Recombination Within a Subclass of Restriction Fragment Length Polymorphisms May Help Link Classical and Molecular Genetics

Richard B. Meagher, Michael D. McLean and Jonathan Arnold

Department of Genetics, University of Georgia, Athens, Georgia 30602

Manuscript received March 21,1988
Accepted July 14, 1988

ABSTRACT

Restriction fragment length polymorphisms (RFLPs) are being used to construct complete linkage maps for many eukaryotic genomes. These RFLP maps can be used to predict the inheritance of important phenotypic loci and will assist in the molecular cloning of linked gene(s) which affect phenotypes of scientific, medical and agronomic importance. However, genetic linkage implies very little about the actual physical distances between loci. An assay is described which uses genetic recombinants to measure physical distance from a DNA probe to linked phenotypic loci. We have defined the subset of all RFLPs which have polymorphic restriction sites at both ends as class II RFLPs. The frequency of class II RFLPs is computed as a function of sequence divergence and total RFLP frequency for highly divergent genomes. Useful frequencies exist between organisms which differ by more than 7% in DNA sequence. Recombination within class II RFLPs will produce fragments of novel sizes which can be assayed by pulsed field electrophoresis to estimate physical distance in kilobase pairs between linked RFLP and phenotypic loci. This proposed assay should have particular applications to crop plants where highly divergent and polymorphic species are often genetically compatible and thus, where class II RFLPs will be most frequent.

RESTRICTION fragment length polymorphisms (RFLPs) are being used to develop complete linkage maps of complex eukaryotic genomes (DONIS-KELLER *et al.* 1987; BERNATZKY and TANKSLEY 1986a; HELENTJARIS *et al.* 1986). RFLP maps identify DNA probes which can be used to predict the inheritance of genetic loci for important phenotypes such as genetic diseases and useful agricultural traits (LANDER and BOTSTEIN 1986, 1987; NIENHUIS *et al.* 1987). The phenotypes for many of these traits are difficult to score directly in genetic experiments, however closely linked RFLP markers can be used to monitor any portion of a genome in simple assays. Another assumed benefit is that the RFLP probes will also be used to isolate clones of the linked genes which encode important phenotypes. However, there is no way to know if a centimorgan, the unit of recombination which measures the genetic distance between loci, represents 10 kb or 100,000 kb of DNA in the specific region of the genome being examined. It would be scientifically naive to begin a chromosome walk with a linked DNA probe in an attempt to clone to a phenotypic locus based strictly on genetic linkage data and without extensive molecular information about the genome being examined including an estimate of the actual physical distance between loci. Walking (STEINMETZ, STEPHAN and LINDAHL 1986) or even jumping (POUSTKA *et al.* 1987) a few thousand kilobase pairs is very tedious in any large eukaryotic genome but it will be even more difficult in large plant genomes which can have haploid DNA contents from 2 to 126 pg (BENNETT, SMITH and HESLOP-HARRISON 1982) of which greater than 60% is repetitive sequence (FLAVELL 1982). The gene at the phenotypic locus might be isolated on a very large (*e.g.*, 1000 kb) fragment by preparative pulsed field gel electrophoresis (GARDINER, LAAS and PATTERSON 1986; MICHIELS, BURMEISTER and LEHRACH 1987), or cloned in an artificial yeast chromosome vector (BURKE, CARLE and OLSON 1987), or isolated via chromosome mediated gene transfer (PORTEOUS 1987). However, these methods also depend on knowing that the RFLP and phenotypic loci are both physically linked on the same large DNA fragment. Class II RFLPs, described herein, used in combination with classical genetics and pulsed field gel electrophoresis (PFGE) could determine the physical distance between a DNA probe and a genetically linked locus which affects an important phenotype. Once the physical distance between a DNA probe and a phenotypic marker has been estimated, a strategy for isolating clones of the desired gene can be designed. The relationship of physical distance to genetic distance has been examined indirectly and directly in a limited number of eukaryotic systems. Some of these data are discussed below.

PHYSICAL *VS.* GENETIC DISTANCES

Recombination frequencies have been used to determine the genetic distances between genes in viral,

prokaryotic and eukaryotic genomes. A direct relationship between physical distance and frequency of recombination is assumed when constructing these maps. This correlation is well established in *Escherichia coli* (SMITH *et al.* 1987). However, it is incorrect to assume that the ratio of physical distance to crossover units is constant in all regions of a genome, or that it is similar between the genomes of different organisms.

Indirect estimates of the relationship between physical and genetic distances: Indirect estimates of the average kb/cM ratio can be made for different organisms by dividing the haploid DNA content reported for the organism in kilobase pairs by the total genetic map lengths defined by recombination frequencies or the number of chiasma per genome. FINCHAM (1983) estimated these average ratios for yeast, Drosophila and maize to be 10 kb/cM, 570 kb/cM, and 4000 kb/cM, respectively. The 400-fold increase in the number of kilobase pairs per map unit parallels an approximate 300-fold increase in haploid genome size when comparing yeast to maize. A similar calculation can be made for the human genome: an estimate of 1000 kb/cM is obtained assuming a haploid genome size of 3×10^9 bp and the total genetic map length of about 3000 cM (DONIS-KELLER *et al.* 1987). This estimate will probably be within a factor of two or three for most mammalian genomes, which vary little in size. These calculations also suggests the daunting possibility that in the largest angiosperm genomes, such as those found in the Liliaceae, the average ratio could be as much as 20,000 kb/cM.

Genomes with small numbers of chromosomes have less total recombination units. Thus Drosophila, which has a genome size only ten times that of yeast but has only four chromosomes, has a large kb/cM ratio of 57 times that of yeast. REES and DURRANT (1986) calculate the number of chiasma per pg (chiasma frequency) as a function of total (haploid) DNA amount for 20 plant species. Within a chromosome complement, physical chromosome length is proportional to the number of chiasma (*i.e.*, this implies kb/cM ratios increase directly with chromosome length within a species). Comparisons among three angiosperm genera show that an increase in total DNA content correlates with a decrease in the chiasma frequency (*i.e.*, this implies kb/cM ratios will increase with an increase in total DNA content when comparing distant organisms). A generalization of these data is that there are usually 50 to 300 map units per chromosome, no matter what the chromosome size. This suggests that the average kb/cM ratio will be first, a direct function of genome and chromosome sizes, and second, an inverse function of the total number of chromosomes.

Direct measurements of physical distance: The ratio of the actual physical distance (kb), measured directly by agarose gel electrophoresis, to the genetic distance (cM) has been made in only a few eukaryotic systems and for only a few pairs of loci. In yeast, ratios of approximately 10 kb/cM were measured for a number of loci, but this ratio holds only for loci that are distant from the centromere (CLARKE and CARBON 1980; CLARKE *et al.* 1986; NAKASEKO *et al.* 1986). At the *white*, *rosy* and *ultrabithorax* regions of the Drosophila genome, this ratio was determined to be approximately 350, 900 and 1800 kb/cM, respectively (BENDER *et al.* 1983; ZACHAR and BINGHAM 1982). The similarity between these values and the indirect estimates made by FINCHAM (1983) for yeast and Drosophila is reasonable. However, assuming that the average kb/cM ratio determined for a species applies to all loci could result in tremendous errors in estimating physical distance between many loci. In yeast, for example, there are large variations in this ratio for different pairs of loci and especially due to crossover suppression near the centromeres. There is almost no measurable recombination occurring over a 60-kb region near the centromere of chromosome *II* (CLARKE *et al.* 1986; NAKASEKO *et al.* 1986). Major histocompatibility complexes (MHC) are hot spots for recombination in eukaryotes. Estimates of approximately 300 to 600 kb/cM can be made for the region between the *K* and *I-E* MHC in hybrids between *Mus musculus castaneus* and varieties of laboratory mice (STEINMETZ, STEPHEN and LINDAHL 1986). This ratio is close to that calculated above for most mammals. However, in crosses between different strains of laboratory mice, which lack these recombination hot spots within the MHC, the ratio can be estimated to be 30,000 kb/cM. The *bronze* (*Bz*) locus in maize has 100-fold higher levels of recombination as compared with regions between the flanking marker loci (DOONER *et al.* 1985). DOONER (1986) estimates 14 kb/cM within *Bz* as compared to the average value of 3000 kb/cM estimated above for the entire maize genome. Therefore, it should not be assumed that indirect estimations of kb/cM ratios can be applied to specific pairs of loci due to local sequences which affect recombination frequencies. Many sequence and structural components of chromosomes are known to affect recombination frequencies (LUCCHESI and SUZUKI 1968) including inversions, translocations, knobbed and compound chromosomes, aneuploidy, repetitive sequences and the degree of sequence divergence between recombining genomes. Even the sex of the parent in which recombination takes place can have a strong effect on recombination frequencies (DONIS-KELLER *et al.* 1987).

Planning the isolation of a phenotypic locus from a large genome requires a direct measurement of physical distance from a RFLP locus. In making the direct measurements cited above, DNA probes were available for loci at both ends of the linkage relationship. Obviously this cannot be done when estimating the distance to a novel phenotypic marker. The recom-

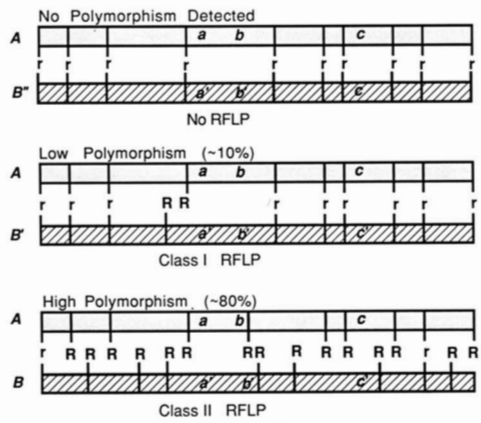

FIGURE 1.—Comparison of the restriction site divergence which might exist between a hypothetical chromosome A with divergent homologous chromosomes B″, B′ and B. Conserved restriction endonuclease sites are indicated by **r**; unique sites found only in one homolog by ***R***; a sequence for which there is a DNA probe by [*a*]; linked phenotypic markers by [*b*] and [*c*]. The percent polymorphism indicated is an estimate which can be obtained from the approximate percent of all the restriction fragments measured which gave RFLPs after probing a large number of different restriction endonucleases digests of the two genomes with [*a*]. A class I RFLP (polymorphic at one end only) and a class II RFLP (polymorphic at both ends) are indicated.

binant class II RFLPs we define will allow a direct estimate of distance to be made with a probe for only one member of the pair of linked loci.

CLASS II RFLPs

Definition of class I and class II RFLPs: Figure 1 shows the comparisons of a hypothetical chromosome A to the divergent hypothetical homologs B″, B′, and B. These three pairs of chromosomes have different degrees of sequence divergence and thus different numbers of RFLPs. The first pair, A and B″, represent two genomes with small sequence divergence. After digesting A and B″ with a large number of restriction endonucleases and probing with the [*a*] locus probe, no RFLPs are detected between A and B″. Between the second hypothetical pair of chromosomes, A and B′, slightly more sequence divergence has occurred (Figure 1). After examining a large number of restriction endonuclease digestions, only 10% of the digestions examined in the two contrasting genomes yield fragments which are polymorphic when probed with [*a*]. Thus for the A/B′ pair it is likely that the RFLP observed with any one restriction endonuclease will have a single restriction site difference at one end of the polymorphic fragments. We will define two polymorphic fragments which differ at only one end as a class I RFLP. The frequency of polymorphisms is too low to expect a significant number of polymorphic fragments to be generated due to restriction site changes at both ends of the fragments. Between the third hypothetical pair of chromosomes, A and B, great sequence divergence has occurred (Figure 1). If 80% of the restriction fragments which hybridize to the [*a*] probe are polymorphic after examining a large number of restriction endonuclease digestions, then it is likely that the RFLP measured with the [*a*] probe for any one particular restriction endonuclease is polymorphic for sites at both ends of the fragments. We define two polymorphic fragments with restriction site differences at both ends as a class II RFLP. Since the percentage of total RFLPs is a measure of sequence divergence between two chromosomes, then higher frequencies of total RFLPs imply higher frequencies of class II RFLPs. How the frequency of class II RFLPs relates to sequence divergence will be discussed below.

Recombination within class II RFLPs generates novel fragment sizes: In a hybrid organism containing the A and B homologs, when each homolog contains hypothetical loci [*a*] and [*b*] within a single restriction fragment of a class II RFLP, a recombination event between the probe locus [*a*] and the phenotypic marker [*b*] will yield recombinant fragments of novel sizes which are different from either of the two parental fragments. (Examine the recombinant A* and B* chromosomes in Figure 2A and 2B.) In other words, a novel RFLP pattern will be generated when recombination occurs within a class II RFLP. The polymorphic fragments of a class II RFLP can be related in two possible ways. First, one fragment can be contained entirely within another with the larger fragment overlapping both ends of the smaller one (Figure 2A). In this case, when one parental class II fragment is an internal subset of the other, recombination should generate two new fragments of intermediate sizes between the two parental fragment sizes (Figure 2A, A* and B*). The sizes of the novel recombinant fragments generated depends upon the sizes of the two parental fragments and upon the position of the smaller fragment relative to the overlapping ends of the larger fragment. Second, the two polymorphic fragments can be asymmetrically positioned relative to one another (Figure 2B). In this case, when the ends of the parental fragments are staggered relative to each other, recombination should generate one fragment larger and one fragment smaller than the two parental fragments (Figure 2B, A* and B*). SENECOFF and COX (1986) assayed for the novel fragments generated *in vitro* by recombination of class II fragments containing the site-specific recombination site of the yeast 2-μm plasmid. In this specialized case, recombination between two fragments of the same size, which are asymmetric relative to each other in sequence, yields one fragment larger and one fragment smaller than either of the parental fragments as shown in the general model (Figure 2B).

It is instructive to consider examples where [*a*] and [*b*] are not contained on the same class II RFLP and where [*a*] and [*b*] are both on the fragments of a class

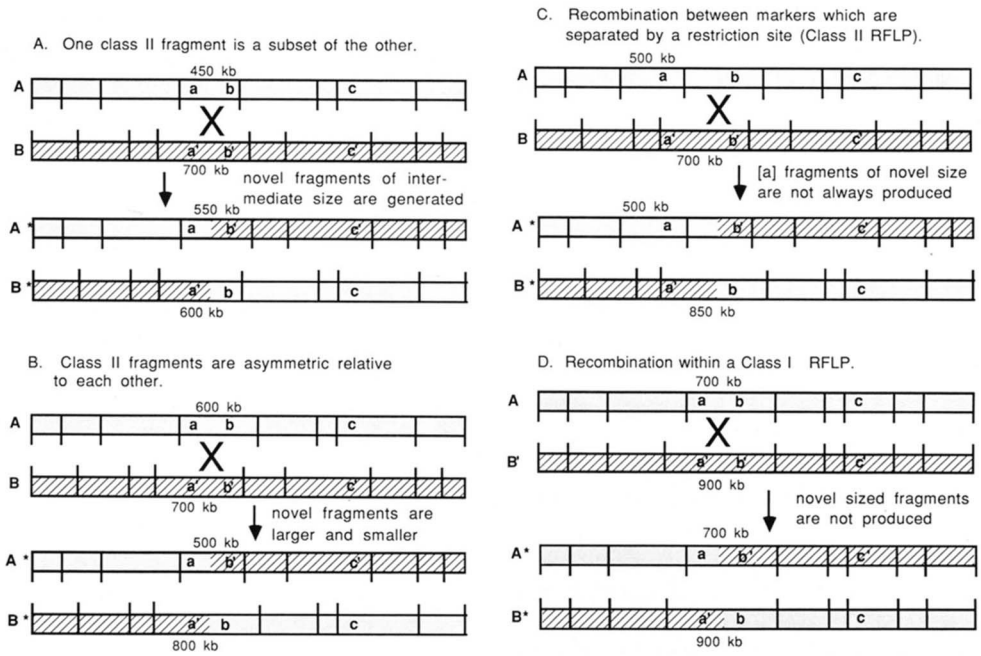

FIGURE 2.—Recombination between two markers contained within the same fragments of a class II RFLP produces novel fragments differing in size from either of the parental RFLP fragments. A, When one class II parental fragment is a subset of the other, recombination between [*a*] and [*b*] produces novel fragments intermediate in size between the parental fragments. B, When the two parental fragments are staggered or asymmetric relative to each other, recombination between [*a*] and [*b*] produces one novel fragment larger and one novel fragment smaller than either of the parental fragments. C, Recombination between [*a*] and [*b*], which are separated by a restriction site in one parental chromosome may not produce a novel sized fragment which will be detected with the [*a*] probe. D, Recombination between [*a*] and [*b*] contained within fragments which are part of a class I RFLP does not produce fragments of novel sizes. In panels A–D, chromosomes marked A* and B* are the recombinant chromosomes. The sequences originally contained on the A chromosome are in gray while those from the B chromosome are indicated by cross-hatching. A sequence for which we have a DNA probe is indicated by [*a*] and the linked phenotypic markers by [*b*] and [*c*]. Sizes of the hypothetical fragments detected by the [*a*] probe before and after recombination are indicated in kilobase (kb) pairs.

I RFLP. When the phenotypic marker [*b*] lies outside the fragment probed with [*a*], in even one of the parental genomes, recombination between [*a*] and [*b*] will not always produce a novel sized fragment after probing with [*a*] (Figure 2C). In this case one new fragment was generated which would be detected with the [*a*] probe and one parental [*a*] fragment remains the same. Recombination between [*a*] and [*c*] (not shown) should produce very few recombinants which would show novel fragments when probed with [*a*] (as discussed below). When recombination occurs between [*a*] and [*b*] on fragments which are part of a class I RFLP (Figure 2D) no novel fragment sizes are generated. In this case the recombinant fragments are the same size as the parental fragments probed with [*a*].

The sizes of the novel recombinant fragments generated depend upon the sizes of the parental fragments and the degree of asymmetric overlap at each end of the polymorphic pair. In all cases the sum of the sizes of the two novel fragments will always equal the sum of the two parental fragments. Recombination within class II RFLPs with a subset relationship will generate novel fragments with an asymmetric relationship, and recombination within asymmetric class II RFLPs will generate novel fragments with a subset relationship. If the class II RFLP recombinant fragments are very large (100–2000 kb), but can be resolved by PFGE, then it should be possible to use the novel sizes of recombinant fragments to estimate the physical distance corresponding to the genetic distance (cM) between these loci ([*a*] and [*b*] in Figure 2). In a typical genetic mapping experiment, the AB hybrid might be backcrossed with the homozygous AA parent. Relevant recombination between the class II RFLPs would occur during meiosis in the AB parent yielding a subset of germ cells with A* or B*. The frequency of offspring from the AB × AA backcross identified as recombinants between [*a*] and [*b*] determine the genetic distance between the loci. These recombinant offspring would always have the one parental A chromosome and either an A* or B* chromosome containing a novel recombinant fragment. The parental and novel recombinant fragments would be revealed when a Southern blot of these DNAs is probed with [*a*]. If every organism containing a chromosome resulting from a recombination event between [*a*] and [*b*], as shown by classical genetic analysis, also had one of the novel recombinant fragments, then the two loci, [*a*] and [*b*], will in almost all cases be on the same DNA fragment. Exceptions involving markers on different fragments are discussed in the following section. If the genetically defined recombinants between [*a*] and [*b*] are examined

with PFGE, and the novel fragments (e.g. see A* or B* in Figure 2A) are always detected when examining recombination between the 450- and 700-kb fragments, then it is highly likely that [*a*] and [*b*] are on the same fragment and the maximum distance between these loci can be estimated to be no more than 450 kb. This is a reasonable physical distance to consider a chromosome walk from [*a*] to [*b*]. If the genetically defined recombinants between [*a*] and a phenotypic marker [*c*] are also examined and some recombinants do not give novel fragment sizes, then [*a*] and [*c*] are not on the same restriction fragment of a class II RFLP. The parental chromosomes in Figures 2A to 2D show [*a*] separated by several restriction cleavage sites from [*c*] as they then might be positioned in genome. Recombinants between [*a*] and [*c*] are discussed below.

The potential use of class II RFLP recombination assays depends both upon the design of classical genetic experiments, since the individuals examined must be sufficiently polymorphic to contain class II RFLPs, and upon identifying individuals containing recombination events in the region of interest. In other words, the individuals with recombination between [*a*] and [*b*] in class II RFLPs become the key to measuring physical distance. These individuals will often be identified in the progeny of classical genetic experiments which measure the recombination between the DNA probe and phenotypic loci. Only those organisms already identified as recombinants need to be assayed at the DNA level. The recombinant individuals can be used in direct determinations of the physical distance between the DNA probe and phenotypic loci, and in proof of physical linkage between these loci on the same DNA fragment. This assay would also be useful in organisms which are polyploid because the novel class II RFLP recombinants will hybridize independent of parental RFLPs which might still be present. Class II RFLPs may even be useful in proving standard linkage relationships in systems where linkage might otherwise be obscure, such as in determining the locations of genes encoding quantitative traits. Organisms with saturated RFLP maps are ideal for these studies because the estimate of physical distance to the phenotypic marker can be made from DNA probes from both directions.

Considering a linked phenotypic marker which is not on the same PFGE fragment as the probe: When the physical distance between a probe and the phenotypic marker is very great, then the two loci will probably not be on the same large PFGE fragment but may be several large fragments apart on the genome. For example, assume that RFLP probe [*a*] maps 5 cM from phenotypic marker [*c*] (shown in Figure 2B) and assume that A and B are from two highly polymorphic genomes. This genetic distance could easily represent 1,000 to 50,000 kb in a large eukaryotic genome. Most of this range is beyond direct measurement by present PFGE technology. However fragments of 100 to 2,000 kb can be measured. When a large number of recombinants can be examined, it is likely in some individuals that recombination will occur between [*a*] and [*c*] within the 600-kb fragment containing [*a*], the RFLP probe (Figure 2B). If four of ten individuals with genetic recombination between [*a*] and [*c*] display novel fragment sizes as a result of recombination in these class II RFLP fragments, then it could be estimated that these two loci may be less than approximately 1500 kb apart. This result is generated from a small sample size and as such is subject to large statistical variation. The assay could be improved by examining large numbers of recombinant individuals identified by classical genetic experiments between easily scorable phenotypic markers in the region of the DNA probe. This would be feasible in organisms such as maize, tomato, Petunia or Drosophila where numerous phenotypic markers have been mapped on the genome. Due to the greater genetic distance between the two distant loci, a larger number of recombinant organisms will be generated from a single cross making the determination more feasible.

These working estimates are subject to the same problems that are described for estimating average kb/cM ratios. This approach depends upon the random distribution of recombination events between [*a*] and [*c*]. In the examples shown, this assay requires that recombination frequencies be constant over a 5 cM region. If among the several genetic recombinants between [*a*] and [*c*], the [*a*] probe did not hybridize to any recombinant class II RFLP fragments after examining digestions with several different restriction endonucleases, then it could be interpreted as meaning that [*a*] and [*c*] may be physically very far apart, perhaps too far to plan a chromosome walk. These data could also mean that a hot spot for recombination exists between [*a*] and [*c*] in one of the fragments which did not hybridize to [*a*]. A hot spot for recombination closer to [*c*] could result in an overestimation of physical distance. A hot spot for recombination on the fragment probed for by [*a*] could result in all recombinants between [*a*] and [*c*] generating novel fragment sizes, thus [*a*] and [*c*] could be misinterpreted as being on the same fragment. To control for this possibility, organisms with a saturated RFLP map have the tremendous advantage that the physical distances can be estimated from both directions. Partial digestions, described below, should also allow these larger distances to be examined more accurately.

ANALYSIS OF CLASS II RFLPs

The success of these assays of physical distance depends upon examining the size of large recombinant RFLPs by PFGE. This and related electrophoretic techniques allow DNA molecules ranging in size

from small genes (1 to 10 kb) to small chromosomes (2000 kb) to be resolved on agarose gels (CARLE and OLSEN 1985; CARLE, FRANK and OLSON 1986; CHU, VOLLRATH and DAVIS 1986; SCHWARTZ and CANTOR 1984; GEMMILL *et al.* 1987). A number of regions in animal genomes originally described at the genetic level or at the gene level have now been mapped physically using PFGE. The mouse MHC has been mapped over a 600-kb region using PFGE in combination with chromosome walking (STEINMETZ, STEPHAN and LINDAHL 1986; MULLER *et al.* 1987). The human MHC map, which was examined primarily by PFGE analyses of large overlapping restriction fragments, covers a 3000-kb region (LAWRANCE *et al.* 1987). PFGE has been used to generate a large scale restriction map of the amplified DNA sequences in human neuroblastomas (SHILOH *et al.* 1985), to map a 4000-kb region of DNA controlling Duchenne muscular dystrophy (KENWRICK *et al.* 1987), and to generate a restriction map of the entire *Escherichia coli* genome (SMITH *et al.* 1987). These distances exceed the estimates of the physical distance (kb) corresponding to a recombination unit (cM) for all but the largest plant genomes.

Most of the studies cited examine many single fragments between 100 and 800 kb and use overlapping fragment mapping techniques to obtain maps of larger dimensions. Two potential problems in examining larger fragments are the limits of PFGE technology and the ability to generate larger restriction fragments. Separations of very large fragments, including the chromosomes in the yeast *Saccharomyces pombe*, which are 3000, 6000 and 9000 kb, have already been demonstrated (BARLOW and LEHRACH 1987), so PFGE separation technology will probably not be limiting. Although only two 8-base recognition restriction endonucleases, *Sfi*I and *Not*I, have been identified, it is likely that many more will be identified using PFGE screening techniques and larger substrate molecules. Partial digestions with *Sfi*I, *Not*I or rare 6-base cutters can also be used to yield fragments near the desired size range (NGUYEN *et al.* 1987; SMITH *et al.* 1987). In Figure 2, recombinants between the A and B chromosomes are shown. Partial digestion profiles of the A* and B* chromosomes are very different than those generated from the A and B chromosomes, respectively. Even if recombination takes place close to [*c*], some major differences in the partial digestions will be detected in such a comparison. As long as [*c*] is within a few large restriction fragments of [*a*], the pattern should not be too complex to identify novel recombinant class II RFLPs.

Since most of the restriction endonuclease recognition sequences are guanine plus cytosine (GC) rich, the GC poor composition of most plant genomes will make it easier to find enzymes that generate large fragments. In soybean, for example, which has a genome composed of 35% GC, eight-base recognition restriction endonucleases like *Not*I with all Gs or Cs in their recognition sequence should generate exceptionally large fragments. Based on Equation 2 described below and in PHILLIPS, ARNOLD and IVARIE (1987a, b), which also considers tri- and tetranucleotide frequencies and based on the existing 60 kb of soybean sequence in the NIH Data Base, the average predicted size for a *Not*I fragment is 1600 kb. Partial digestions with such enzymes should cover the desired distances needed in these assays. Methods for blocking cleavage at all but a small subset of sites (MCCLELLAND, KESSLER and BITTNER 1984) and a new chemical cleavage technique may allow even more control over the number of fragments generated in any genome (MOSER and DERVAN 1987).

FREQUENCY OF CLASS II RFLPs

Predicted frequencies: The practical significance of recombination assays between class II RFLPs will be limited by the availability of genetically compatible organisms with sufficiently divergent genomes. The following derivations describe the relationship between the estimated sequence divergence within a species and the predicted frequency of class II polymorphisms. The development follows NEI and LI (1979) and UPHOLT (1977) with the nomenclature following that of NEI (1987). These formulas generally describe this relationship and may be applied to highly divergent genomes where class II RFLPs would be most frequent.

The corrected nucleotide divergence will be called d. It represents the expected number of nucleotide substitutions which have occurred per site in sequences contained in the two homologous chromosomes A and B (Figure 2). These nucleotide substitutions have occurred since the two homologs descended from a common ancestral chromosome t time units ago. The corrected nucleotide divergence is expected to increase linearly with time according to the molecular clock hypothesis such that $d = 2\lambda t$, where λ is the substitution rate of the molecular clock. The corrected nucleotide divergence is now routinely reported for a number of organisms and is estimated either directly from sequence data, indirectly from restriction maps, or from the number of shared fragments among restriction enzyme digestion profiles (NEI 1987). The quantity d is related to the observed percent sequence divergence, $100 \cdot p$, between A and B by correcting for multiple hit kinetics (JUKES and CANTOR 1969):

$$d = -\tfrac{3}{4} \ln(1 - 4p/3) \quad \text{or} \qquad p = \tfrac{3}{4}(1 - e^{-4/3d}). \tag{1}$$

The occurrence of a restriction endonuclease site in the A and B homologs will be considered. This could, for example, be a restriction endonuclease cleavage site flanking the probe [*a*] (Figures 1 and 2). The probability α of occurrence of a restriction site can be estimated from the oligonucleotide composition of the probe region and the composition of the recognition site (PHILLIPS, ARNOLD and IVARIE 1987a, b; ARNOLD *et al.* 1988). This probability α is also the predicted frequency of occurrence of a site in the genome. For example, the probability that a given 6-bp sequence is an *Eco*RI site (GAATTC) can be approximated from the base composition by the following formulas (*Note*: Let f(G) denote the frequency of nucleotide G in the genome and let f(GAAT) denote the frequency of the tetranucleotide, GAAT, in the genome, etc.):

$$\alpha = \text{f(G)f(A)f(A)f(T)f(T)f(C)}$$

or more accurately described from overlapping tetra- and trinucleotide frequencies:

$$\alpha = \frac{\text{f(GAAT)f(AATT)f(ATTC)}}{\text{f(AAT)f(ATT)}}. \qquad (2)$$

The method used for calculating the probability of there being a restriction site depends on how much sequence data are available on the test organism. Alternatively, α can be measured empirically from densitometric scans of restriction endonuclease digested genomic DNA (PETERSON 1988), where $1/\alpha$ is the average fragment size. In practice one would expect some statistical variation in the frequency of sites in a particular region of the genome.

In general, with P being the probability of loss of a site, the three possible fates of a given site which occurs with a probability α on both homologs A and B are: (i) both sites remain unchanged, which has probability $(1 - P)^2$; (ii) the site on either chromosome A or B is lost and the site on the other chromosome remains unchanged, which has probability $2P(1 - P)$; or (iii) the sites on both homologs are lost, which has probability P^2. If event (ii) occurs at a restriction endonuclease recognition site, then a RFLP will be generated. If the site is lost from both homologs (iii), there may or may not be a polymorphism. This will depend upon whether an adjacent site is different between homologs, given that the first site is lost from both homologs. If this second site is lost from both homologs, then the next adjacent site must be considered, and so on. Treating the chromosomes examined as effectively infinite in length, the probability that the event (iii) occurs and generates a polymorphism is:

$$\alpha P^2[2\alpha P(1 - P) + (\alpha P^2)2\alpha P(1 - P) + (\alpha P^2)^2\, 2\alpha P(1 - P) + \ldots]$$

or

$$\alpha P^2 2\alpha P(1 - P)[1 + \alpha P^2 + (\alpha P^2)^2 + (\alpha P^2)^3 + \ldots].$$

The series is geometric and can be summed to yield the probability that event (iii) occurs and yields a polymorphism:

$$2\alpha P(1 - P)\alpha P^2/(1 - \alpha P^2).$$

Detection of the more remote sites is limited by the size of fragments which can be resolved by PFGE; however, the contribution of successively more remote sites to this sum is small.

Adding this probability to the probability that event (ii) occurs, producing a polymorphism, yields the probability $\pi(\alpha)$ of a polymorphism at a given site which is dependent upon nucleotide composition:

$$\begin{aligned}\pi(\alpha) &= 2\alpha P(1 - P)\{1 + [\alpha P^2/(1 - \alpha P^2)]\} \\ &= 2\alpha P(1 - P)/(1 - \alpha P^2).\end{aligned} \qquad (3)$$

Another way in which this formula is useful is in calculating the average distance from one polymorphic restriction site (*e.g.*, *Eco*RI) to the next polymorphic *Eco*RI site. This distance can be expressed as $1/\pi(\alpha)$. It will depend upon both the degree of divergence between two genomes and the oligonucleotide composition of the genome.

Repeating the above argument without placing conditions on the nucleotide composition of the site or that of the genome yields the probability of a polymorphism as a function of the probability of loss of a site:

$$\begin{aligned}\pi &= 2P(1 - P)\{1 + [P^2/(1 - P^2)]\} \\ &= 2P/(1 + P).\end{aligned} \qquad (4)$$

If the probability of a polymorphism, π, is small, then logically the occurrence of class II RFLPs will be small. Furthermore, the average spacing between polymorphisms may be expressed as $1/\pi$. This is not dependent on oligonucleotide composition because we are not specifying the sequence of the site.

What remains to be established is a relation between the probability of loss of a site P and percent sequence divergence $100 \cdot p$ (Equation 1). This will give the relation between the probability of polymorphism π and corrected nucleotide divergence, d. The probability, $1 - P$, that a site of length r remains unchanged from the ancestral chromosome is:

$$1 - P = (1 - p)^r. \qquad (5)$$

Substituting (1) into (5) yields P in terms of d:

$$P = 1 - (\tfrac{1}{4} + \tfrac{3}{4}\, e^{-4/3d})^r. \qquad (6)$$

Substituting P into (4) yields a formula for the frequency of polymorphism π in terms of the amount of corrected sequence divergence d. For example, suppose $d = 0.01$, then for a 6-base ($r = 6$) recognition restriction endonuclease site, $P = 0.06$. Substituting

the value of $P = 0.06$ into (4) yields a frequency of polymorphism of $\pi = 0.11$.

Class II polymorphic fragments have two restriction sites which are polymorphic at both ends of a probed fragment (Figure 2). If the nucleotide substitution process is independent at different sites, then the probability of a class II polymorphism at a given location is simply the square of the probability of a single polymorphism:

$$\pi_{(II)}(\alpha) = [\pi(\alpha)]^2.$$

If no conditions are placed on the nucleotide composition of the site or the genome, the probability of a class II polymorphism is simply:

$$\pi_{(II)} = \pi^2. \quad (7)$$

If for example $d = 0.1$ and $r = 6$, then $P = 0.45$ from (6). Substituting P into (4) yields $\pi = 0.62$ or a probability of any RFLP being a class II polymorphism of $\pi_{(II)} = 0.38$.

Figure 3 presents the predicted frequency of RFLPs and class II RFLPs generated by 6- and 8-base recognition restriction endonucleases based on (4) and (7). These curves can be useful in two ways. First, given the nucleotide divergence d of two separate inbred genomes, one can assess whether the predicted frequency of class II polymorphisms is sufficient for class II RFLP recombination assays. In particular, pairs of inbred organisms with d greater than 0.065 sequence divergence will have a useful frequency of class II RFLPs. For two organisms with 0.065 sequence divergence, 24% of all the restriction fragments generated by 6-base recognition restriction endonucleases should be class II RFLPs. Even more significant to the practical application of class II RFLP recombination assays is that at this level of divergence, 34% of all fragments observed with 8-base recognition restriction endonucleases will be class II RFLPs.

Second, given the percent polymorphism in a probe region, one can estimate the nucleotide divergence of two genomes without resorting to restriction mapping or sequencing. For example, if the observed frequency of RFLPs for 6-base recognition enzymes were 62%, then from Figure 3 the estimated nucleotide divergence would be $d = 0.10$ and the frequency of class II RFLPs with 8-base recognition enzymes would be expected to be 50%.

Observed levels of divergence: The RFLPs observed between related pairs of organisms can be generated by several mechanisms of genome divergence including base substitution, slippage repair between related but nonhomologous sequences, insertions and deletions and the various mechanisms of genome turnover (BAIRD and MEAGHER 1987; FLAVELL 1982; MURRAY, PETERS and THOMPSON 1981). In two inbred Petunia varieties, 87% of the fragments generated by primarily 6-base recognition endonucleases which hybridized to actin gene probes were RFLPs (MCLEAN *et al.* 1988). Based on the above calculation and the assumption that this divergence is due to nucleotide sequence divergence of $d = 0.26$ we can predict that 86% of all 8-bp fragments will be class II RFLPs between these two genomes. However, it is not possible to determine from RFLP data what mechanism generated this high level restriction site polymorphism. DNA sequence data, genomic Southern blots and genetic data on the Petunia actin genes suggest that within this superfamily of genes substantial sequence divergence exists between closely related inbred Petunia varieties, and any or all of these mechanisms could have contributed to the level of polymorphism (BAIRD and MEAGHER 1987; MCLEAN *et al.* 1988). Even if the high degree of polymorphism observed in small fragments produced with 6-base enzymes in the Petunia system was generated primarily by large numbers of insertions and deletions the larger fragments examined with 8-base recognition restriction endonucleases would contain dozens and perhaps hundreds of insertion deletion differences. Since the numerous insertion events in each divergent parental genome can be assumed to be independent, it is likely large fragments from these genomes would also be polymorphic in size. The general behavior of class II RFLPs in the class II RFLP recombination assays should be the same regardless of whether the polymorphisms were generated by base substitutions or large numbers of small insertions and deletions.

In many plant genera it is possible to find divergent species which may be drastically different in genome content (BENNETT, SMITH and HESLOP-HARRISON

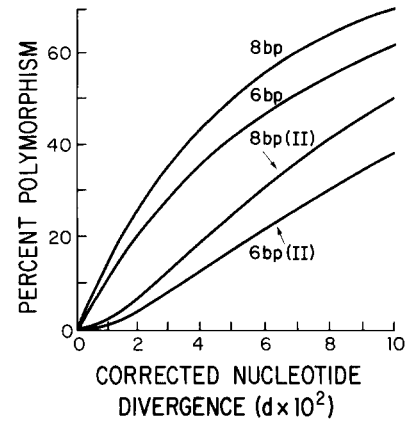

FIGURE 3.—Frequency of Class II RFLPs. The percent restriction fragment length polymorphisms are graphed as a function of corrected nucleotide divergence for 6- and 8-bp recognition restriction endonucleases, labeled 6 bp and 8 bp, respectively. The percent class II restriction fragment length polymorphisms are also graphed for 6- and 8-base recognition endonucleases and are labeled 6 bp(II) and 8 bp(II), respectively. Formulas (4) and (7) presented in the text were used to generate these curves. A more extensive graph examining percent polymorphism at higher values of d is available from the authors on request.

1982; SAMUEL, SMITH and BÉNNETT 1986) but are still genetically compatible. Stable and often fertile interspecific hybrids can be made at reasonable frequencies between individuals differing substantially in genome content. Such hybrids should contain high frequencies of RFLPs for a large number of independent loci. BERNATZKY and TANKSLEY (1986a, b) demonstrated that for two divergent but genetically compatible tomato species, *Lycopersicon esculentum* and *Lycopersicon penellii*, 34 randomly selected cDNA clones hybridized to RFLPs for many restriction endonucleases. In the examination of two phenotypically divergent Petunia varieties, V23 and R51, RFLPs were found for 48 of 55 comparisons (MCLEAN *et al.* 1988). In six races of maize, 16 of 18 endosperm cDNA probes were found to yield RFLPs (BURR *et al.* 1983). Sucrose synthase (*Sh1*) and alcohol dehydrogenase loci (*Adh1*) are also known to be highly polymorphic between closely related maize plants (EVOLA, BURR and BURR 1986; JOHNS, STROMMER and FREELING 1983).

Conclusion: The use of class II RFLPs in recombination assays will help to physically link classical genetic data with RFLP maps and RFLP probes. It appears useful to set up classical genetic mapping studies using RFLPs with class II RFLP recombination assays in mind. The application of RFLP maps to the isolation of phenotypic marker loci and identification of hot spots for recombination is strongly dependent upon the availability of class II RFLPs in these systems.

We would like to thank TOM GERATS, WYATT ANDERSON, STEVE TANKSLEY and MIKE MURRAY for their help and suggestions in the preparation of the manuscript. MARJORIE ASMUSSEN deserves particular mention for her help with the mathematics and preparation of the manuscript. This work was supported by grants to RBM from the National Institutes of Health (GM 36397-03) and the U.S. Department of Agriculture (88-37234-3548), and to J.A. from DECOR and the National Science Foundation (BSR-8716804).

LITERATURE CITED

ARNOLD, J., A. J. CUTICCHIA, D. A. NEWSOME, W. W. JENNINGS and R. IVARIE, 1988 Mono- through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis. Nucleic Acids Res. **16:** 7145–7158.

BAIRD, W. V., and R. B. MEAGHER, 1987 A complex gene superfamily encodes actin in Petunia. EMBO J. **6:** 3223–3231.

BARLOW, D. P., and H. LEHRACH, 1987 Genetics by gel electrophoresis: the impact of pulsed field gel electrophoresis on mammalian genetics. Trends Genet. **3:** 167–171.

BENDER, W., M. AKAM, F. KARCH, P. A. BEACHY, M. PEIFER, P. SPIERER, E. B. LEWIS and D. S. HOGNESS, 1983 Molecular genetics of the *bithorax* complex in *Drosophila melanogaster*. Science **221:** 23–29.

BENNETT, M. D., J. B. SMITH and J. S. HESLOP-HARRISON, 1982 Nuclear DNA amounts in angiosperms. Proc. R. Soc. Lond. B. **216:** 179–199.

BERNATZKY, R., and S. D. TANKSLEY, 1986a Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. Genetics **112:** 887–898.

BERNATZKY, R., and S. D. TANKSLEY 1986b Genetics of actin-related sequences in tomato. Theor. Appl. Genet. **72:** 314–321.

BURKE, D. T., G. F. CARLE and M. V. OLSON, 1987 Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science **236:** 806–812.

BURR, B., S. V. EVOLA, F. A. BURR and J. S. BECKMANN, 1983 The application of restriction fragment length polymorphism to plant breeding. pp. 45–59. In: *Genetic Engineering Principles and Methods*, Vol. 5, Edited by J. K. SETLOW and A. HOLLAENDER. Plenum Press, New York.

CARLE, G. F., and M. V. OLSON, 1985 An electrophoretic karyotype for yeast. Proc. Natl. Acad. Sci. USA **82:** 3756–3760.

CARLE, G. F., M. FRANK and M. V. OLSON, 1986 Electrophoretic separations of large DNA molecules by periodic inversion of the electric field. Science **232:** 65–68.

CLARKE, L., and J. CARBON, 1980 Isolation of the centromere-linked *CDC10* gene by complementation in yeast. Proc. Natl. Acad. Sci. USA **77:** 2173–2177.

CLARKE, L., H. AMSTUTZ, B. FISHEL and J. CARBON, 1986 Analysis of centromeric DNA in the fission yeast *Schizosaccharomyces pombe*. Proc. Natl. Acad. Sci. USA **83:** 8253–8257.

CHU, G., D. VOLLRATH and R. W. DAVIS, 1986 Separation of large DNA molecules by contour-clamped homogenous electric fields. Science **234:** 1582–1585.

DONIS-KELLER, H., P. GREEN, C. HELMS, S. CARTINHOUR, B. WEIFFENBACH, K. STEPHENS, T. P. KEITH, D. W. BOWDEN, D. R. SMITH, E. S. LANDER, D. BOTSTEIN, G. AKOTS, K. S. REDIKER, T. GRAVIUS, V. A. BROWN, M. B. RISING, C. PARKER, J. A. POWERS, D. E. WATT, E. R. KAUFFMAN, A. BRICKER, P. PHIPPS, H. MULLER-KAHLE, T. R. FULTON, S. NG, J. W. SCHUMM, J. C. BRAMAN, R. G. KNOWLTON, D. F. BARKER, S. M. CROOKS, S. E. LINCOLN, M. J. DALY and J. ABRAHAMSON, 1987 A genetic linkage map of the human genome. Cell **51:** 319–337.

DOONER, H. K., 1986 Genetic fine structure of the *Bronze* locus in maize. Genetics **113:** 1021–1036.

DOONER, H. K., E. WECK, S. ADAMS, E. RALSTON, M. FAVREAU and J. ENGLISH, 1985 A molecular genetic analysis of insertions in the *Bronze* locus in maize. Mol. Gen. Genet. **200:** 240–246.

EVOLA, S. V., F. A. BURR and B. BURR, 1986 The suitability of restriction fragment length polymorphisms as genetic markers in maize. Theor. Appl. Genet. **71:** 765–771.

FINCHAM, J. R. S., 1983 *Genetics*. Jones and Bartlett Publishers, Boston.

FLAVELL, R., 1982 Sequence amplification, deletion and rearrangement: major sources of variation during species divergence. pp. 301–323. In: *Genome Evolution*, Edited by G. A. DOVER and R. B. FLAVELL. Academic Press, London.

GARDINER, K., W. LAAS and D. PATTERSON, 1986 Fractionation of large mammalian DNA restriction fragments using vertical pulsed-field gradient gel electrophoresis. Somatic Cell Mol. Genet. **12:** 185–195.

GEMMILL, R. M., J. F. COYLE-MORRIS, F. D. MCPEEK, JR., L. F. WARE-URIBE and F. HECHT, 1987 Construction of long-range restriction maps in human DNA using pulsed field gel electrophoresis. Gene Anal. Techn. **4:** 119–131.

HELENTJARIS, T., M. SLOCUM, S. WRIGHT, A. SHAEFER and J. NIENHUIS, 1986 Construction of genetic linkage maps in plants using restriction fragment polymorphisms. Theor. Appl. Genet. **72:** 761–769.

JOHNS, M. A., J. N. STROMMER and M. FREELING, 1983 Exceptionally high levels of restriction site polymorphism in DNA near the maize *Adh1* gene. Genetics **105:** 733–743.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules. pp. 21–132. In: *Mammalian Protein Metabolism*, Edited by H. N. MONRO. Academic Press, New York.

KENWRICK, S., M. PATTERSON, A. SPEER, K. FISCHBECK and K. DAVIES, 1987 Molecular analysis of the Duchenne muscular

dystrophy region using pulsed field gel electrophoresis. Cell **48:** 351–357.

LANDER, E. S., and D. BOTSTEIN, 1986 Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc. Natl. Acad. Sci. USA **83:** 7353–7357.

LANDER, E. S., and D. BOTSTEIN, 1987 Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science **236:** 1567–1570.

LAWRANCE, S. K., C. L. SMITH, R. SRIVASTAVA, C. R. CANTOR and S. M. WEISSMAN, 1987 Megabase-scale mapping of the *HLA* gene complex by pulsed field gel electrophoresis. Science **235:** 1387–1389.

LUCCHESI, J. C., and D. T. SUZUKI, 1968 The interchromosomal control of recombination. Annu. Rev. Genet. **2:** 53–86.

MCCLELLAND, M., L. G. KESSLER and M. BITTNER, 1984 Site-specific cleavage of DNA at 8- and 10-base pair sequences. Proc. Natl. Acad. Sci. USA **81:** 983–987.

MCLEAN, M., W. V. BAIRD, A. G. M. GERATS and R. B. MEAGHER, 1988 Determination of copy number and linkage relationships among five actin gene subfamilies in *Petunia hybrida*. Plant Mol. Biol. (in press).

MICHIELS, F., M. BURMEISTER and H. LEHRACH, 1987 Derivation of clones close to *met* by preparative field inversion gel electrophoresis. Science **236:** 1305–1308.

MOSER, H. E., and P. B. DERVAN, 1987 Sequence-specific cleavage of double helical DNA by triple helix formation. Science **238:** 645–650.

MULLER, U., D. STEPHAN, P. PHILIPPSEN and M. STEINMETZ, 1987 Orientation and molecular map position of the complement genes in mouse *MHC*. EMBO J. **6:** 369–373.

MURRAY, M. G., D. L. PETERS and W. F. THOMPSON, 1981 Ancient repeated sequences in the pea and mung bean genomes and implications for genome evolution. J. Mol. Evol. **17:** 31–42.

NAKASEKO, Y., Y. ADACHI, S. FUNAHASHI, O. NIWA and M. YANAGIDA, 1986 Chromosome walking shows a highly homologous repetitive sequence present in all the centromere regions of fission yeast. EMBO J. **5:** 1011–1021.

NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

NEI, M., and W.-H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

NGUYEN, C., P. PONTAROTTI, D. BIRNBAUM, G. CHIMINI, J. A. REY, J.-F. MATTEI and B. R. JORDAN, 1987 Large scale physical mapping in the q27 region of the human X chromosome: the coagulation factor IX gene and the *mcf.2* transforming sequence are separated by at most 270 kilobase pairs and are surrounded by several 'HTF islands.' EMBO J. **6:** 3285–3289.

NIENHUIS, J., T. HELENTJARIS, M. SLOCUM, B. RUGGERO and A. SCHAEFER, 1987 Restriction fragment length polymorphism analysis of loci associated with insect resistance in tomato. Crop Sci. **27:** 797–803.

PETERSON, R. C., 1988 Prediction of the frequencies of restriction endonuclease recognition sequences using di- and mononucleotide frequencies. Biotechniques **6:** 34–39.

PHILLIPS, G. J., J. ARNOLD and R. IVARIE, 1987a Mono- through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. Nucleic Acids Res. **15:** 2611–2626.

PHILLIPS, G. J., J. ARNOLD and R. IVARIE, 1987b The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and under-represented sequences by Markov chain analysis. Nucleic Acids Res. **15:** 2627–2638.

PORTEOUS, D. J., 1987 Chromosome mediated gene transfer: a functional assay for complex loci and an aid to human genome mapping. Trends Genet. **3:** 177–182.

POUSTKA, A., T. M. POHL, D. P. BARLOW, A.-M. FRISCHAUF and H. LEHRACH, 1987 Construction and use of human chromosome jumping libraries from *Not*I-digested DNA. Nature **325:** 353–355.

REES, H., and A. DURRANT, 1986 Recombination and genome size. Theor. Appl. Genet. **73:** 72–76.

SAMUEL, R., J. B. SMITH and M. D. BENNETT, 1986 Nuclear DNA variation in *Piper* (Piperaceae). Can. J. Genet. Cytol. **28:** 1041–1043.

SCHWARTZ, D. C., and C. R. CANTOR, 1984 Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell **37:** 67–75.

SENECOFF, J. F., and M. M. COX, 1986 Directionality in FLP protein-promoted site specific recombination is mediated by DNA-DNA pairing. J. Biol. Chem. **261:** 7380–7386.

SHILOH, Y., J. SHIPLEY, G. M. BRODEUR, G. BRUNS, B. KORF, T. DONLON, R. R. SCHRECK, R. SEEGER, K. SAKAI and S. A. LATT, 1985 Differential amplification, assembly, and relocation of multiple DNA sequences in human neuroblastomas and neuroblastoma cell lines. Proc. Natl. Acad. Sci. USA **82:** 3761–3765.

SMITH, C. L., J. G. ECONOME, A. SCHUTT, S. KLCO and C. R. CANTOR, 1987 A physical map of the *Escherichia coli* K12 genome. Science **236:** 1448–1453.

STEINMETZ, M., D. STEPHAN and K. F. LINDAHL, 1986 Gene organization and recombinational hotspots in the murine major histocompatibility complex. Cell **44:** 895–904.

UPHOLT, W. B., 1977 Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. Nucleic Acids Res. **4:** 1257–1265.

ZACHAR, Z., and P. M. BINGHAM, 1982 Regulation of *white* locus expression: the structure of mutant alleles at the *white* locus of *Drosophila melanogaster*. Cell **30:** 529–541.

Communicating editor: M. T. CLEGG