable. Roses (2004) has calculated that haplotype mapping of as few as 10–20 individuals exhibiting adverse effects (safety pharmacogenomics) could enable "risk" genes to be identified.

Personalized medicine involves matching genotypes to therapy

From the material presented in the early part of the chapter it should be clear that we are beginning to understand the genetic and biochemical causes of common but complex diseases. Initially, this will lead to better classification of the different subtypes of a disease and hence to better diagnoses. This in turn will facilitate selection of the most appropriate therapies. Later, when we know the exact cause of each disease, we should be able to develop drugs that will treat the cause rather than the symptoms as we do at present. Ultimately, genetic analysis of affected individuals will suggest what drugs *could* be used to treat the disease and a second, pharmacogenomic analysis will determine which drugs *should* be used (Fig. 25.13).

A different way of using pharmacogenomic data has been suggested by a study on the chemosensitivity of different cancer cell lines. Staunton *et al.* (2001) used oligonucleotide chips to study the expression levels of 6817 genes in a panel of 60 human cancer cell lines for which the chemosensitivity profiles had been determined. Their objective was to determine if the gene-expression signatures of untreated cells were sufficient for the prediction of chemosensitivity. Gene-expression-based classifiers of sensitivity or resistance for 232 compounds were generated and in independent tests were found to be predictive for 88 of the compounds, irrespective of the tissue of origin of the cells. These results could open the door to the development of more effective chemotherapy regimes for cancer patients.

Suggested reading

Glazier A.M., Nadeau J.H. & Altman T.J. (2002) Finding genes that underlie complex traits. *Science* **298**, 2345–9.

This review provides an excellent introduction to the material covered in this chapter.

Russell R.K., Nimmo E.R. & Satsangi J. (2004) Molecular genetics of Crohn's disease. *Current Opinion in Genetics and Development* **14**, 264–70.

This paper presents a more detailed account of the material presented on Crohn's disease in this chapter.

Paran I. & Zamir D. (2003) Quantitative traits in plants: beyond the QTL. *Trends in Genetics* **19**, 303–6.
This short review describes not only the tomato work described in this chapter but the excellent work done on understanding the evolution of maize.

Barton N.H. & Keightley P.D. (2002) Understanding quantitative genetic variation. *Nature Reviews Genetics* **3**, 11–21.

Flint J., Vladar W., Shifman S. & Mott R. (2005) Strategies for mapping and cloning quantitative trait genes in rodents. *Nature Reviews Genetics* **6**, 271–86.
Two excellent reviews that describe in detail the work done in a number of classical model systems.

Singer J.B., Hill A.E., Burrage L.C., *et al.* (2004) Genetic dissection of complex traits with chromosome substitution strains of mice. *Science* **304**, 445–8.
Like many papers from Eric Lander and his group, this one is destined to become a classic.

Morley M., Moloney C.M., Weber T.M., *et al.* (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–7.
This is another paper that is destined to become a classic.

Roses A.D. (2004) Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nature Reviews Genetics* **5**, 645–56.
This excellent review is just one of a number of reviews on pharmacogenomics that appeared in the September 2004 issue of Nature Reviews Genetics.

Crawford D.C. & Nickerson D.A. (2005) Definition and clinical importance of haplotypes. *Annual Review of Medicine* **56**, 303–20.