
Chromosome Maps

T.P. Speed

*Department of Statistics, University of California at Berkeley, USA
Genetics and Bioinformatics Group, The Walter & Eliza Hall Institute of Medical
Research, Royal Melbourne Hospital, Australia*

and

H. Zhao

*Department of Epidemiology and Public Health, Yale University School of Medicine,
USA*

Chromosome maps are a natural way of organizing genetic data about chromosomes. Existing chromosome maps can be broadly divided into four categories: genetic maps, physical maps, radiation hybrid maps, and gene maps. Although they all make reference to the same biological entity, namely chromosomes, these maps differ substantially in the types of genetic experiments conducted and the types of genetic data collected. They further differ in the metrics employed to define distances and the resolution achievable. Collectively, these maps provide essential tools to further our understanding of the organization and function of the genome. In this review, we first describe the biological principles behind each type of chromosome map and then outline the statistical models and methods that have been developed to construct it. The current state of each chromosome map is summarized, and links to mapping software are provided for readers interested in getting hands-on experience with chromosome mapping.

1.1 INTRODUCTION

Chromosome maps are a natural way of organizing genetic data about chromosomes, in very much the same way that ordinary (cartographic) maps organize geographic data about continents, countries or cities. Geneticists have long constructed different types of maps to order genes or markers, breakpoints, deletions and other features in relation to one another and to landmarks along chromosomes such as centromeres and telomeres. *Genetic maps* were the first type of map constructed to position genes along chromosomes, with the

distances between pairs of genes being defined in terms of recombination fractions. Thus genetic maps were unusual for at least two reasons. Firstly, the objects being mapped – genes, later polymorphic markers, collectively described as loci – were frequently abstract, in the sense that data concerning them was only indirectly observed; the genes themselves were never seen. And secondly, the distance was only relative, and defined statistically. Since the rate of recombination varies along chromosomes, genetic map distance is not proportional to actual physical distance, although there are useful average relationships for different organisms.

Physical maps take a number of forms, but common to all is the fact that the objects being mapped are concrete, usually assayable, and the distances are physical, most recently thousands (kb) or millions (Mb) of base-pairs, reflecting the fact that chromosomes are long DNA molecules. However, the first physical maps were not exactly of this type. Early examples of physical maps are those based on the salivary gland polytene chromosomes of insects belonging to the order Diptera, such as *Drosophila melanogaster*. In these maps the positioning is provided by the visible bands. The familiar cytogenetic maps of human (see Figure 1.1) and other mammalian chromosomes created by staining metaphase chromosomes all have a similar character, again with bands providing positioning information.

Slightly confusingly, physical maps refer not only to maps of loci along chromosomes, but also to organized collections of chromosomal segments, such as restriction fragments and more general ordered sets of cloned fragments of a chromosome.

When recombination is used to define distances, the genes or markers must be polymorphic in order to be mappable. By contrast, the physical mapping of loci only requires probes for recognizing specific chromosomal sites or for detecting fragment

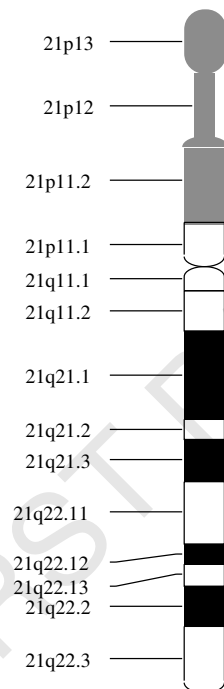


Figure 1.1 Ideogram of human chromosome 21. (Source: www.gdb.org/hugo/chr21/integratedMaps.html.)

overlap. The most useful probe in this context is the *sequence tagged site* (STS), this being defined by a pair of 20–30 base-pairs polymerase chain reaction (PCR) primers which reliably amplify a unique segment in a genome.

Maps of human (and other) chromosomes intermediate in resolution between the genetic and physical maps are the *radiation hybrid* (RH) maps. These are based on assay data from human–rodent somatic cell hybrids containing small fragments of human chromosomes. They have their own metric, namely the average number of breaks per unit physical distance, reflecting the radiation dose used to fragment the chromosomes.

Next in resolution, although different in character, are *gene maps*. These maps are currently constructed by clustering *expressed sequence tags* (ESTs), more specifically short DNA sequences in the 3'-untranslated regions of complementary DNAs (cDNAs), and then locating the clusters on chromosomes using STSs in these regions. Ideally, each point on such a map corresponds to a unique gene, and so gene maps should locate genes on physical maps. Until a genome is completely sequenced, these maps provide the best usable description of the locations of the genes of an organism.

With the exception of the polytene chromosome and cytological maps, all of the kinds of maps we have mentioned make use of statistical methods in their definition, in their construction, or in the assignment of the corresponding map distance. In this chapter we give a review of some of the statistical models and methods used in chromosome mapping. It will be partial in the sense that we discuss genetic mapping much more fully, in part because it is the most thoroughly developed mapping method, going back nearly a century.

1.2 GENETIC MAPS

1.2.1 Mendel's Two Laws

Modern genetics began with the work of Mendel on garden peas in the 1860s (Mendel 1866). In his experiments, Mendel studied a number of heritable traits in peas, including seed color. He interpreted his experiments with this trait by postulating the existence of things which we now call genes. He said that two gene variants controlled color in his two lines, y (for yellow) and g (for green), and that the color gene-pair in the seed determines what color the seed will be. His experiments led him to believe that all cells in the mature plant contain the seed's color gene-pair, with the exception of sex cells, which contain only one of the pair. If the seed's gene-pair is y/g , then half the pollen cells get y and half get g ; similarly for egg cells. Mendel was able to explain his observations with this theory, and it is largely what we believe today. It is often called Mendel's first law, the law of segregation. Mendel's first law says that each adult pea plant has a *gene-pair* (say, y and g) for each character studied, and that the pair y and g segregate from each other into gametes, so half the gametes will carry y , the other half will carry g .

Mendel also considered two or more heritable traits together, for example seed color and seed shape. Denoting the two variants of seed shape by s (for smooth) and w (for wrinkled), he first established that, when considered on its own, seed shape inheritance was also explained by supposing that each cell had one gene-pair, in this case one of the pairs s/s , s/w or w/w , and that sex cells had just one of s or w , effectively chosen at random from the pair generally present. Mendel then carried out experiments to determine

how the two traits, seed shape and seed color, were inherited together. He concluded that each sex cell received one gene from each gene-pair, chosen at random from the available pair, independently for the two gene-pairs. For example, if the mature organism's cells generally possessed gene-pairs y/g and s/w , then its sex cells received ys , yw , gs or gw with equal frequency $\frac{1}{4}$. Let us see the sense in which this last statement is true. Consider an organism P whose gene-pairs for two traits are y/g and s/w , that is descended from a parent GF that was y/y and s/s , and a parent GM that was g/g and w/w . Then P is ys/gw , getting y and s from GF, and g and w from GM. In a natural sense, y and s were combined together in GF, as were g and w in GM, while y and g (and s and w) were separated at that generation, being in different individuals. With peas, when y , g , s , and w corresponded to seed color and shape as above, Mendel saw that this togetherness or separateness in the G-generation had no impact on the choice of genes that P passed on to its offspring C: ys , yw , gs , and gw were found to be passed on with equal frequency. Mendel's second law says that during gamete formation, the segregation of one gene-pair is independent of other gene-pairs. When two gene-pairs, say (y, g) and (s, w) , segregate, each (haploid) gamete will be equally likely to have genotypes (y, s) , (y, w) , (g, s) , and (g, w) .

The above observation, sometimes known as Mendel's law of independent segregation, turns out to hold for some, but not all, pairs of genes. The exceptions are the biological basis for genetic mapping. In the early 1900s, deviations from Mendel's second law were observed by Bateson *et al.* (1905) in the sweet pea, and by Morgan (1911) in *Drosophila*: some genotypes appeared more often than other genotypes, indicating that the gene pairs were not segregating independently. There are many pairs of traits whose genes do not recombine freely, but tend to stay together, in the sense that the parent P above with composition ys/gw would be more likely to pass on the pairs ys and gw to its offspring C, than the pairs yw and gs . This phenomenon is known as *linkage*: genes which came to P together from the G-generation are preferentially passed on together to offspring in the C-generation. In the most extreme case, C would receive each of P's parental combinations ys and gw with frequency $\frac{1}{2}$, and never receive yw or gs . We would then say that the genes are *completely linked*; no recombining takes place. For a given pair of traits such as seed color and seed shape, with heritable variants (*alleles*) such as y , g and s , w , we define their recombination fraction to be the frequency with which P's nonparental combinations yw and gs are passed on; with Mendel's examples this fraction was always $\frac{1}{2}$. In the early part of the twentieth century examples were found where this fraction was noticeably smaller than $\frac{1}{2}$, and to this day, pairs of genes for traits separate into those which freely recombine, and those for which the recombination fraction is less than $\frac{1}{2}$. Using the then much-debated chromosome theory of Mendelian heredity, Morgan explained this nonindependent segregation by supposing the two pairs of genes lie on the same chromosome. A chromosomal exchange between these two genes will result in a recombination between them. Morgan inferred that genes on the same chromosome tend to remain together much more often than if they are on different chromosomes, and called this principle the third law of heredity. He also hypothesized that the cross-shaped structure (called *chiasma*) seen during the diplotene phase of meiosis is a manifestation of *crossing over*. It is now known that crossovers are precise breakage-and-reunion events which are essential for proper segregation, and can promote genetic variation.

1.2.2 Basic Principles in Genetic Mapping

In the following discussion, we make no distinction between *gene*, *marker*, and *locus*, which all refer to some region on the chromosome. Consider two genes \mathcal{A} (with alleles A and a) and \mathcal{B} (with alleles B and b) and a diploid cell with AB and ab on homologous chromosomes. There are four possible meiotic products, namely, AB , ab , Ab , and aB . The first two are called *parental* types or *nonrecombinants*, because both AB and ab retain the configuration of one of the homologous chromosomes. The other two types, Ab and aB , are called *recombinants*. If two markers are recombined in a meiotic product, then during meiosis an odd number of crossovers must have occurred between the two markers on the strand carrying them. The recombination fraction, r_{AB} , is defined as the proportion of recombinants. It was Sturtevant (1913) who first used the variations in the strength of linkage to determine the sequence in the linear dimension of the chromosome. He argued that if the arrangement of the genes in the chromosome is linear and the recombination frequencies depend on the physical distance between them, then genes can be arranged like dots in a straight line at distances apart proportional to the recombination fraction. For example, for three genes, y (yellow gene), w (white gene), and mi (miniature gene) on the sex chromosome of *Drosophila*, the observed recombination fraction between y and w was $r_{y,w} = 1.3\%$, that between w and mi was $r_{w,mi} = 32.6\%$, and that between y and mi was $r_{y,mi} = 33.8\%$. Because $r_{y,mi} \approx r_{y,w} + r_{w,mi}$, the white gene can be inferred to lie between the yellow and miniature genes.

For three genes \mathcal{A} , \mathcal{B} , and \mathcal{C} on the same chromosome in the order $A-B-C$, the additivity among the three recombination fractions, i.e. $r_{AC} = r_{AB} + r_{BC}$, generally holds when the recombination fractions are small (less than 10%). However, as noted by Sturtevant (1913), the additivity in general does not hold when larger recombination fraction values are involved, and usually $r_{AC} < r_{AB} + r_{BC}$. Deviations from additivity are due to the existence of double crossovers. The next major development in genetic mapping was Haldane's definition (1919) of the genetic distance between two loci as the average number of crossovers between the loci per meiosis. This gave geneticists an additive distance along chromosomes, albeit one which was rapidly found not to correlate precisely with any apparent physical distance. The unit of genetic distance is the centimorgan (cM). Two markers are 1 cM apart if on average there is one crossover occurring between these two markers on a single strand for every 100 meioses.

Therefore, we have two basic concepts in genetic mapping: the recombination fraction, which can be estimated from data on the offspring of suitable parents; and map distance, which will be based upon the same data, but can only be estimated using a probabilistic model for recombination. With experimental organisms such as the fruit fly, maize, mice, fungi and yeast, establishing linkage and estimating recombination fractions was generally straightforward, because crosses could be planned, and large numbers of offspring examined. With humans, even establishing linkage between a pair of genes was a major achievement in the classical era, and estimating recombination fractions was a challenging statistical problem. Part of the reason for this lies in the longer generation times, and hence the difficulty in obtaining large sets of data, and part lies in the fact that matings are not subject to experimental control, forcing the human geneticist to make use of nonrandomly sampled family or pedigree data. One further complication with human data was the existence of genes with only an indirect relationship between genotype and phenotype, the issue of penetrance. Dominant and recessive traits are instances of what

are termed incompletely penetrant traits, and there are many human genetic diseases with quite complex patterns of penetrance, including age and sex dependence.

1.2.3 Meiosis, Chromatid Interference, Chiasma Interference, and Crossover Interference

Before we describe in detail statistical methods for genetic mapping, we briefly review the process of meiosis and the genetic concepts relevant to genetic mapping. At the start of meiosis two chromosome sets are present, one coming from each parent in the previous generation. Each chromosome thus has a partner called a *homolog*. During the pachytene and diplotene phases of meiosis, homologous chromosomes pair and each of the paired chromosomes duplicates, resulting in a bundle of four homologous *chromatids*. Chromatids which are copies of the same chromosome are called *sister chromatids*, and those originating from homologous chromosomes are called *nonsister chromatids*. Crossovers take place after the formation of this four-strand structure, with each crossover involving two nonsister chromatids. The number and locations of crossovers vary from chromosome to chromosome for the same meiosis, and from meiosis to meiosis for the same chromosome.

Most genetic mapping efforts have focused on the case where data from only one of the four products of any given meiosis can be observed. Extending terminology from fungal genetics, we call this *single spore data* in recognition of the fact that in organisms such as *Saccharomyces cerevisiae* (baker's yeast) and *Neurospora crassa* (red bread mold) all four products of a single meiosis can be recovered together in what are known as tetrads or octads. Genetic studies on these organisms have contributed greatly to our knowledge of many biological mechanisms. Some interesting statistical models that have been developed using tetrad and octad data will be discussed in later sections.

Mather (1933) distinguished two aspects of crossing over which are relevant to the observed recombination outcome: the distribution of crossover events along the bundle of four chromatids; and the pairs of nonsister chromatids to be involved in crossovers. To distinguish crossover events occurring on the four-strand bundle and crossover events on single strands in the following, we describe crossover events on the four-strand bundle as chiasmata, and those on single strands as crossovers. Chiasma interference refers to nonrandom distribution of chiasmata on the four-strand bundle, whereas crossover interference refers to nonrandom distribution of crossover locations along single strands. Muller (1916) first noted that simultaneous recombinations are not independent, e.g., double recombinations take place at a frequency below that expected under the independence assumption. For example, for the three genes discussed above – yellow, white, and miniature – the expected double recombination frequency is $1.3\% \times 32.6\% = 0.43\%$. However, the observed frequency was only 0.045% . This suggests that the occurrence of one recombination reduces the chance of other recombinations in the nearby region. Crossover interference is seen in almost all organisms, including humans, and the presence of one crossover usually inhibits the formation of crossovers in a nearby region. The biological nature of crossover interference is still not well understood.

With respect to the pairs of nonsister chromatids involved in crossovers, we say there is no chromatid interference (NCI) if any pair of non-sister chromatids are equally likely to be involved in any chiasma, independent of which pairs were involved in other chiasmata.

The observation of crossover interference on the meiotic products (single strands) can be the result of chiasma interference alone, of chromatid interference alone, or of both types of interference. Zhao and Speed (1996) noted that the operation of two types of

interference can lead to no apparent crossover interference, therefore these two types of interference cannot be separated based on single-strand recombination data. In contrast, tetrad data carries information to distinguish these two types of interference.

1.2.4 Genetic Map Functions

Until the mid-1980s, most linkage mapping was two-point, that is, involved the estimation or testing of a single recombination fraction. For two-point data, we can infer the unobservable genetic distance between two markers from the observable recombination fraction through genetic map functions. Under the assumption of NCI, Mather (1935) showed that given k (≥ 1) chiasmata between two markers on the four-strand bundle, the probability of observing recombination between these two markers is $\frac{1}{2}$. Therefore, the overall recombination fraction between two markers is $\frac{1}{2}(1 - p_0)$, where p_0 is the probability of having zero chiasmata between these two markers. This is called Mather's formula. Assuming chiasmata occurring independently of each other, Haldane derived the now well-known Haldane map function relating recombination fraction and map distance: $r = \frac{1}{2}(1 - e^{-2d})$ with inverse $d = -\frac{1}{2} \log(1 - 2r)$. Nearly 90 years later, this approach has proved to be very satisfactory for a wide variety of organisms. Note that Haldane derived his map function under the two-strand model, i.e., assuming only two strands (the two homologous chromosomes) are involved in the crossover process. Although this assumption is incorrect, we would arrive at the same map function under the four-strand model with no chromatid interference.

In addition to deriving the Haldane map function in his seminal 1919 paper, Haldane also proposed the empirical inverse map function $d = 0.7r + 0.3(-\frac{1}{2} \log(1 - 2r))$ to account for crossover interference in the data then available, and introduced a differential equation method which permitted the construction of a variety of map functions. A variety of other genetic map functions embodying different degrees of crossover interference have been proposed, including Ludwig (1934), Kosambi (1944), Carter and Falconer (1951), Sturt (1976), Rao *et al.* (1977), Felsenstein (1979), and Karlin and Liberman (1978).

For all these map functions, genetic distance is very close to recombination fraction when the latter is small, and map distances can be (and in the fly group were) estimated without a model – provided the pair of genes were connected by a sequence of closely linked genes – by adding small recombination fractions.

1.2.5 Genetic Mapping for Three Markers

Historically, the first formal linkage analysis involving more than two loci was given by Fisher (1922). He showed how to combine data from a number of two-point analyses in order to obtain efficient estimates of a set of recombination fractions. For three markers A , B , and C in an arbitrary but fixed order, the joint recombination probabilities may be denoted by $\mathbf{p} = (p_{i_1 i_2})$, where the subscript $i_k = 1$ corresponds to recombination across the k th interval, and $i_k = 0$ corresponds to no recombination across the same interval. Therefore, we have four probabilities $\mathbf{p} = (p_{00}, p_{01}, p_{10}, p_{11})$ for three markers, corresponding to the four patterns of recombination or not across $A-B$ and $B-C$. Although the data were all two-point, Fisher needed to express the recombination fraction across the union $A-C$ of two adjacent intervals $A-B$ and $B-C$ in terms of their individual recombination fractions. He did so by making the assumption of *complete interference*, that is, by assuming that at most one recombination could occur across any pair of adjacent

intervals. This is equivalent to the following joint distribution:

$$p_{00} = 1 - r_1 - r_2; \quad p_{01} = r_2; \quad p_{10} = r_1; \quad p_{11} = 0,$$

where r_1 and r_2 are the recombination fractions across $\mathcal{A}-\mathcal{B}$ and $\mathcal{B}-\mathcal{C}$, respectively. This model would not be appropriate for the analysis of three-point data in which double recombinants are observed, but it has been used in modern times with very short intervals. In human linkage analysis one finds almost exclusive use of the extremely tractable Poisson or *no-interference* model, whose joint probabilities for three loci take the form

$$p_{i_1 i_2} = r_1^{i_1} (1 - r_1)^{1-i_1} r_2^{i_2} (1 - r_2)^{1-i_2},$$

where, for $i = 1, 2$, the recombination fractions r_i may be expressed in terms of genetic distances d_i by

$$r_i = \frac{1}{2}(1 - e^{-2d_i}).$$

It seems that although this model and its extension to more than three loci fail to fit most data sets of any size, the recombination fractions and locus orderings obtained are generally satisfactory; see Speed *et al.* (1992). Any map function, $r = M(d)$, can be used to analyze three-point data. This is because $r_1 = p_{10} + p_{11} = M(d_1)$, $r_2 = p_{01} + p_{11} = M(d_2)$, and $p_{11} = \frac{1}{2}[M(d_1) + M(d_2) - M(d_1 + d_2)]$, and we can derive all the $p_{i_1 i_2}$ from a given map function. Therefore, likelihood functions for the observed data can be constructed and maximum likelihood estimates of genetic distances can be obtained.

For an arbitrary crossover process model, under the assumption of NCI, Speed *et al.* (1992) derived a set of inequality constraints and showed the robustness of the ordering. The order with respect to which these probabilities are defined does not need to be the true one, and if we change it, the probabilities need only be relabelled. For example, if we go from the order $O : \mathcal{A}-\mathcal{B}-\mathcal{C}$ with probabilities \mathbf{p} , to $O' : \mathcal{A}-\mathcal{C}-\mathcal{B}$ with probabilities \mathbf{p}' , then \mathbf{p}' is related to \mathbf{p} as follows:

$$p'_{00} = p_{00}, \quad p'_{10} = p_{10}, \quad p'_{01} = p_{11}, \quad p'_{11} = p_{01}.$$

Three-point phase known crosses (in which allelic combinations across loci are together on the same chromosome) have been used for decades to order loci in experimental organisms, without any explicit model assumptions. This works because, under very general conditions, the smallest of the four probabilities ($p_{i_1 i_2}$) corresponds to the event of double recombination across two consecutive intervals when the loci are correctly ordered. For example, if the correct order is $O : \mathcal{A}-\mathcal{B}-\mathcal{C}$, then (assuming no chromatid interference)

$$p_{11} \leq p_{10}, p_{01} \leq p_{00}.$$

If, on the other hand, $O' : \mathcal{A}-\mathcal{C}-\mathcal{B}$ is the correct order, but we have written our probabilities relative to O , then $p'_{11} = p_{01}$ will be the smallest probability. It follows that with sufficiently large samples of data, any set of loci can be ordered by inspection, with only a small chance of error. Naturally this is also possible using only the pairwise recombination fractions, but that would take more data to achieve the same level of confidence in the ordering. More generally, it is possible to show that, under the assumption of NCI, a multipoint recombination probability decreases, or at least does not increase, when any nonrecombinant interval is changed to recombinant status; see Speed *et al.* (1992).

1.2.6 Genetic Mapping for Multiple Markers

Although inefficient from the statistical viewpoint, three or more loci can be mapped using only two-point data, since linear maps are determined by pairwise distances. When there are plenty of data, such as with *Drosophila*, multipoint analyses may be unnecessary. However, in most contexts, data are scarce. In such cases, multipoint linkage analysis can be viewed as an attempt to make more efficient use of recombination data to further the aims of linkage analysis; see Lathrop *et al.* (1984) and Thompson (1984).

Multipoint linkage analyses make fuller use of available data, and can achieve greater precision or power. They are more complex than two-point analyses in several important ways. First, they require the specification of an order for the loci. Second, they require the specification of a joint distribution for all possible recombination patterns: for n loci, there are 2^{n-1} such patterns (including the parental one). Third, from the perspective of parametric statistical inference, joint distributions over recombination patterns corresponding to distinct orderings of the loci define noncomparable statistical models. Most of the difficulties of multipoint linkage analysis stem from these facts, particularly the rate of increase of the number of orders or patterns with the number of loci. When linkage analysis is being done using pedigree data, the size (number of individuals) and complexity (presence of one or more loops) of the pedigrees are additional limiting factors.

At the initial stage of genetic mapping, linkage groups have to be defined. Two markers are in linkage if the recombination fraction between them is less than $\frac{1}{2}$. A linkage group is defined as a set of markers where each marker is linked to at least one other marker in the same set. With enough markers covering the genome, each linkage group will correspond to a chromosome. However, for three markers A , B , and C , it is possible that A and B are genetically linked, B and C are genetically linked, yet the recombination fraction between A and C may be approaching $\frac{1}{2}$ if they are sufficiently far apart from each other on a chromosome. Although linkage groups have been well defined for humans and some experimental organisms, linkage group construction remains the critical first step for many organisms at the early stage of genetic mapping.

To define whether two markers are in linkage is to test whether the recombination fraction between these two markers is less than $\frac{1}{2}$. This hypothesis testing problem can be carried out using the likelihood ratio test; see Ott (1999). The LOD (log-odds) score is often used to assess the evidence for linkage. It differs from the usual likelihood ratio statistic by a constant factor and is defined as

$$\text{LOD} = \log_{10} \frac{L(\text{data}|r)}{L(\text{data}|r = \frac{1}{2})}.$$

A LOD score of 3 has been used as the threshold for linkage testing. The justification of this threshold is discussed by Ott (1999) and Risch (1991).

After linkage groups are defined, the next task is to order genetic markers within each group. The locus ordering problem resembles the traveling salesman problem (TSP) widely discussed in the field of combinatorial optimization, (see Johnson, 1990), in which there are a large number of discrete states, each of which can be assigned a numerical value by a cost or objective function. The calculation of the objective function can depend either on information from pairwise data (e.g., pairwise LOD scores) or on joint genetic information (e.g., multipoint LOD scores, discussed later). For example, Speed *et al.* (1992) showed that under the assumption of NCI, a given order imposes linear

constraints among multilocus recombination probabilities. Maximum likelihood under these constraints for each order can be used as the objective function.

With n markers, the ideal ordering approach would be to compute the objective function for each of the $\frac{1}{2}n!$ orders, and then to rank the orders, choosing as the best order the one with the largest objective function. With a few markers, all possible orderings can be considered. However, this quickly becomes impossible with many genetic markers. There is no evidence to suggest that a method exists which is generally better than choosing that order which maximizes the likelihood of the data using a suitable recombination model, at least not when the calculation of the likelihoods corresponding to each of the $\frac{1}{2}n!$ distinct orders is possible. The Poisson or no-interference model is the one typically used in this context. Although there does not appear to be a systematic study of this issue, the available evidence suggests that only small gains in the efficiency of ordering loci are to be found by using a more suitable model when interference exists; see Lathrop *et al.* (1984), Bishop and Thompson (1988), Goldgar and Fain (1988), Speed *et al.* (1992) and Goldstein *et al.* (1995) for related results. Different heuristic ordering strategies were reviewed in Weeks (1991), but the programs that are currently widely used to order loci based on human or other pedigree data make little or no use of recent research from the field of combinatorial optimization.

In our previous discussion on genetic distance estimation from two-point or three-point genetic data, we described how map functions can be used to estimate genetic distances. However, when there are more than three markers, the multilocus recombination probabilities cannot be uniquely determined from the map function. A crossover process model is needed to derive joint multilocus recombination probabilities. Several point process models (Fisher *et al.*, 1947; Karlin and Liberman, 1979; Risch and Lange, 1979; and King and Mortimer, 1990) have been proposed to incorporate crossover interference in modeling the crossover process. The first satisfactory class of recombination models were the chi-square renewal process models discussed by Fisher and his students and colleagues (Fisher *et al.*, 1947). Bailey (1961) gave a good overview of this research. The simplest of these joint probabilities is too complex to be given here, and this is probably the reason why this class of models has not been used with human data until recently (Lin and Speed, 1996). The chi-square model has been extended to the Poisson-skip model which has the chi-square model as its special cases and can also incorporate negative crossover interference; see Lange *et al.* (1997). The major alternatives to the chi-square renewal models are due (independently) to Karlin and Liberman (1979) and Risch and Lange (1979), called count-location or generalized no-interference models, and the model of Goldgar and Fain (1988). For a review and comparison of different stochastic models for recombination, see McPeck and Speed (1995). One approach that does not depend on specific models for recombination was developed by Weinstein (1936) and was recently used to study human meiosis by Lamb *et al.* (1997), Zhao *et al.* (2000) and Li *et al.* (2001). The only assumption employed by this approach is that there is at most one chiasma in each marker interval, which is likely to be satisfied when many markers are studied on a chromosome. Although substantially more parameters are involved, the results from Weinstein's approach can be used to assess the goodness-of-fit of different crossover process models and to identify anomalous features of these models.

Liberman and Karlin (1984) proposed to extend genetic map functions to four or more marker cases by embodying the assumption that, for a pair of noncontiguous intervals, the probabilities for joint recombination patterns across these intervals do not depend on

the distance between the intervals, something which is not consistent with observations. Those map functions which can be extended to multilocus data through this approach have been (inappropriately) called ‘multilocus feasible’ by Liberman and Karlin (1984). This criterion excludes many functions which were found to fit well to recombination data, such as the Kosambi map function proposed by Kosambi (1944). However, Zhao and Speed (1996) showed that there exist stationary renewal processes which give rise to most map functions in the literature (including the Kosambi map function). Therefore, these map functions are compatible with the analysis of multilocus data via this approach. Moreover, the inter-event distributions of the stationary renewal processes corresponding to most map functions can be closely approximated by gamma distributions.

We have discussed the cases where recombination or nonrecombination can be unambiguously scored. For human pedigrees, matters are more complicated at many levels. As with two-point linkage analyses, a major complication in multipoint linkage analyses can be the incompleteness of data. For example, there may be missing data due to some individuals not being typed. All data may be available, but phenotype may not determine genotype, as with dominant traits and other types of incomplete penetrance. Genotypes may be known, but haplotypes may not. That is, phase may be unknown. With known genotypes at n loci, there are 2^{n-1} possible haplotypes. While these incompleteness problems can slow down two-point analyses, they can quickly make exact multipoint analyses impossible. On the other hand, multipoint analyses can make use of data that cannot be used in two-point analyses, for example, when only uninformative data are available at a locus intermediate between two fully informative loci; see Lathrop *et al.* (1985) and Ott (1999). In multipoint linkage analysis using pedigree data, the feasibility of an exact analysis will depend on the number of loci, the size and complexity of the pedigrees involved, and the nature and extent of incompleteness in the data.

For pedigrees with simple structures or with a few genetic markers, the likelihood for a pedigree can be calculated exactly. The exact calculations can be divided into two types of algorithms: the Elston–Stewart algorithm (Elston and Stewart, 1971) and the Lander–Green algorithm (Lander and Green, 1987). Consider a pedigree with m individuals, where $\mathbf{x} = (x_1, x_2, \dots, x_m)$ is the observed phenotypes for the pedigree. If G_i is the set of genotypes g_i compatible with the phenotype of person i , then the likelihood of the pedigree can be written as a sum of products:

$$\sum_{g_1 \in G_1} \cdots \sum_{g_m \in G_m} \prod_i P(x_i | g_i) \prod_{k \text{ founders}} P(g_k) \prod_{\{i_1, i_2, i_3\}} P(g_{i_1} | g_{i_2}, g_{i_3}),$$

where $\{i_1, i_2, i_3\}$ is an offspring–parent triad and i refers to the individuals with observed phenotypes. The probability $P(x_i | g_i)$ is the probability of an individual with genotype g_i having phenotype x_i . For codominant genetic markers, the probability is either 1 or 0. The founder probability $P(g_k)$ is a function of population gene allele frequencies. The Elston–Stewart algorithm can be viewed as a method for choosing an order to perform the iterated sum to minimize the total number of additions and multiplications. The number of calculations in the Elston–Stewart algorithm scales linearly with the number of individuals in the pedigrees but exponentially with the number of markers.

The Lander–Green algorithm works as follows. Let $\mathbf{x}_L = (\mathbf{x}_{L_1}, \mathbf{x}_{L_2}, \dots, \mathbf{x}_{L_N})$ denote the collection of phenotypes at locus i , and $\mathbf{g}_L = (\mathbf{g}_{L_1}, \mathbf{g}_{L_2}, \dots, \mathbf{g}_{L_N})$ denote the collection of ordered genotypes at these loci for the individuals. Then the likelihood for the pedigree

can be written as

$$\sum_{g_{L_1} \in L_1} \cdots \sum_{g_{L_N} \in L_N} \left[\prod_i P(x_{L_i} | g_{L_i}) \right] P(g_{L_N} | g_{L_{N-1}}, g_{L_{N-2}}, \dots, g_{L_1}) \cdots P(g_{L_2} | g_{L_1}) P(g_{L_1}).$$

Assuming no crossover interference, then the likelihood is

$$\sum_{g_{L_1} \in L_1} \cdots \sum_{g_{L_N} \in L_N} \left[\prod_i P(x_{L_i} | g_{L_i}) \right] P(g_{L_N} | g_{L_{N-1}}) \cdots P(g_{L_2} | g_{L_1}) P(g_{L_1}).$$

The Lander–Green algorithm can be extended to incorporate the chi-square model in linkage analysis. The Elston–Stewart algorithm is mostly useful for large pedigrees but only a limited number of markers, whereas the Lander–Green algorithm is useful for multiple markers but is limited in the number individuals in each pedigree. This likelihood can be efficiently evaluated using the forward–backward algorithm of the hidden Markov model methodology. In addition, parameter estimates can be obtained using the EM algorithm. The number of operations scales linearly with the number of markers but exponentially with the number of individuals in the pedigree.

Both algorithms will fail if we have large pedigrees with many markers typed, and simulation methods have been proposed to approximate the likelihood by Thompson (1994) and Sobel and Lange (1996). In a recent review, Lin (1996) discussed both the sequential imputation approach of Irwin *et al.* (1994) and Markov chain Monte Carlo methods of Lin and Wijsman (1994).

1.2.7 Tetrads

Recall that tetrads and octads refer to the case where all four products of a single meiosis can be recovered together, such as in yeast and bread mold. Octads are generated from tetrads following one mitosis, and the octads can usually be represented by tetrads, except when gene conversions occur. If we ignore the possibility of gene conversions, we need make no distinction between tetrads and octads in the following discussion, and refer to both as tetrads. Genetic studies using tetrad data are very valuable in studying the crossovers during meiosis. Compared to single-spore data, tetrad data have several advantages. First, with tetrad data chromatid interference and chiasma interference can be distinguished. Second, when chromatid interference is absent, chiasma interference can be detected with only two markers, whereas at least three markers are needed for single-spore data. Chiasma interference can even be detected with one marker in some studies. Third, the position of the centromere can be inferred. In some organisms, such as *Neurospora crassa*, the tetrads are produced in a linear order corresponding to the meiotic divisions; these are called ordered tetrads. In others, such as *Saccharomyces cerevisiae*, the tetrads are produced as a group without order, and are called unordered tetrads.

If a cross involves two strains differing with respect to two genes, geneticists distinguish three possible tetrad types: parental ditype with two representatives of each of the two parental types; nonparental ditype, where all four strands show recombinant types; and tetratype, where two of the four strands show parental types and the other two strands show recombinant types. For tetrad data involving two genetic markers, let P , T , and N denote the proportion of tetrads having parental ditype, tetratype, and nonparental ditype, respectively. The recombination fraction between the two markers can then be estimated

by $N + \frac{1}{2}T$. Although the genetic distance can be estimated from this recombination fraction through a genetic map function, there is more information in the raw tetrad data. Under the assumption of NCI, given two chiasmata between two markers, the probabilities of observing parental ditype, tetratype, and nonparental ditype are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively. If there is a single chiasma between two markers, the resulting tetrad always have tetratype. Therefore, we can estimate the probability of having two chiasmata by $4N$, and the probability of having one chiasma by $T - 2N$. This leads to an estimated distance of $\frac{1}{2}(T + 6N)$ if we assume there are no more than two chiasmata between the two markers. This formula was first proposed by Perkins (1949).

Under the assumption of NCI, Mather (1935) showed that if $k \geq 1$ chiasmata occur between a pair of markers, then the conditional probabilities p_0^k , p_1^k , and p_2^k of observing a tetrad with parental ditype, tetratype, and nonparental ditype, respectively, are

$$\begin{aligned} p_0^k &= \frac{1}{3} \left(\frac{1}{2} + \left(-\frac{1}{2}\right)^k \right), \\ p_1^k &= \frac{2}{3} \left(1 - \left(-\frac{1}{2}\right)^k \right), \\ p_2^k &= \frac{1}{3} \left(\frac{1}{2} + \left(-\frac{1}{2}\right)^k \right). \end{aligned}$$

For a given crossover process model, the above relations can be used to relate the probabilities of three tetrad patterns to the genetic distance between two markers. For example, under the Poisson model, $p_0 = \frac{1}{6}(1 + 2e^{-3d} + 3e^{-2d})$, $p_1 = \frac{2}{3}(1 - e^{-3d})$, and $p_2 = \frac{1}{6}(1 + 2e^{-3d} - 3e^{-2d})$, where p_0 , p_1 , and p_2 are the probabilities of parental ditype, tetratype, and nonparental ditype between two markers, respectively, and d is the genetic distance; see Haldane (1931).

One unique feature of ordered tetrad analysis is that there is information on centromeres. The distance between a single marker and its centromere can be estimated using data from a single marker. For marker A with alleles A and a inherited from two parents, there are six distinguishable configurations, as illustrated in Table 1.1. Because spindle-centromere attachment during meiosis is random (see Griffiths *et al.*, 1996), types 1 and 6 have equal probability because of random spindle-centromere attachment at the first meiotic division, whereas types 2–5 have the same probability because of random spindle-centromere attachment at the second meiotic division. Types 1 and 6 are called first-division segregation (FDS) pattern and types 2–5 are called second-division segregation (SDS) pattern (Griffiths *et al.*, 1996).

The probabilities of FDS and SDS can be related to the genetic distance between A and its centromere if a chiasma process model is specified. Let $S_A = P(\text{SDS})$; then

Table 1.1 Six distinguishable patterns for marker A . Strands 1 and 2 are attached to one centromere and strands 3 and 4 are attached to the other.

Strand	S_1	S_2	S_3	S_4	S_5	S_6
1	A	A	A	a	a	a
2	A	a	a	A	A	a
3	a	A	a	A	a	A
4	a	a	A	a	A	A

$S_A = c_1 = 2d$ under the complete interference model, where c_1 denotes the probability of having one chiasma. For the Poisson model, $S_A(d) = 1 - F_A(d) = \frac{2}{3}(1 - e^{-3d})$. Several chiasma models and various map functions derived from these models were studied by Zhao and Speed (1998a). It was found that most map functions proposed in the literature can be well approximated by the map functions under the chi-square model. Centromeres can also be mapped with three markers on three different chromosomes using unordered tetrads, as shown by Whitehouse (1957). For three markers A , B , and C , denote the frequencies of SDS for these three loci by x , y , and z . Then the probability of tetratype between A and B is $T_{AB} = x + y - \frac{3}{2}xy$ when A and B are on different chromosomes. Similarly, $T_{AC} = x + z - \frac{3}{2}xz$ and $T_{BC} = y + z - \frac{3}{2}yz$. These three equations can be used to solve for three unknown parameters. For example,

$$x = \frac{2}{3} \left\{ 1 \pm \sqrt{\frac{4 - 6T_{AB} - 6T_{AC} + 9T_{AB}T_{AC}}{4 - 6T_{BC}}} \right\}.$$

However, this method only has reasonable precision when at least two of the three loci are fairly close to their respective centromeres.

For a cross involving three markers A , B , and C on the same chromosome, if both marker intervals ($A-B$ and $B-C$) show tetratypes, there are three types that can be distinguished according to the pattern between A and C : parental ditype, tetratype, and nonparental ditype (often called two-strand, three-strand, and four-strand doubles). Under the assumption of NCI, the ratio of these three types is 1 : 2 : 1. A significant deviation from the expected ratio can be attributed to the presence of chromatid interference. Geneticists have examined this ratio in different organisms and, overall, found no consistent evidence against the NCI assumption; see Fincham *et al.* (1979). Although most studies on ordered tetrads and unordered tetrads used only three loci for the detection of chromatid interference, some information is lost when only the 1 : 2 : 1 ratio is examined for each pair of marker intervals. Zhao *et al.* (1995b) derived a set of linear equality and inequality constraints on the probabilities of unordered tetrad patterns with an arbitrary number of loci under the assumption of NCI. For example, for two markers, NCI imposes that $p_0 \geq p_2$ and $p_1 \geq 2p_2$. Similar constraints were derived for ordered tetrads by Zhao and Speed (1998a). These constraints can be used to test the presence of chromatid interference without assuming any model for the chiasma process.

To perform genetic mapping using multiple markers simultaneously, we need to be able to evaluate the probability for any multilocus tetrad pattern under a given model for the chiasma process. Both the count-location model (Risch and Lange, 1983) and the chi-square model (Zhao *et al.*, 1995a; Zhao and Speed, 1998a) have been applied to analyze tetrad data. Detailed procedures can be found in these papers.

1.2.8 Half-tetrads

Half-tetrads can arise either from meiosis I or meiosis II nondisjunctions. The first well-studied half-tetrad data were attached-X chromosomes in *Drosophila* (Beadle and Emerson, 1935). Half-tetrads were also constructed using autosomes in *Drosophila* (Baldwin and Chovnick, 1967), and have been used in the study of many other organisms, including maize, potatoes, leopard frog, rainbow trout, salmonid fish, catfish, and zebrafish. In mammals, half-tetrads can be studied in the form of uniparental disomy

(Robinson *et al.*, 1993), autosomal trisomies (Morton *et al.*, 1990), nondisjunction in ovarian teratomas (Eppig and Eicher, 1983), and PCR analysis of meiosis I products in individual secondary oocytes (Cui *et al.*, 1992).

Genetic mapping using genetic marker data from human nondisjunction data was discussed by Shahar and Morton (1986), Chakravarti and Slaugenhaupt (1987), Chakravarti *et al.* (1989), and Feingold *et al.* (2000). Map distances, as well as LOD scores for these distances, can be calculated from the observed patterns of nonreduction (heterozygous genotype) and reduction (homozygous genotype) of markers along the nondisjoined chromosome pair. Zhao and Speed (1998b) derived the general relationship between multilocus half-tetrad probabilities and multilocus ordered tetrad probabilities. These relationships can be used for likelihood analysis of half-tetrads, and the same methods have been extended to study uniparental disomy and trisomy.

1.2.9 Other Types of Data

Genetic maps can also be constructed using other types of data, including organisms with more than two copies of chromosomes (Bailey, 1961; Wu *et al.*, 2001), bacterial and bacteriophage (Stahl, 1979), and recombinant inbred strains (Green, 1981). Genetic background and statistical methods for these types of data can be found in these references.

1.2.10 Current State of Genetic Maps

Statistical methods for establishing linkage and estimating recombination fractions in humans were pioneered by Bernstein (1931), and developed intensively by the British school centered around Fisher and Haldane during the 1930s and 1940s. The first human linkage to be established was between the X-linked genes for hemophilia and red-green color blindness by Bell and Haldane (1937); two decades later, Mohr (1954) found linkage between two blood groups on an autosome. Early ways of establishing linkage were based upon what are now known as score tests, and a method using sib-pairs, while likelihood methods quickly came into use for estimation. Several methods of correcting for sampling biases were also developed. All of these ideas continue to be important today.

A major limitation in human genetic mapping before the 1980s was the shortage of genetic markers. Markers are Mendelian factors, often but not necessarily genes in the modern sense, which segregate in human populations. For many years human genetic markers were mainly blood cell antigens and proteins. They provided the basis of human genetic maps, and were a framework within which new genes could be mapped. Despite there being a fair number of known genetic diseases and Mendelian markers such as those just mentioned, the human genetic map was still very sparse in the 1970s. However, it was during this period that the first good algorithms for calculating probabilities over pedigrees were developed, motivated initially by problems in genetic counseling, and later by the desire to carry out segregation analyses on large pedigrees. Programs based upon these algorithms continue to play a very important role in modern genetic mapping.

Around 1980, the idea of treating DNA sequence differences as genetic markers arose. It was quickly developed, and the present wide availability of what are collectively known as molecular markers has revolutionized human genetic mapping, and that of many other organisms. Development of the CEPH reference families (Dausset *et al.*, 1990) was a critical step in genetic map construction. The first fairly complete human map was published in 1987, and consisted of about 400 restriction fragment length polymorphisms

(RFLPs), mapped using DNA from a panel of 21 three-generation families from the CEPH consortium (Donis-Keller *et al.*, 1987). In order to build this map, new multilocus methods for mapping were developed. The mapping of many loci simultaneously was first carried out by Fisher in 1922, but it was only following the availability of cheap, fast computers and suitable algorithms that this idea became widely adopted.

At the time the 1987 map was being developed, the PCR was beginning to revolutionize molecular genetics. Several new types of genetic markers have been developed using PCR, with acronyms such as RAPD (random amplified polymorphic DNAs), STRP (short tandem repeat polymorphism) and SSCP (single-strand conformation polymorphism), and the latest genetic maps include several thousand readily assayed markers. It is now possible to carry out genome-wide scans, effectively searching the entire genome for linkage between a trait and markers. Searches of this kind have been remarkably successful in locating genes contributing to a wide range of disease and other phenotypes. They also raise many new statistical questions, especially as interest now focuses on complex and quantitative traits. Such traits are believed to be influenced by a number of genes, as well as the environment, and mapping these genes with available data remains a challenging task.

In recent genetic maps constructed from these pedigrees, more than 8000 STRPs were mapped to the human genome by Broman *et al.* (1998). This map not only provides guidelines for disease gene mapping, but also allows a very detailed comparison between male and female genetic maps (Broman *et al.*, 1998) and the study of crossover interference (Broman and Weber, 2000). More recently, Kong *et al.* (2002) estimated recombination rates across the human genome through 5136 microsatellite markers typed for 146 families, with a total of 1257 meiotic events. They detected 'systematic differences in recombination rates between mothers and between gametes from the same mother, suggesting that there is some underlying component determined by both genetic and environmental factors that affects maternal recombination rates'. In Figure 1.2 we show some features of portions of several genetic maps for human chromosome 21.

Genetic maps for pigs, cows, tomatoes, rice, pine trees and many other species of commercial or scientific interest have followed quickly behind the human maps.

1.2.11 Programs for Genetic Mapping

The programs described here can all be found in the website maintained by the Ott group at Rockefeller University, at <http://linkage.rockefeller.edu/soft/list.html>. As discussed above, algorithms for carrying out multipoint linkage analysis with human (and other) pedigree data are of two kinds: those based upon the Elston and Stewart (1971) approach, using what is known as peeling; and those based upon the Lander and Green (1987) hidden Markov model formulation. Each of these classes of algorithms has its strengths and weaknesses, and there are problems that cannot be solved exactly with either of them. The Elston–Stewart approach underlies most of the algorithms discussed in Terwilliger and Ott (1994). For a recent improvement of the implementation of these algorithms, see O'Connell and Weeks (1995). A suite of genetic mapping programs that have gained much popularity uses the basic Lander–Green algorithm in a number of different human linkage problems. These include MAPMAKER for crosses among inbred strains (Lander *et al.*, 1987), analyses with sib-pairs (Kruglyak and Lander, 1995), the analysis of recessive traits with nuclear families (Kruglyak *et al.*, 1995), and multipoint linkage with many markers for general pedigrees of moderate

size (Kruglyak *et al.*, 1996). Similar statistical methods underlie the program CRI-MAP, which is most suitable for CEPH-type pedigrees (Green, 1988). MultiMap is another program that assists with map construction (Matise *et al.*, 1994). It consists of framework construction and comprehensive map construction. MultiMap recently incorporated the Gene Mapping System algorithm (Lathrop *et al.*, 1988; Marinov *et al.*, 1999), which is based on identifying and permuting linkage groups within an initial order of all loci. Other programs include Map/Map+ for map integration (Morton *et al.*, 1992), JoinMap for plants (Stam, 1993), and OUTMAP for outbred populations (Ling, 2000).

When exact linkage analysis methods fail because of time or space constraints, Monte Carlo methods may be used. At present these are more research tools than approaches suitable for routine use, but they are developing rapidly, and should become more widely used in the near future. Some of these simulation methods have been implemented in SIMWALK (Sobel and Lange, 1996) and Morgan (Thompson, 1994).

1.3 PHYSICAL MAPS

Physical mapping is the process of determining the locations of 'sites' such as restriction sites (4–8 bp), sequence-tagged sites (20–30 bp) and cloned fragments (kb to Mb) on a larger DNA molecule or a chromosome. Among other things, maps of such sites are helpful if not essential for cloning genes and for sequencing large stretches of DNA, and have been very widely used in recent years. To quote from an early successful paper in the area, Olson *et al.* (1986, p. 7830):

a strong case can be made for the value of constructing physical maps of the genomes of intensively studied organisms. We expect the main value of these maps to lie in facilitating the organization of molecular genetic information. Just as conventional cartography provides an indispensable framework for organizing data in fields as disparate as demography and geophysics, it is reasonable to suppose that 'DNA cartography' will prove equally useful in organizing the vast quantities of molecular genetic data that may be expected to accumulate in the coming decades. Furthermore, the principal by-product of these projects – global clone collections that are cross-indexed to the physical maps – could be expected to improve the efficiency of subsequent structural and functional studies of local regions.

The two most common approaches to physical mapping are termed top-down, producing a macrorestriction map, and bottom-up, resulting in a contig map. With either strategy, the maps represent ordered sets of DNA fragments that are generated by cutting genomic DNA.

However, the first physical maps were made from microscope images, and although their construction and interpretation involve no statistics, we discuss them briefly for completeness.

1.3.1 Polytene Chromosomes

Polytene chromosomes are many-stranded chromosomes resulting from repeated chromosomal replication, without the subsequent separation of sister chromatids. Up to 1024

chromatids can be present, giving giant chromosomes visible under a microscope in nondividing cells. Most widely known are those of the salivary glands of insects of the order Diptera, and a classic reference to the polytene chromosomes in the salivary glands of *Drosophila melanogaster* is Bridges (1935).

After appropriate staining, the *Drosophila* polytene chromosomes have distinctive banding patterns which have been cataloged, and have proved invaluable for localizing structural alterations such as deletions, and for use with more recent *in situ* hybridization techniques. For further details, we refer to Saura *et al.* (1997), and to FlyBase (<http://flybase.bio.indiana.edu>).

1.3.2 Cytogenetic Maps

A closely related class of physical maps are the familiar cytological maps whose human versions are frequently represented in ideogrammatic form (see Figure 1.1). Such maps were originally derived in a variety of organisms by associating mutant phenotypes with chromosomal defects visible by direct microscopic examination. In this way, genes can be physically located on chromosomes, at least to a low level of resolution.

In the late 1960s, staining techniques were discovered which led quickly to the adoption of banding patterns of human chromosomes now widely used; see, for example, Vogel and Motulsky (1997). This field has evolved greatly in recent years with the advent of *fluorescent in situ hybridization* (FISH) and multiple colouring of chromosomes; see Trask (1998).

1.3.3 Restriction Maps

A *restriction site* is the location of a sequence, typically 4–6 bp long, where a particular restriction enzyme will cut DNA. Isolated from various bacteria, restriction enzymes recognize short DNA sequences and cut DNA molecules at specific sites in the sequence. Ignoring for the moment variations from uniform base composition, restriction enzymes with 4 bp recognition sites will yield pieces – termed *restriction fragments* – on average about 256 bp long, while those with 6 or 8 bp recognition sites will yield pieces of average length 4 or 64 kb, respectively. Since hundreds of different restriction enzymes have been characterized, and they can be used together, DNA can now be cut with them into fragments of many different sizes in many different ways. A restriction map describes the order and distance between restriction sites.

In *top-down mapping*, a chromosome is cut into large DNA fragments using restriction enzymes having rare restriction sites. The fragments are separated by size and assigned to regions by hybridization with genetically or cytogenetically mapped DNA probes. Then the fragments are assembled into contiguous blocks, resulting in a macrorestriction map. Such fragments may average 1 Mb in size. For a finer map, the ordered fragments may be taken one at a time and dissected with more frequently cutting restriction enzymes.

The simplest way to construct a restriction map is to compare the fragment sizes produced when a DNA molecule is digested with two different restriction enzymes; see Waterman (1995, Chapters 2–4) for a discussion of some of the computational issues here. Restriction maps are easy to generate if there are relatively few cut sites with the enzymes being used, but most enzymes cut frequently, generating a large number of small fragments (from less than 100 bp to 1 kb). Therefore, such mapping is more applicable to

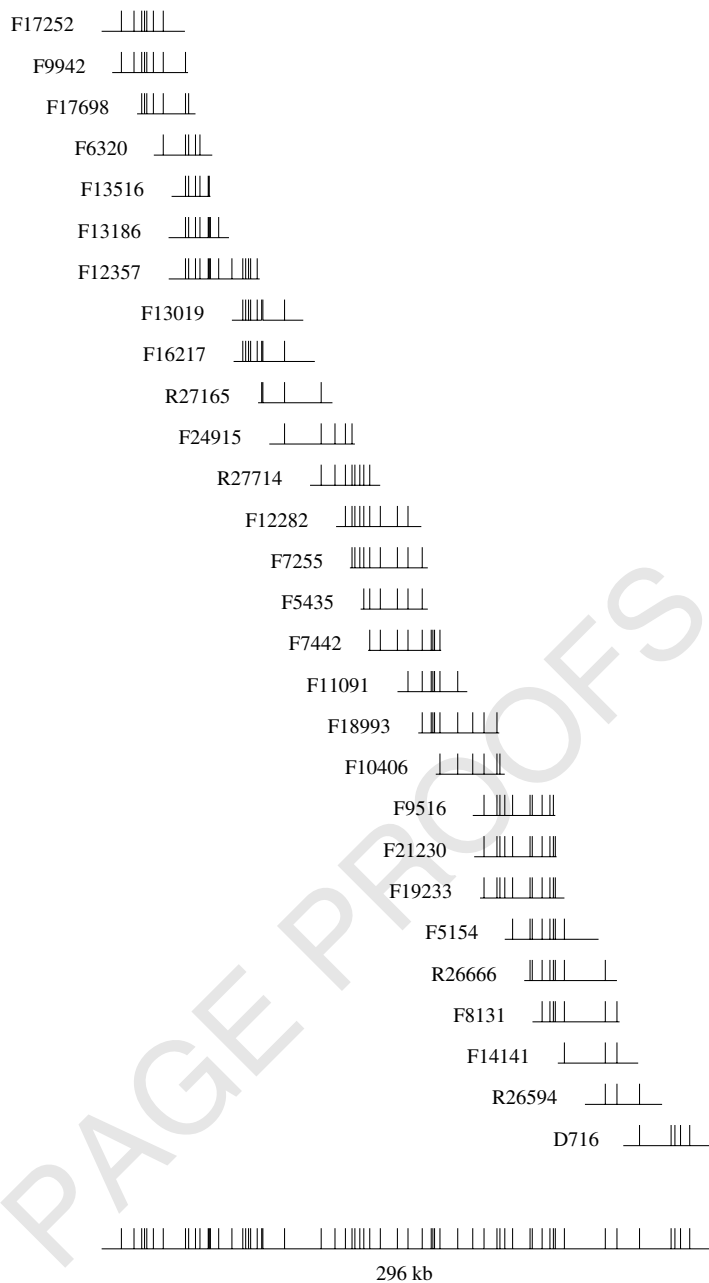


Figure 1.3 Restriction map of cosmid clones. (Source: Lawrence Livermore National Laboratory.)

small molecules, such viral and organelle genomes, or to genomic DNA that has already been cloned.

A major advantage of a restriction map (like that in Figure 1.3) is that accurate lengths are known between sets of reference points. We can preserve an overview of the target

and we can reach a nearly complete map relatively quickly. The disadvantage of most restriction mapping efforts is that they do not produce the DNA in a convenient form. This approach yields maps with more continuity and fewer gaps between fragments than contig maps, but map resolution is lower and may not be useful in finding particular genes. Currently, this approach allows DNA pieces to be located in regions measuring from about 0.1 Mb to 1 Mb.

1.3.4 Restriction Mapping Via Optical Mapping

Optical mapping is a single-molecule approach for the construction of ordered restriction maps developed by Schwartz *et al.* (1993). It uses light microscopy to directly image individual DNA molecules which are bound to specially derivatized surfaces and then cleaved by restriction enzymes. Cleaved fragments retain their original order, and restriction sites are flagged by small, visible gaps. Optical mapping solves the problem of determining fragment order.

The statistical analysis of optical mapping data is relatively new and quite complex, so we do not attempt to summarize it here. A first solution to the problem can be found in Anantharaman *et al.* (1997). These authors take a Bayesian approach, constructing a prior model for an ordered restriction map and a probability model for restriction map data from single molecules. They then approximate the mode of the posterior distribution of the parameters. Orientation, false cuts and sizing errors are among the issues to be dealt with. A second, hierarchical Bayes approach to the same problem using reversible-jump Markov chain Monte Carlo can be found in Lee *et al.* (1998). Finally, a successful application of the method to a whole genome is published in Lai *et al.* (1999).

1.3.5 Ordered Clone Maps

Clones – more fully, cloned DNA fragments – are generated by first breaking a large number of identical chromosomes into fragments, either by physical means such as sonication, compression or irradiation, or by chemical means, typically complete or partial digestion with one or more restriction enzymes. Individual fragments (inserts) of appropriate sizes are then joined to another DNA molecule (the vector) and the result is incorporated into a (host) organism such as *Escherichia coli* or yeast. The average size of the insert varies widely among different hosts and incorporation methods. Yeast artificial chromosomes (YACs) in yeast may have DNA fragments ranging from 100 kb to 1 Mb. Cosmids in *E. coli* may have fragments ranging from 35 to 45 kb, while the now widely used bacterial artificial chromosomes (BACs) have inserts of sizes in the range 100–200 kb. The hosts are separated from each other and allowed to grow into colonies, with the fragment in each host being replicated along with the host's DNA during cell divisions. After enough divisions, each host colony can be harvested, resulting in a library of cloned DNA fragments, where each fragment is present in large enough quantities to permit isolation and purification for subsequent biochemical analyses.

The bottom-up approach to physical mapping is usually carried out by breaking up the DNA molecule of interest, cloning selected fragments, and subjecting each clone to one more experiments – restriction digestions, hybridizations or PCR assays with unique or repetitive probes, or sequencing – to obtain what is called a *fingerprint* of the clone. These fingerprints are then used to solve the combinatorial puzzle of inferring the arrangement of clones along the molecule with the help of these data. The ordered

fragments form contiguous DNA blocks which are called *contigs*. Clone ordering usually begins by comparing clones to each other, in order to determine the strength of evidence that any pair of clones overlap, and it is here that statistical ideas enter.

Currently, clone libraries ordered in this way have inserts which vary in size from a few thousand base-pairs up to 1 Mb. Contig maps thus consist of a linked library of overlapping clones representing a complete chromosomal segment. An advantage of this approach is the accessibility of these clones to other researchers. While useful for finding genes localized to a small region (under 2 Mb), contig maps can be difficult to extend over large stretches of a chromosome because not all regions are clonable.

The statistical analysis of overlap and the estimation of distances will differ somewhat with different fingerprinting techniques. In a hybridization experiment, the fingerprint will be the list of probes that hybridize to the clone; with restriction digestion the fingerprint is a list of observed fragment sizes resulting from the digestion of the clone, while with STS-content mapping (see below) the fingerprint consists of an enumeration of the STSs found to be located on that clone. An example of an STS-based clone map is given in Figure 1.4.

1.3.6 Contig Mapping using Restriction Fragments

One approach, due by Coulson *et al.* (1986), begins with the calculation, for each pair of clones, of the probability of the observed level of matching of fragment sizes, up to a prescribed tolerance, arising by chance, i.e. when the clones do not in fact overlap. This probability – essentially a p -value, but called the probability of coincidence – is used for selecting possibly overlapping clone pairs. Clones are then assembled into contigs using a variety of *ad hoc* rules based on these probabilities. For details, we refer to Sulston *et al.* (1988). A modified version of this procedure is embodied in the program FPC (Soderlund *et al.*, 1997) which was widely used in preparing physical maps for human genome sequencing.

An alternative approach involving a likelihood ratio or Bayes posterior odds was initiated by Michiels *et al.* (1987), and then more fully developed by Branscomb *et al.* (1990). We sketch it now, referring the reader to Nelson and Speed (1994a) and Nelson *et al.* (1997) and the references cited there for fuller details of the trinomial model.

For each clone in a library of N clones, we create DNA fingerprint data by restriction digestion, electrophoresis, and sizing. This consists of a list of fragment lengths. For a particular length l , there are four patterns when we compare two clones: (1,1) if a fragment of length l is observed in both clones; (1,0) if a fragment of length l is observed in the first clone but not the second one; (0,1) if a fragment of length l is observed in the second clone but not the first one; and (0,0) if a fragment of length l is observed in neither clone. The probability of each outcome under a simple randomness model can be approximated by

$$\begin{aligned} p_{00} &= q^{L_1+L_2-\theta}, \\ p_{01} &= q^{L_1}(1 - q^{L_2-\theta}), \\ p_{10} &= q^{L_2}(1 - q^{L_1-\theta}), \\ p_{11} &= 1 - q^{L_1} - q^{L_2} + q^{L_1+L_2-\theta}, \end{aligned}$$

where L_1 and L_2 are the lengths of the two clones, θ is the overlapping amount, $q = e^{-\lambda_l}$, and λ_l is the intensity for fragments of size l . When all the clones are of the same size,

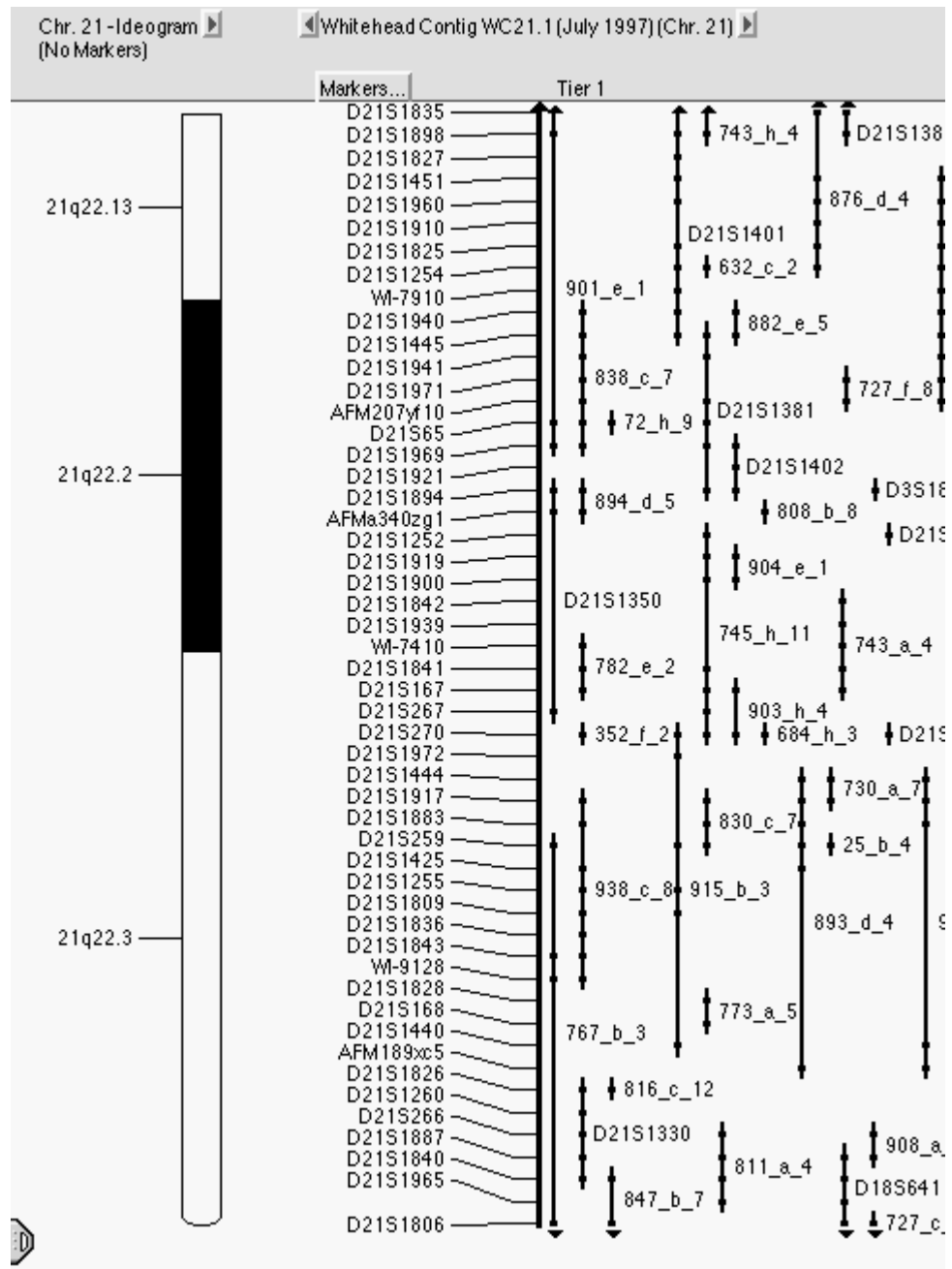


Figure 1.4 STS map of portion of human chromosome 21. (Source: Mapview at www.gdb.org/hugo/chr21/.)

$p_{01} = p_{10}$ and the data can be reduced to a trinomial variable $(n_{00}, n_{01} + n_{10}, n_{11})$. To decide whether two clones overlap, the likelihood ratio test value $L(\theta)/L(\theta = 0)$ can be calculated. Alternatively, with prior information on the θ , posterior odds can be calculated to decide if two clones overlap. The above simple assumptions can be loosened to allow different intensities and errors in fragment size detections. With pairwise similarity measures such as the log posterior odds for overlap, clustering algorithms can be used to build contigs.

1.3.7 Sequence-Tagged Site Maps

An STS is defined by two short sequences, each typically 20–25 bp in length, that have been designed from a region of sequence that appears as a single copy in the human genome. These sequences can act as primers in a PCR assay to score for presence or absence of the site in any DNA sample. One of the aims of the human genome project was to build a high-resolution RH map using STSs as landmarks throughout the human genome. Geneticists would then be able to use the map to isolate genes through nearby landmarks. Sequencers would be able to decide where to prepare clones for the actual sequencing. In addition, the STSs would become part of a common set of markers that can be screened against maps created using different mapping techniques, helping to integrate the efforts of mapping teams world-wide.

For STS content mapping, the data can be summarized as an incidence matrix with N rows corresponding to N clones and M columns corresponding to M STSs. The (i, j) entry is 1 if the j th STS hits the i th clone, and is 0 otherwise. If there are no errors, the problem can be solved by testing whether this incidence matrix has the consecutive 1s property. An incidence matrix has the consecutive 1s property for rows if its columns can be permuted so as to make all the 1s in each row appear consecutively. Booth and Lueker (1976) described linear-time algorithms for determining if a matrix has the consecutive 1s property, and they provided a compact description of all possible consistent permutations in the form of a PQ-tree. Therefore, the problem is completely solvable in linear time if the data are error-free.

However, real data sets are never error-free, and the consecutive 1s property no longer holds. Before we discuss various algorithms in the literature, most of which use combinatorial approaches to optimizing an objective function of orderings, we consider a likelihood-based approach for STS ordering following Yeh (1999). For a pair of STSs (s_i, s_j) , distance D apart, we can count the number of clones retaining both STSs (n_{11}), the first STS but not the second STS (n_{10}), the second STS but not the first STS (n_{01}), and neither STS (n_{00}). Define the set of coretenion probabilities for (s_i, s_j) as $\mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})$, where p_{11} is the probability that a clone contains both s_i and s_j , p_{10} that it retains s_i only, p_{01} that it retains s_j only, and p_{00} that it retains neither. Assuming clones are random line segments of length L on the chromosome (genome) of size G , the coretenion probabilities are

$$p_{11} = l - d,$$

$$p_{10} = d,$$

$$p_{01} = d,$$

$$p_{00} = 1 - l - d,$$

when $d < l$, and

$$\begin{aligned} p_{11} &= 0, \\ p_{10} &= l, \\ p_{01} &= l, \\ p_{00} &= 1 - 2l, \end{aligned}$$

when $d \geq l$, where l is the normalized clone length L/G and d is the normalized STS distance D/G . These probabilities allow us to evaluate the likelihood for each ordering of the STSs. The estimated order is the one that maximizes the overall likelihood. Thus the STS ordering problem is equivalent to a TSP with the STSs as vertices and the pairwise likelihoods as the distances. Using arguments similar to those in marker ordering in genetic mapping (Speed *et al.*, 1992), Yeh (1999) showed that this procedure will recover the true order with probability 1 when the number of clones is large. The objective function here is the same as that discussed by Green and Green (1991), which is the first major paper on this topic, and which describes in outline the widely-used program SEGMAP. However, that program goes considerably further, including the solution of a linear programming problem to find bounds on the distance between any pair of points in the map (STSs or clone ends).

In addition to obtaining distance estimates among the STSs, this likelihood approach is robust when the error rates are not too high. Mott *et al.* (1993) used simulated annealing to minimize the total discrepancy among adjacent STSs, where the discrepancy between two STSs a and b is defined as:

$$d(a, b) = 1 - \frac{\#(\text{clones positive for } a \text{ and } b)}{\#(\text{clones positive for } a \text{ or } b)}.$$

Alizadeh *et al.* (1995) used the TSP algorithms to minimize the total Hamming distance of the clone-probe incidence matrix, which corresponds to the number of gaps of the probe ordering. They also proposed an alternative objective function based on a weighted sum of the number of chimeric clones, the number of false positives, and the number of negatives. Christof *et al.* (1997) formulated the problem as a weighted betweenness problem, assuming the probes are from both ends of all clones. Alizadeh *et al.* (1995) and Nelson *et al.* (1997) described statistical procedures for evaluating overlapping configurations involving more than two clones.

Once STSs are ordered, clones are ordered with respect to the probes by maximizing a measure of fit between the probe data for that clone and the list of ordered probes. Unlike physical maps constructed from restriction maps, the map constructed using STS content mapping would not be tied to a particular set of clones, thus could be used to order any subsequently generated library.

In the early 1990s, considerable effort was put into the generation of clone contig maps, using STS screening. The major achievement of this phase of physical mapping was the publication of a clone contig map of the entire genome, consisting of 33 000 YACs containing fragments with an average size of 0.9 Mb (Cohen *et al.*, 1993).

The combined STS maps now include positions for almost 7000 simple sequence length polymorphisms that have already been mapped onto the genome by genetic means. As a

result, the physical and genetic maps can be directly compared, and the clone contig maps that include STS data can be anchored on to both maps.

1.4 RADIATION HYBRID MAPPING

A radiation hybrid is a rodent cell that contains fragments of chromosomes from a second organism. The technology was first developed in the 1970s by Goss and Harris (1975; 1977) and reintroduced by Cox *et al.* (1990) based on the observation that exposure of human cells to X-rays causes the chromosomes to break up randomly into fragments, and these chromosome fragments can then be propagated if the irradiated cells are fused with nonirradiated hamster or other rodent cells. A routine selection process is used to screen out hamster cells without human chromosome fragments. In its simplest form, a single human chromosome is exposed to a radiation source. For the whole genome radiation mapping (WG-RH), a normal diploid human cell is used as the donor by Walter *et al.* (1994). The WG-RH mapping has the advantage that pieces of many different human chromosomes may be contained in the same hybrid and so a single panel of WG-RHs can be used to map any region of the human genome. Detailed mapping of the entire human genome can be accomplished with fewer than 100 WG-RHs. The resulting panels can be screened for human-specific markers. Data for RH mapping also can be summarized in an incidence matrix like the one for STS content mapping.

Like STS content mapping, the basic premise of RH mapping is that the closer two loci are on the human chromosome, the less likely it is that they will be broken by irradiation. The retention patterns from the various hybrid clones give clues for determining locus order and for estimating the distance between adjacent loci for a given order.

One criterion that quantifies this heuristic is the minimum obligate breaks criterion. For a given order of the loci, we can count the number of changes from 1 to 0 and from 0 to 1. If we sum these changes over all the clones, we get the total number of obligate breaks. The objective is to find the order that minimizes the total number of obligate breaks across all clones. The advantage of the minimum breaks criterion is that it does not depend on any assumptions about how breaks occur and how fragments are retained. Assuming the same retention rate, one can again use arguments similar to those in marker ordering for genetic mapping and show that this criterion is strongly statistically consistent (Lange, 1997, Chapter 11).

1.4.1 Haploid Data

We now turn to probabilistic models for RH mapping. In RH mapping, the distance between sites can be expressed in units (centirays) representing the percentage probability of separation by breakage with a given irradiation dosage. This gives a better measure of physical distance than genetic distance, because the vulnerability to breakage seems to be fairly constant along the whole length of the chromosome. Therefore, in models for RH mapping, the breakage process along the chromosome can be modeled as a Poisson process (Cox *et al.*, 1990). For any two loci, the probability of at least one break θ and the physical distance δ are related by

$$1 - \theta = e^{-\lambda\delta},$$

where the value of λ depends on the irradiation dose. This function is similar to the Haldane map function used in genetic mapping. Note that the parameters δ and λ cannot be separated from the estimation. For RH mapping, in addition to considering breakage, we also need to take retention into account. It is normally assumed that different segments are retained independently; however, different fragments may be allowed to have different retention probabilities.

For two markers, there are four possibilities for a haploid RH: (1,1) when both markers are present; (1,0) when the first marker is present but not the second marker; (0,1) when the second marker is present but not the first one; and (0,0) when neither marker is present. The probabilities for these four patterns are:

$$\begin{aligned} p_{11} &= \theta P_A P_B + (1 - \theta) P_{AB}, \\ p_{10} &= \theta P_A (1 - P_B), \\ p_{01} &= \theta P_B (1 - P_A), \\ p_{00} &= \theta (1 - P_A)(1 - P_B) + (1 - \theta)(1 - P_{AB}), \end{aligned}$$

where P_A , P_B , and P_{AB} are the probabilities of a hybrid retaining a fragment with marker A only, with marker B only, and with both markers A and B , respectively (Cox *et al.*, 1990). In the general case of many markers, Boehnke *et al.* (1991) derived the probability for a hybrid with any retention pattern.

Note that when $P_A = P_B = P_{AB} = r$ in the above equations, the probabilities reduce to

$$\begin{aligned} p_{11} &= \theta r^2 + (1 - \theta)r, \\ p_{10} &= \theta r(1 - r), \\ p_{01} &= \theta r(1 - r), \\ p_{00} &= \theta(1 - r)^2 + (1 - \theta)(1 - r). \end{aligned}$$

Therefore, the probabilities are simply a reparametrization of those we derived when we discussed ordering STSs: simply put $l = 1 - r$ and $d = \theta r(1 - r)$.

1.4.2 Diploid Data

For the WG-RH mapping, two chromosomes instead of one, are involved. For a pair of markers, we have the same four possibilities for an RH. Assuming the same retention rate r for all fragments, the probabilities for the four possibilities are

$$\begin{aligned} p_{11} &= 1 - 2(1 - r)^2 + [(1 - r)(1 - \theta r)]^2, \\ p_{10} &= (1 - r)^2 - [(1 - r)(1 - \theta r)]^2, \\ p_{01} &= (1 - r)^2 - [(1 - r)(1 - \theta r)]^2, \\ p_{00} &= [(1 - r)(1 - \theta r)]^2. \end{aligned}$$

For an arbitrary number of markers, hidden Markov models were used by Lange *et al.* (1995) to calculate the probability of any retention pattern.

When the retention probabilities are allowed to vary for different fragments, the number of parameters involved increases quadratically with the number of markers examined.

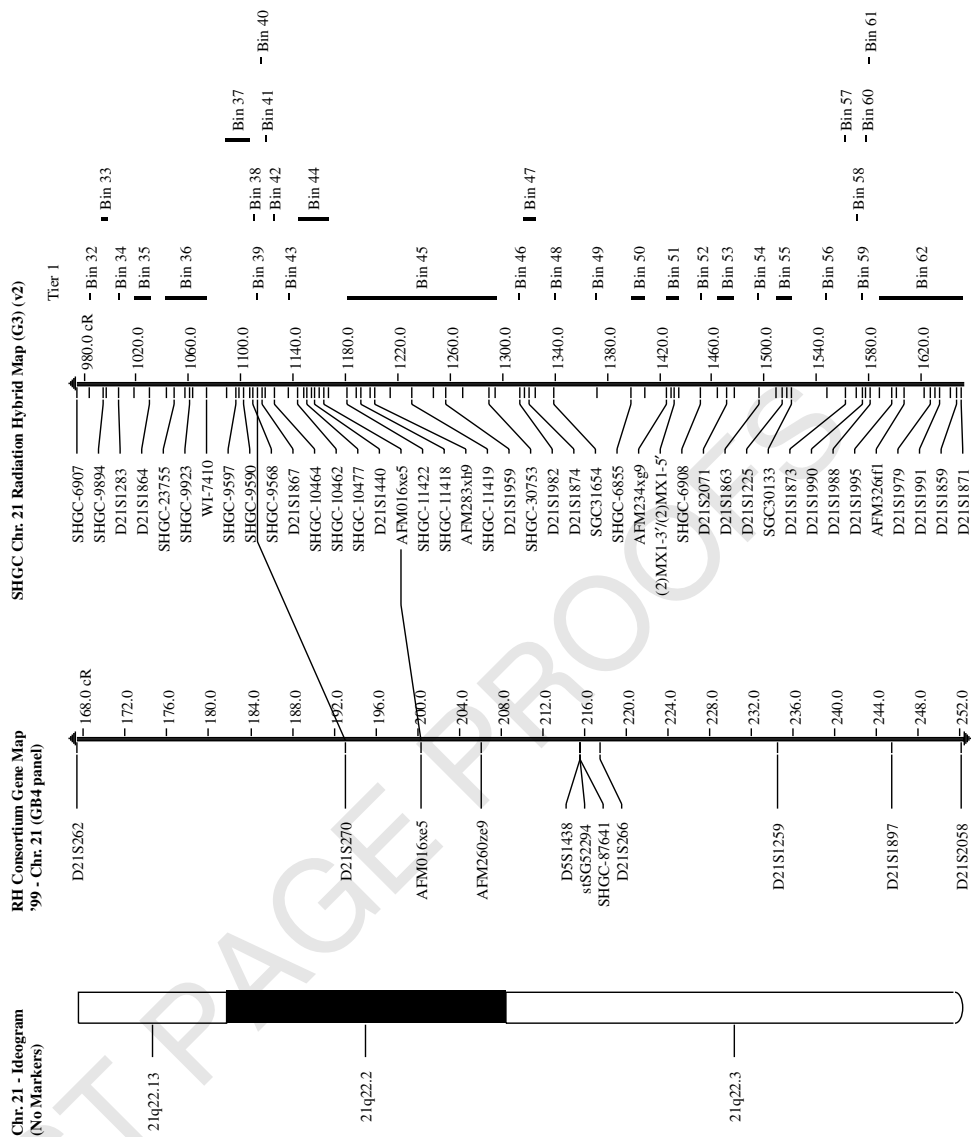


Figure 1.5 RH map of part of human chromosome 21. (Source: Mapview at www.gdb.org/hugo/chr21/.)

Although a number of computational methods can be used for such problems (see Boehnke *et al.*, 1991; Leach and O'Connell, 1995), this raises a serious optimization problem. If the retention rate is assumed to be constant, the calculation can be simplified. Jones (1997) found that adopting simple models generally does not affect the ability to recover the true locus order, but affects the estimation of distances among the loci.

Bayesian methods have also been developed for RH mapping by Lange and Boehnke (1992) and Guerra *et al.* (1992). Tibshirani *et al.* (1999) proposed to maximize a pseudo-likelihood based on information from all marker pairs and then to use multidimensional scaling to provide starting positions for the markers. Lange (1997) has an excellent chapter on RH mapping, and also recommended is the review by Jones (2000). More recent methods include Agarwala *et al.* (2000) and Ben-Dor *et al.* (2000).

There are three human radiation hybrid panels available: Genebridge4 (93 hybrids), Stanford G3 (83 hybrids), and Stanford TNG4 (90 hybrids). These panels differ in radiation dose and retention probabilities. RH panels have also been made for mouse, rat, cow, pig, zebrafish, dog, cat, baboon, and horse. More detailed information can be found at <http://linkage.rockefeller.edu/tara/rhmap>. An example of the RH maps are given in Figure 1.5.

1.5 OTHER PHYSICAL MAPPING APPROACHES

In *directed mapping* (see Palazzolo *et al.*, 1991; Mizukami *et al.*, 1993), random seed clones are selected first, and contigs are extended by anchors generated from contig ends. Then an unmapped clone is selected, and STSs from its ends are constructed and used to find the set of overlapping clones, usually by a PCR assay. The process continues until all clones have been either selected or identified as overlapping some selected clones. Nelson and Speed (1994b) found that a project using a directed strategy makes slower progress in the beginning, but closes the gaps much faster in later stages.

A *double end-sequencing strategy* combines sequencing and mapping by sequencing both ends of subclones and inferring clone overlaps from end-sequence comparisons; see Chen *et al.* (1993) and Roach (1995). A double end-sequencing strategy combined with directed finishing provides an efficient approach to sequencing a large piece of DNA (Yeh, 1999).

High-resolution mapping using FISH uses two or more fluorescent probes to hybridize to chromosomes at a particular stage in the cell cycle. The distance between the fluorescent dots in each cell is measured. The data from FISH experiments consist of a series of distance measurements between two or more probes. Such mapping data are now being linked to more traditional physical mapping data; see Kirsch *et al.* (2000).

Recently 11 pharmaceutical and technology companies, one large scientific trust, and four major academic centers joined efforts to identify and map single nucleotide polymorphisms (SNPs). Their mission is to identify 300 000 SNPs distributed evenly throughout the human genome. The latest release (October 24, 2002) consists of nearly 1.8 million SNPs, all of which have been anchored to the human genome by RH mapping and/or 'in silico' mapping to the genomic working draft. The web site for the SNP consortium is <http://snp.cshl.org>.

1.6 GENE MAPS

Because of the possibilities of having two or more noncontiguous DNA fragments in a single clone, in 1994 the International Radiation Hybrid Mapping Consortium was formed to construct a human gene map in which cDNA-based STS markers from 3'-untranslated regions of cDNAs were physically mapped and then integrated with the genetic map of polymorphic microsatellite markers. The consortium initially reported a map with about 16 000 genes by Schuler *et al.* (1996); a later map constructed by Deloukas *et al.* (1998) contains 30 181 gene-based markers. The resulting map density approached the target of one marker per 100 kb set as the objective for physical mapping at the outset of the human genome project. The GeneMap '98 or Human Transcript Map STSs are derived from transcribed sequences. Finally, we mention the Integrated Molecular Analysis of Genomes and their Expression (IMAGE) Consortium, 'the world's largest public collection of genes'.

1.7 PROGRAMS FOR PHYSICAL MAPPING

For physical mapping, there is SEGMAP, Site-Content Map Assembly Software, at <http://www.genome.washington.edu/UWGC/analysistools/segmap.htm>. FPC (<http://www.sanger.ac.uk/Software/fpc>) is an interactive program for building contigs from fingerprinted clones, where the fingerprint for a clone is a set of restriction fragments. Software for optical mapping can be found at <http://schwartzlab.biotech.wisc.edu/omm/omm.html>. Finally, for radiation hybrid mapping, many computer programs have been developed and the links to their web sites are collected at <http://linkage.rockefeller.edu/tara/rhmap>. Programs for physical mapping are also discussed in Chapters 6 and 7 of Bishop (1998).

Acknowledgments

This work has been supported in part by NIH grant 8R1GM59506A to T.P. Speed, and NIH grants GM59507 and HD36834 and research grant FY98-0752 from the March of Dimes Birth Defects Foundation to H. Zhao.

REFERENCES

- Agarwala, R., Applegate, D.L., Maglott, D., Schuler, G.D. and Schäffer, A.A. (2000). A fast and scalable radiation hybrid map construction and integration strategy. *Genome Research* **10**, 350–364.
- Alizadeh, F., Karp, R.M., Newberg, L.A. and Weissner, D.K. (1995). Physical mapping of chromosomes: a combinatorial problem in molecular biology. *Algorithmica* **13**, 52–76.
- Anantharaman, T.S., Mishra, B. and Schwartz, D.C. (1997). Genomics via optical mapping. 2. Ordered restriction maps. *Journal of Computational Biology* **4**, 91–118.
- Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.

- Baldmin, M. and Chovnick, A. (1967). Autosomal half-tetrad analysis in *Drosophila melanogaster*. *Genetics* **55**, 277–293.
- Bateson, W., Saunders, E.R. and Punnett, R.C. (1905). Experimental studies in the physiology of heredity. *Reports to the Evolution Committee of the Royal Society* **2**, 1–55 and 88–99.
- Beadle, G.W. and Emerson, S. (1935). Further studies of crossing over in attached-X chromosomes of *Drosophila melanogaster*. *Genetics* **20**, 192–206.
- Bell, J. and Haldane, J.B.S. (1937). The linkage between the genes for colour blindness and haemophilia in man. *Proceedings of the Royal Society London B* **123**, 119–150.
- Ben-Dor, A., Chor, B. and Pelleg, D. (2000). RHO – radiation hybrid ordering. *Genome Research* **10**, 365–378.
- Bernstein, F. (1931). Zur Grundlegung der Chromosomentheorie der Vererbung beim Menschen. *Zeitschrift für induktive Abstammungs- und Vererbungslehre* **57**, 113–138.
- Bishop, D.T. and Thompson, E.A. (1988). Linkage information and bias in the presence of interference. *Genetic Epidemiology* **5**, 107–119.
- Bishop, M.J. (1998). *Guide to Human Genome Computing*, 2nd edition. Academic Press, San Diego, CA.
- Boehnke, M., Lange, K. and Cox, D.R. (1991). Statistical methods for multipoint radiation hybrid mapping. *American Journal of Human Genetics* **49**, 1174–1188.
- Booth, K.S. and Lueker, G.S. (1976). Testing for the consecutive 1s property, interval graphs, and graph planarity using PQ-tree algorithm. *Journal of Computer and System Sciences* **13**, 335–379.
- Branscomb, E., Slezak, T., Pae, R., Galas, D., Carrano, A.V. and Waterman, M. (1990). Optimizing restriction fragment fingerprinting methods for ordering large genomic libraries. *Genomics* **8**, 351–366.
- Bridges, C.B. (1935). Salivary chromosome maps. *Journal of Heredity* **26**, 60–64.
- Broman, K.W. and Weber, J.L. (2000). Human crossover interference. *American Journal of Human Genetics* **66**, 1911–1926.
- Broman, K.W., Murray, J.C., Sheffield, V.C., White, R.L. and Weber, J.L. (1998). Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *American Journal of Human Genetics* **63**, 861–869.
- Carter, T.C. and Falconer, D.S. (1951). Stocks for detecting linkage in the mouse and the theory of their design. *Journal of Genetics* **50**, 307–323.
- Chakravarti, A. and Slaugenhaupt, S.A. (1987). Methods for studying recombination on chromosomes that undergo nondisjunction. *Genomics* **1**, 35–42.
- Chakravarti, A., Majumder, P.P., Slaugenhaupt, S.A., Deka, R., Warren, A.C., Surti, U., Ferrell, R.E. and Antonarakis, S.E. (1989). In *Molecular and Cytogenetic Studies of Nondisjunction: Proceedings of the Fifth Annual National Down Syndrome Society Symposium*, T.J. Hassold and C.J. Epstein, eds. Alan R. Liss, New York, pp. 35–42.
- Chen, E.Y., Schlessinger, D. and Kere, J. (1993). Ordered shotgun sequencing: a strategy for integrating mapping and sequencing of YAC clones. *Genomics* **17**, 651–656.
- Christof, T., Jünger, M., Kececioglu, J., Mutzel, P. and Reinelt, G. (1997). A branch-and-cut approach to physical mapping of chromosome-by-unique end-probes. *Journal of Computational Biology* **4**, 433–447.
- Cohen, D., Chumakov, I. and Weissenbach, J. (1993). A first-generation map of the human genome. *Nature* **366**, 698–701.
- Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986). Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences (USA)* **83**, 7821–7825.
- Cox, D.R., Burmeister, M., Price, E.R., Kim, S. and Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**, 245–250.
- Cui, X., Gerwin, J., Navidi, W., Li, H., Kuehn, M. et al. (1992). Gene-centromere linkage mapping by PCR analysis of individual oocytes. *Genomics* **13**, 713–717.

- Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M. and White, R. (1990). Program description – Centre-d'Étude-du-Polymorphisme-Humain (CEPH) – collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577.
- Deloukas, P., Schuler, G.D., Gyapay, G. *et al.* (1998). A physical map of 30000 genes. *Science* **282**, 744–746.
- Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P. *et al.* (1987). A genetic linkage map of the human genome. *Cell* **51**, 319–337.
- Elston, R.C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.
- Eppig, J.T. and Eicher, E.M. (1983). Application of the ovarian teratoma mapping method in the mouse. *Genetics* **103**, 797–812.
- Feingold, E., Brown, A.S. and Sherman, S.L. (2000). Multipoint estimation of genetic maps for human trisomies with one parent or other partial data. *American Journal of Human Genetics* **66**, 958–968.
- Felsenstein, J. (1979). A mathematically tractable family of genetic mapping functions with different amount of interference. *Genetics* **91**, 769–775.
- Fincham, J.R.S., Day, P.R. and Radford, A. (1979). *Fungal Genetics*. University of California Press, Berkeley.
- Fisher, R.A. (1922). The systematic location of genes by means of crossover relations. *American Naturalist* **56**, 406–411.
- Fisher, R.A., Lyon, M.F. and Owen, A.R.G. (1947). The sex chromosome of the house mouse. *Heredity* **1**, 335–365.
- Goldgar, D.E. and Fain, P.R. (1988). Models of multilocus recombination: non-randomness in chiasma number and crossover location. *American Journal of Human Genetics* **43**, 38–45.
- Goldstein, D.R., Zhao, H. and Speed, T.P. (1995). Relative efficiencies of chi-square models of recombination for exclusion mapping and gene ordering. *Genomics* **27**, 265–273.
- Goss, S.J. and Harris, H. (1975). New method for mapping genes in human chromosomes. *Nature* **255**, 680–684.
- Goss, S.J. and Harris, H. (1977). Gene transfer by means of cell fusion II. The mapping of 8 loci on human chromosome 1 by statistical analysis of gene assortment in somatic cell hybrids. *Journal of Cell Science* **25**, 39–57.
- Green, E. and Green, P. (1991). Sequence-tagged sites (STS) content mapping of human chromosomes: theoretical considerations and early experiences. *PCR Methods and Applications* **1**, 77–90.
- Green, M.C. (1981). In *The Mouse in Biomedical Research*, Vol. 1, H.L. Foster, J.D. Small and J.G. Fox, eds. Academic Press, New York, pp. 105–117.
- Green, P. (1988). Rapid construction of multilocus genetic linkage maps. I. Maximum likelihood estimation. Draft manuscript.
- Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C. and Gelbart, W.M. (1996). *An Introduction to Genetic Analysis*, 6th edition. W.H. Freeman, New York.
- Guerra, R., McPeck, M.S., Speed, T.P. and Stewart, P.M. (1992). A Bayesian analysis for mapping from radiation hybrid data. *Cytogenetics and Cell Genetics* **59**, 104–106.
- Haldane, J.B.S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8**, 299–309.
- Haldane, J.B.S. (1931). The cytological basis of genetical interference. *Cytologia* **3**, 54–65.
- Irwin, M., Cox, N. and Kong, A. (1994). Sequential imputation for multilocus linkage analysis. *Proceedings of the National Academy of Sciences (USA)* **91**, 11 684–11 688.
- Johnson, D.S. (1990). In *Automata, Languages, and Programming*, M.S. Paterson, ed. Springer-Verlag, Berlin, pp. 446–461.
- Jones, H.B. (1997). Estimating physical distance using radiation hybrid mapping data. *Genomics* **43**, 258–266.

- Jones, H.B. (2000). A review of statistical methods for genome mapping. *International Statistical Review* **68**, 5–21.
- Karlin, S. and Liberman, U. (1978). Classification and comparisons of multilocus recombination distributions. *Proceedings of the National Academy of Sciences (USA)* **75**, 6332–6336.
- Karlin, S. and Liberman, U. (1979). A natural class of multilocus recombination processes and related measure of crossover interference. *Advances in Applied Probability* **11**, 479–501.
- King, J.S. and Mortimer, R.K. (1990). A polymerization model of chiasma interference and corresponding computer simulation. *Genetics* **126**, 1127–1138.
- Kirsch, I.R., Green, E.D., Yonescu, R., Strausberg, R., Carter, N., Bentley, D., Leversha, M.A., Dunham, I., Braden, V.V., Hilgenfeld, E., Schuler, G., Lash, A.E., Shen, G.L., Martelli, M., Kuehl, W.M., Klausner, R.D. and Ried, T. (2000). A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome. *Nature Genetics* **24**, 339–340.
- Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S.T., Frigge, M.L., Thorgeirsson, T.E., Gulcher, J.R. and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nature Genetics* **31**, 241–247.
- Kosambi, D.D. (1944). The estimation of the map distance from recombination values. *Annals of Eugenics* **12**, 172–175.
- Kruglyak, L. and Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439–454.
- Kruglyak, L., Daly, M.J. and Lander, E.S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *American Journal of Human Genetics* **56**, 519–527.
- Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **58**, 1347–1363.
- Lai, Z.W., Jing, J.P., Aston, C., Clarke, V. *et al.* (1999). A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nature Genetics* **23**, 309–313.
- Lamb, N.E., Feingold, E. and Sherman, S.L. (1997). Estimating meiotic exchange patterns from recombination data: an application to humans. *Genetics* **146**, 1011–1017.
- Lander, E.S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences (USA)* **84**, 2363–2367.
- Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E. and Newburg, L. (1987). MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**, 174–181.
- Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. Springer-Verlag, New York.
- Lange, K. and Boehnke, M. (1992). Bayesian methods and optimal experimental design for gene mapping by radiation hybrids. *Annals of Human Genetics* **56**, 119–144.
- Lange, K., Boehnke, M., Cox, D.R. and Lunetta, K.L. (1995). Statistical analysis for polyploid radiation hybrid mapping. *Genome Research* **5**, 136–150.
- Lange, K., Zhao, H. and Speed, T.P. (1997). The Poisson-skip model of crossing-over. *Annals of Applied Probability* **7**, 299–313.
- Lathrop, G.M., Lalouel, J.-M., Julier, C. and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences (USA)* **81**, 3443–3446.
- Lathrop, G.M., Lalouel, J.-M., Julier, C. and Ott, J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American Journal of Human Genetics* **37**, 482–498.
- Lathrop, M., Nakamura, Y., Cartwright, P., O'Connell, P., Leppert, M., Jones, C., Tateishi, H., Bragg, T., Lalouel, J.M. and White, R. (1988). A primary genetic map of markers for human chromosome 10. *Genomics* **2**, 157–164.

- Leach, R.J. and O'Connell, P. (1995). Mapping of mammalian genome with radiation (Goss and Harris) hybrids. *Advances in Genetics* **33**, 63–99.
- Lee, J.K., Dancik, V. and Waterman, M.S. (1998). Estimation for restriction sites observed by optical mapping using reversible-jump Markov chain Monte Carlo. *Journal of Computational Biology* **5**, 505–515.
- Li, J.M., Sherman, S.L., Lamb, N. and Zhao, H.Y. (2001). Multipoint genetic mapping with trisomy data. *American Journal of Human Genetics* **69**, 1255–1265.
- Lieberman, U. and Karlin, S. (1984). Theoretical models of genetic map functions. *Theoretical Population Biology* **25**, 331–346.
- Lin, S. (1996). In *Genetic Mapping and DNA Sequencing*, T.P. Speed and M.S. Waterman, eds. Springer-Verlag, New York, pp. 15–38.
- Lin, S. and Speed, T.P. (1996). Incorporating crossover interference into pedigree analysis using the chi-square model. *Human Heredity* **46**, 315–322.
- Lin, S. and Wijsman, E. (1994). Monte Carlo multipoint linkage analysis. *American Journal of Human Genetics* **55**, A40.
- Ling, S. (2000). Constructing genetic maps for outbred experimental crosses. Ph.D. dissertation, UC Berkeley.
- Ludwig, W. (1934). Über numerische Beziehungen der Crossover-Werte untereinander. *Zeitschrift für induktive Abstammungs- und Vererbungslehre* **67**, 58–95.
- Marinov, M., Matisse, T.C., Lathrop, G.M. and Weeks, D.E. (1999). A comparison of two algorithms, MultiMap and Gene Mapping System, for automated construction of genetic linkage maps. *Genetic Epidemiology* **17**, S649–S654.
- Mather, K. (1933). The relationship between chiasmata and crossing-over in diploid and triploid *Drosophila melanogaster*. *Journal of Genetics* **27**, 243–259.
- Mather, K. (1935). Reductional and equational separation of the chromosomes in bivalents and multivalents. *Journal of Genetics* **30**, 53–78.
- Matisse, T.C., Perlin, M. and Chakravarti, A. (1994). Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome map. *Nature Genetics* **6**, 384–390.
- McPeck, M.S. and Speed, T.P. (1995). Modeling interference in genetic recombination. *Genetics* **139**, 1031–1044.
- Mendel, G. (1866). Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereines in Brünn* **4**, 3–47.
- Michiels, F., Craig, A.G., Zehetner, G., Smith, G.P. and Lehrach, H. (1987). Molecular approaches to genome analysis: a strategy for the construction of ordered overlapping clone libraries. *CABIOS* **3**, 203–210.
- Mizukami, T., Chang, W.I., Garkavstev, I., Kaplan, N., Lombardi, D., Matsumoto, T., Niwa, O., Kounosu, A., Yanagida, M., Marr, T.G. and Beach, D. (1993). A 13kb resolution cosmid map of the 14Mb fission yeast genome by nonrandom sequence-tagged site mapping. *Cell* **73**, 121–132.
- Mohr, J. (1954). *A Study of Linkage in Man*. Munksgaard, Copenhagen.
- Morgan, T.H. (1911). An attempt to analyze the constitution of the chromosomes on the basis of sex limited inheritance in *Drosophila*. *Journal of Experimental Zoology* **11**, 365–414.
- Morton, N.E., Keats, B.J., Jacobs, P.A., Hassold, T., Pettay, D. *et al.* (1990). A centromere map of the X chromosome from trisomies of maternal origin. *Annals of Human Genetics* **54**, 39–47.
- Morton, N.E., Collins, A., Lawrence, S. and Shields, D.C. (1992). Algorithms for a location database. *Annals of Human Genetics* **56**, 223–232.
- Mott, R., Grigoriev, A., Maier, E., Hoheisel, J. and Lehrach, H. (1993). Algorithms and software tools for ordering clone libraries: application to the mapping of the genome of *Schizosaccharomyces pombe*. *Nucleic Acids Research* **21**, 1965–1974.
- Muller, H.J. (1916). The mechanism of crossing over. *American Naturalist* **50**, 193–221; 284–305; 350–366; 421–434.
- Nelson, D.O. and Speed, T.P. (1994a). Statistical issues in constructing high resolution physical maps. *Statistical Science* **9**, 334–354.

- Nelson, D.O. and Speed, T.P. (1994b). Predicting progress in directed mapping projects. *Genomics* **24**, 41–52.
- Nelson, D.O., Speed, T.P. and Yu, B. (1997). The limits of random fingerprinting. *Genomics* **40**, 1–12.
- O'Connell, J.R. and Weeks, D.E. (1995). The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recording and fuzzy inheritance. *Nature Genetics* **11**, 402–408.
- Olson, M.V., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R. and Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences (USA)* **83**, 7826–7830.
- Ott, J. (1999). *Analysis of Human Genetic Linkage*, 3rd edition. Johns Hopkins University Press, Baltimore, MD.
- Palazzolo, M.J., Sawyer, S.A., Martin, C.H., Smoller, D.A. and Hartl, D.L. (1991). Optimized strategies for sequence-tagged-site selection in genome mapping. *Proceedings of National Academy of Sciences (USA)* **88**, 8034–8038.
- Perkins, D.D. (1949). Biochemical mutants in the smut fungus *Ustilago maydis*. *Genetics* **34**, 607–626.
- Rao, D.C., Morton, N.E., Lindsten, J., Hulten, M. and Yee, S. (1977). A mapping function for man. *Human Heredity* **27**, 99–104.
- Risch, N. (1991). A note on multiple testing procedures in linkage analysis. *American Journal of Human Genetics* **48**, 1058–1064.
- Risch, N. and Lange, K. (1979). An alternative model of recombination and interference. *Annals of Human Genetics* **43**, 61–70.
- Risch, N. and Lange, K. (1983). Statistical analysis of multilocus recombination. *Biometrics* **39**, 949–963.
- Roach, J. (1995). Random subcloning. *Genome Research* **5**, 464–473.
- Robinson, W.P., Bernascoli, F., Mutirangura, A., Ledbetter, D.H., Langlois, S., Malcolm, S., Morris, M.A. and Schinzel, A.A. (1993). Nondisjunction of chromosome 5: origin and recombination. *American Journal of Human Genetics* **53**, 740–751.
- Saura, A.O., Saura, A.J. and Sorsa, V. (1997). Electron micrographs maps of *Drosophila melanogaster* polytene chromosomes. <http://www.helsinki.fi/~saura/EM/>.
- Schuler, G.D., Boguski, M.S., Stewart, E.A. et al. (1996). A gene map of the human genome. *Science* **274**, 540–546.
- Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J. and Wang, Y.-K. (1993). Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. *Science* **262**, 110–114.
- Shahar, S. and Morton, N.E. (1986). Origin of teratomas and twins. *Human Genetics* **74**, 215–218.
- Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotype analysis, location scores, and marker sharing statistics. *American Journal of Human Genetics* **58**, 1323–1337.
- Soderlund, C., Longden, I. and Mott, R. (1997). FPC: a system for building contigs from restriction fingerprinted clones. *CABIOS* **13**, 523–535.
- Speed, T.P., McPeck, M.S. and Evans, S.N. (1992). Robustness of the no-interference model for ordering genetic markers. *Proceedings of the National Academy of Sciences (USA)* **89**, 3103–3106.
- Stahl, F.W. (1979). *Genetic Recombination: Thinking about It in Phage and Fungi*. Freeman, San Francisco.
- Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *The Plant Journal* **3**, 739–744.
- Sturt, E. (1976). A mapping function for human chromosomes. *Annals of Human Genetics* **40**, 147–163.
- Sturtevant, A.H. (1913). The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43–59.

- Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T. and Coulson, A. (1988). Software for genome mapping by fingerprinting techniques. *CABIOS* **4**, 125–132.
- Terwilliger, J.D. and Ott, J. (1994). *Handbook of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, MD.
- Thompson, E.A. (1984). Information gain in joint linkage analysis. *IMA Journal of Mathematics Applied in Medicine and Biology* **1**, 31–49.
- Thompson, E.A. (1994). Monte Carlo likelihood in genetic analysis. In *Probability, Statistics, Optimization: a tribute to Peter Whittle*, F.P. Kelley, ed. Wiley, New York, pp. 281–293.
- Tibshirani, R., Lazzaroni, L., Hastie, T., Olshen, A. and Cox, D.R. (1999). The global pairwise approach to radiation hybrid mapping. Technical Report 201, Division of Biostatistics, Stanford University.
- Trask, B.J. (1998). In *Mapping Genomes. Genome Analysis: A Laboratory Manual Series, vol. 4*, B. Birren, E.D. Green, P. Hieter, S. Klapholz, R.M. Myers, H. Riethman, and J. Roskams, eds. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 303–413.
- Vogel, F. and Motulsky, A.G. (1997). *Human Genetics: Problems and Approaches*, 3rd edition. Springer-Verlag, Berlin.
- Walter, M.A., Spillett, D.J., Thomas, P. and Goodfellow, P.N. (1994). A method for constructing radiation hybrid maps of whole genomes. *Nature Genetics* **7**, 22–28.
- Waterman, M.S. (1995). *Introduction to Computational Biology: Maps Sequences and Genomes*. Chapman & Hall, London.
- Weeks, D.E. (1991). In *Advanced Techniques in Chromosome Research*, K.W. Adolph, ed. Marcel Dekker, New York, pp. 297–330.
- Weinstein, A. (1936). The theory of multiple-strand crossing over. *Genetics* **21**, 155–199.
- Whitehouse, H.L.K. (1957). Mapping chromosome centromeres from tetratype frequencies. *Journal of Genetics* **55**, 348–360.
- Wu, S.S., Wu, R.L., Ma, C.X., Zeng, Z.-B., Yang, M.C.K. and Casella, G. (2001). A multivalent pairing model of linkage analysis in autotetraploids. *Genetics* **159**, 1339–1350.
- Yeh, R.-F. (1999). Statistical issues in genomic mapping and sequencing. Ph.D. dissertation, UC Berkeley.
- Zhao, H. and Speed, T.P. (1996). On genetic map functions. *Genetics* **142**, 1369–1377.
- Zhao, H. and Speed, T.P. (1998a). Statistical analysis of half-tetrads. *Genetics* **150**, 473–485.
- Zhao, H. and Speed, T.P. (1998b). Statistical analysis of ordered tetrads. *Genetics* **150**, 459–472.
- Zhao, H., McPeck, M.S. and Speed, T.P. (1995a). Statistical analysis of chromatid interference. *Genetics* **139**, 1057–1065.
- Zhao, H., Speed, T.P. and McPeck, M.S. (1995b). Statistical analysis of crossover interference using the chi-square model. *Genetics* **139**, 1045–1056.
- Zhao, H.Y., Li, J.M. and Robinson, W.P. (2000). Multipoint genetic mapping with uniparental disomy data. *American Journals of Human Genetics* **67**, 851–861.

KEYWORDS: genetic maps, linkage maps, physical maps, gene maps, radiation hybrids

FIRST PAGE PROOFS

QUERIES TO BE ANSWERED BY AUTHOR (SEE MARGINAL MARKS)

IMPORTANT NOTE: Please mark your corrections and answers to these queries directly onto the proof at the relevant place. Do NOT mark your corrections on this query sheet.

Query No.	Query
TS1	Please clarify if the term 'chromosome-by' should be retained as such or if it should be changed to 'chromosome by'.

FIRST PAGE PROOFS