Am J Hum Genet 32:314–331, 1980

Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms

DAVID BOTSTEIN,[1] RAYMOND L. WHITE,[2] MARK SKOLNICK,[3] AND RONALD W. DAVIS[4]

SUMMARY

We describe a new basis for the construction of a genetic linkage map of the human genome. The basic principle of the mapping scheme is to develop, by recombinant DNA techniques, random single-copy DNA probes capable of detecting DNA sequence polymorphisms, when hybridized to restriction digests of an individual's DNA. Each of these probes will define a locus. Loci can be expanded or contracted to include more or less polymorphism by further application of recombinant DNA technology. Suitably polymorphic loci can be tested for linkage relationships in human pedigrees by established methods; and loci can be arranged into linkage groups to form a true genetic map of "DNA marker loci." Pedigrees in which inherited traits are known to be segregating can then be analyzed, making possible the mapping of the gene(s) responsible for the trait with respect to the DNA marker loci, without requiring direct access to a specified gene's DNA. For inherited diseases mapped in this way, linked DNA marker loci can be used predictively for genetic counseling.

INTRODUCTION

Although it is possible to detect linkage among simple Mendelian traits in humans, no method of systematically mapping human genes has been devised, largely because of the paucity of highly polymorphic marker loci.

Received November 20, 1979; revised January 4, 1980.

This project was supported by grants CA-16573 from the National Institutes of Health, and NP-193B, NP-286, and VC-245 from the American Cancer Society.

[1] Department of Biology, Massachusetts Institute of Technology, Cambridge, Mass.

[2] Department of Microbiology, University of Massachusetts Medical School, Worcester.

[3] Department of Medical Biophysics & Computing, University of Utah College of Medicine and LDS Hospital, Salt Lake City.

[4] Department of Biochemistry, Stanford University School of Medicine, Stanford, Calif.

The advent of recombinant DNA technology has suggested a theoretically possible way to define an arbitrarily large number of arbitrarily polymorphic marker loci. In this paper, we discuss this general approach to the construction of a human genetic linkage map using recombinant DNA probes to define marker loci which are polymorphic in DNA sequences; a subset of such DNA polymorphisms can readily be detected directly as differences in the length of DNA fragments after digestion with DNA sequence-specific restriction endonucleases. These restriction fragment length polymorphisms (RFLPs) can be easily assayed in individuals, facilitating large population studies. Small volumes of peripheral blood should provide sufficient lymphocyte DNA for analysis.

RFLPs should be inherited as simple Mendelian codominant markers; this can be verified in family studies. Linkage relationships among RFLPs can be established using pedigree analysis [1–3]. Evaluation of many DNA marker loci should allow the establishment of a set of well-spaced, highly polymorphic genetic markers covering the entire human genome. Since the RFLPs are being used simply as genetic markers, any trait caused wholly or partially by a major locus segregating in a pedigree can be mapped. Such a procedure would not require any knowledge of the biochemical nature of the trait or of the nature of the alterations in the DNA responsible for the trait. No specific gene isolation is required, and the RFLPs can be random sequences functionally unrelated and physically distant from the DNA encoding the locus of interest. In fact, on the average, the physical distance represented by 1 cM (.01 recombination fraction) is about 1,000,000 base pairs (1,000 kilobases [kb]) [4]; we suggest the need for linked markers about 20 cM apart (i.e., 20 million base pairs).

Such a mapping procedure will also allow better definition and substantiation of models of inheritance for the many familial traits which have been refractory to simple genetic analysis in humans. The usefulness of a tightly linked polymorphic locus in resolving modes of inheritance has recently been demonstrated [5].

RESTRICTION FRAGMENT LENGTH POLYMORPHISM (RFLP): DEFINITION

DNA restriction enzymes (see [6] for review) recognize specific sequences in DNA and catalyze endonucleolytic cleavages, yielding fragments of defined lengths. Restriction fragments may be displayed by electrophoresis in agarose gels, separating the fragments according to their molecular size. Differences among individuals in the lengths of a particular restriction fragment could result from many kinds of genotypic differences: one or more individual bases could differ, resulting in loss of a cleavage site or formation of a new one; alternatively, insertion or deletion of blocks of DNA within a fragment could alter its size. These genotypic changes can all be recognized by the altered mobility of restriction fragments on agarose gel electrophoresis.

Fragments encoding *specific* sequences from within a large and complex population of DNA fragments can be detected by hybridization using the method of Southern [7]. The DNA from an agarose gel is transferred onto nitrocellulose paper and hybridized with radioactive probe sequences. Recently, the use of probe sequences of very high specific radioactivity has permitted the identification by this method of single-copy sequences in mammalian DNA [8]. This powerful new technology makes possible the identification of variants from within a specific region of the genome, using digest of total human DNA.

To illustrate, figure 1*a* shows a schematic representation of the chromosomal arrangement of sites for two restriction enzymes for a pair of homologous chromosomes. Restriction enzyme B cleaves at identical locations (b_i and b_{i+1}) in both chromosomes, yielding only one fragment length homologous to the probe sequence, which is revealed as a single band of hybridization as indicated in figure 1*b*. Restriction enzyme A, on the other hand, is seen to be cleaving the homologous DNA sequences at one identical site, a_i, but the location of the next enzyme A site differs between the two homologs. This difference in location between a_{i+1} and a'_{i+1} generates the pattern (fig. 1*b*) of two bands when hybridized with the probe.

RFLPs AS GENETIC MARKERS

RFLPs were first used as a tool for genetic analysis in 1974. Linkage of temperature-sensitive mutations of adenovirus to specific restriction fragment length differences was used to locate the mutations on a physical map of the restriction fragments [9]. Other studies have shown the maternal inheritance of mammalian mitochondrial DNA [10] and the existence of RFLP within mitochondrial DNAs from two individual humans, as well as from several human cell lines [11].

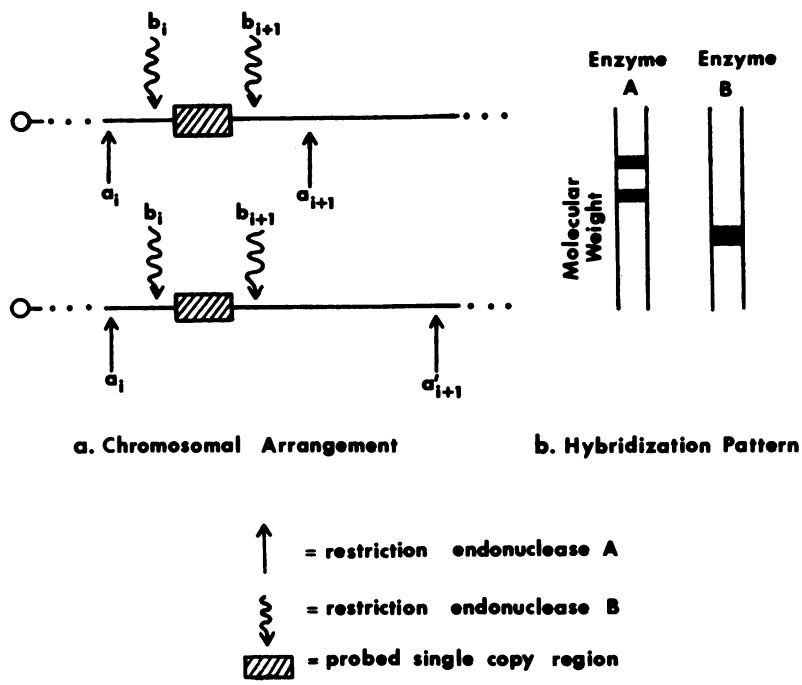

FIG. 1.—*a*, Cuts made in pair of homologous chromosomes by enzyme A and enzyme B; *b*, hybridization pattern of enzymes A and B given cuts of *a*.

Restriction endonuclease fragment polymorphisms have also proven useful in the genetic analysis of chromosomal markers in yeast. Three groups [12–14] discovered RFLPs in their systems of study. Six of eight EcoR1 restriction fragments encoding tyrosine tRNA were mapped using RFLP analysis [13, 15]. An RFLP at a particular tyrosine locus (*SUP4*) was shown to be generated by the presence or absence of a 5.8-kb repetitive DNA segment [14]. By use of a restriction site polymorphism, recombination was examined, and the RFLP was subsequently used to map the DNA locus [12]. RFLP analysis has also been applied to genetic mapping in the nematode worm *Cenorhabditis elegans* (Hirsch, personal communication, 1979).

RFLPs IN HUMAN DNA

More recently, RFLPs have been discovered in human systems. In the region of the human *globin* genes, a heterozygosity for a PstI restriction site has been found within the intervening sequence of the *γ-globin* gene among recombinant DNA clones [16]. Apparently, the two chromosomes of the individual from which the DNA was isolated were heterozygous for this marker. This was confirmed by examination of a Southern transfer of DNA from the individual. HpaI fragment length polymorphic variants associated with sickle-cell traits have also been found [17]. Although most human *β-globin* genes were present on a 7.6-kb HpaI fragment, among individuals of African origin, *β-globin* genes were also found on 7.0- and 13.0-kb HpaI fragments. Of individuals with the *hemoglobin* $+S$ allele, 87% were found to have the 13.0-kb variant fragment. This linked RFLP has been used to diagnose sickle-cell trait in utero [17]. More recently, the abundance of RFLPs associated with the *β-* and *γ-globin* regions have been examined [18, 19]. Among 60 individuals, Jeffreys found one rare and two frequent polymorphisms.

GENETIC MAPPING WITHOUT SPECIFIC GENE ISOLATION

The above RFLPs were recognized by their relationship to the DNA sequence for a gene of interest. These experiments could not have proceeded without prior isolation of the specific gene's DNA. We note, however, that for the more general purpose of mapping genetic loci, RFLPs need not encode the gene of interest, but only be sufficiently nearby to display genetic linkage. This is extremely important, because a recombinant DNA probe need not reveal a restriction fragment containing part of the gene of interest to be useful; that is, one does not have to "isolate the gene" to map it. If a substantial number of polymorphic regions can be identified (and we estimate that 150 will eventually be required), then all genes will be linked to one or another of the regions containing RFLPs and can thereby be mapped.

MAKING THE RFLP MAP

The demonstration of a heritable and detectable RFLP at the *β-globin* locus implies that the genetic mapping scheme based on linkage in family studies of randomly derived RFLPs is possible in principle. It is less clear, however, how difficult it will be to find usefully polymorphic loci and to establish linkage relationships in practice. To address these issues, we must answer a series of questions, many of which can have interdependent and complex answers. Four interrelated questions must be answered:

(1) How many markers are needed? (2) How polymorphic must each marker be? (3) How many families are needed to establish linkage? and (4) How much polymorphism can we expect in the human genome?

For counseling purposes, the number of markers needed and the degree of polymorphism desired are unlimited: more and higher are always better. Although loose linkage to a slightly polymorphic marker is better than no linkage, the tighter the linkage, the more accurately one can predict genotypes, and the higher the degree of polymorphism, the more often one will have informative families necessary for counseling. For the purpose of construction of a genetic linkage map, we can be less demanding and more precise, although here, too, the need for polymorphism and useful marker loci can never be totally fulfilled.

Individuals and lineages spanning multiple generations will be selected from large pedigrees. The restriction fragment fingerprint resulting for each probe, from each individual, will be recorded and analyzed for linkage. The pedigrees under study will already have been typed for HLA, red cell antigens, and isoenzymes. Standard LOD score analysis in pedigrees, explained below, will determine linkages between the probes, and between each probe and other chromosomal markers.

Mostly, we expect RFLPs to be inherited as Mendelian codominant alleles. These might simply be the result of single-base pair changes, although other possibilities can be imagined. Deletions, additions, and other local rearrangements should also manifest Mendelian inheritance. RFLPs within distant translocations, however, will reveal unusual inheritance, since the probe may be detecting sequences that, although homologous, are genetically unlinked.

Another interesting possibility is that RFLPs could result from the activity of DNA modifying enzymes [20]. In this case, pedigree analysis would show us the linkage relationship of the modification genes. Furthermore, these polymorphisms might not give the expected codominant phenotype, but will be expressed as either true dominants or recessives, depending on whether the gene is structural for a modifying enzyme or regulatory; and if regulatory, whether positive or negative. Furthermore, the activity of such a system might affect RFLPs detectable with several probes, possibly known by mapped RFLPs to reflect sequences imbedded in separate linkage groups. In this case again, linkage relationships might yield apparently disparate results. The two described possibilities are distinguished by their dominance characteristics, rearrangements giving codominance, and modifying enzymes being either dominant or recessive. Other, even more complex, possibilities may exist.

Before considering our questions, we must briefly address the issue of how linkage data in humans are analyzed. As RFLPs are identified, they must be examined for a number of statistical properties by pedigree analysis [1–3]. To establish a map of markers, pedigree analysis will be used with qualitative traits (such as allele 1, 2, . . . , n). As diseases are studied later, quantitative traits such as cholesterol, weight, time to an event, or a combination of the above will also be considered. Models are fit by maximizing the likelihood of the parameters given the pedigree structure and the observations. Geneticists compare likelihoods of different models by calculating likelihood ratios. When the likelihood ratio is 1000:1, that is, the odds of one model are 1,000-fold greater than the odds of another, then the first is accepted compared to the

latter. The $\log_{10}$ of the likelihood ratio (LOD score) as described by Morton [21], is usually reported. Thus a LOD score of 1 to 2 is "interesting"; 2 to 3, "suggestive"; and > 3, "proof" of the phenomenon in question.

How Many Markers Are Needed?

To answer this question, we must know how large the human genome is in genetic terms and how far one can be from a marker and still detect linkage. Renwick [4] has estimated that the human genome is 33 Morgans in length. Figure 2, from [22], shows that the expected LOD score for two linked markers in small family structures falls by an order of magnitude between $\theta = .1$ and $\theta = .3$, where θ is the recombination fraction. Therefore, a search for linkage will be very inefficient at greater values of θ. For this discussion, an equivalence between Morgans and recombination fraction is assumed, which is approximately true within this range of values. We will pick $\theta = .2$ as the maximum distance desired between RFLP marker loci which will still allow linkage to be detected. To map a new RFLP, a dominant gene for a disease, or a gene for a carrier trait, RFLPs spaced 0.2 Morgans apart will be sufficient to guarantee that a gene is detectably linked to an RFLP locus at a distance no greater than $\theta = .1$. At this distance, and without further dividing the 33 Morgans into individual chromosome

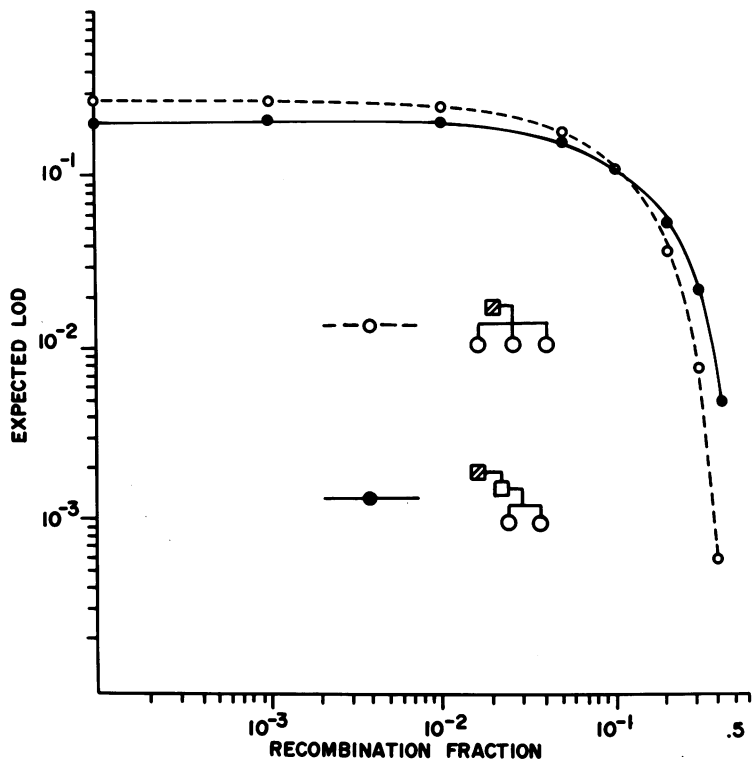

FIG. 2.—Expected LOD score as a function of recombination fraction for two 4-member pedigrees

sizes, about 150 RFLP markers would be required. To create a map, more than 150 polymorphic probes (marker loci) must be isolated and tested and, occasionally, slightly greater than optimal distances will have to be accepted.

How Polymorphic Must Each Marker Be?

To simplify this discussion, we will imagine that we are examining for linkage a pair of loci which we will call the "index" locus and the "marker" locus. We will assume that the index locus contains a rare dominant allele which segregates in the family under study. The allele could be a disease trait or a rare allele of an RFLP locus. The marker locus is a polymorphic locus for which information for linkage studies is to be determined. We assume that all alleles at the marker locus are codominant, which is our expectation for RFLPs. Our assumption of rarity of an allele at the index locus is made only to simplify the discussion and is not a necessity for our analysis. The lineage of the rare allele at the index locus in the family under study is always clear by our assumptions. The question to be answered is: How does the degree of polymorphism at the marker locus influence the probability of detection of linkage to the index locus? Informativeness in this context is represented by the probability that a given offspring of a parent carrying the rare allele at the index locus will allow deduction of the parental genotype at the marker locus. Table 1 gives the mating types, their frequencies, and the probability of informativeness of an offspring for a marker locus, A_i, with n alleles and frequencies p_i.

TABLE 1

MATING TYPES, PROBABILITY THAT OFFSPRING ARE INFORMATIVE, AND FREQUENCY OF EACH TYPE FOR THE LINKAGE MODELS

Genotype of affected parent	Genotype of unaffected mate	Probability offspring informative	Frequency of mating
A_iA_j	A_kA_l	1.0	$2p_i\,p_j\,(1 - p_i - p_j)^2$
A_iA_j	A_kA_i or A_kA_j	1.0	$2p_i\,p_j\,(1 - p_i - p_j)(2p_i + 2p_j)$
A_iA_j	All homozygotes	1.0	$2p_i\,p_j \sum_{i=1}^{n} p_i^2$
A_iA_j	A_iA_j	.5	$2p_i\,p_j \times 2p_i\,p_j$
All homozygotes	All genotypes	0	p_i^2

One can evaluate each marker locus for its polymorphism information content (PIC) by summing in table 1 the mating frequencies multiplied by the probability that an offspring will be informative. Under our assumptions, the expected value of PIC can be calculated as

$$1 - \left(\sum_{i=1}^{n} p_i^2\right) - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$$

A similar index can be found in [49].

Both index and marker loci are considered in nuclear families, ignoring both the advantages and complexities of multigenerational studies which allow one to know the coupling or repulsion of disease and marker loci and, on occasion, the genotypes for

loci with recessive or silent alleles. Multigenerational studies can increase the amount of information which can be derived from a marker locus, but they should not alter our basic impression of the value of a locus for linkage. Complications in the genetic model can make detection of linkage much more difficult and can reduce the PIC. Table 2 presents the PIC and number of alleles for 29 established codominant chromosomal markers. The calculations use gene frequencies in Caucasians. Under this formulation, there are four codominant loci which are highly informative (PIC > .5), fifteen are reasonably informative (.5 > PIC > .25), and ten are only slightly informative (PIC < .25). Loci with many alleles and a PIC near 1 are most desirable. Such loci will probably be essential if the kind of mapping method we propose is to become a practical reality.

How Many Pedigrees?

The number of informative families required can only be estimated. The precise pedigree structure and family size are relevant in determining this number. An analysis of this problem [22] has shown that the amount of information in a family study varies considerably with the precise structure of the pedigree; Thompson et al. quantified two

TABLE 2

PIC OF 29 CODOMINANT HUMAN CHROMOSOME MARKER LOCI IN CAUCASIANS

Locus	PIC	No. alleles	Population	No. subjects	Source
HLA-A	.976	12	European	···	[23]
HLA-B	.976	16	European	···	[23]
MNSs	.638	4	English	1000	[24]*
Rh	.580	6	English	154	[25]*
AP	.448	4	English	597	[26]
TC II	.400	4	Caucasian	131	[27]
GPT	.382	3	Caucasian	528	[28]
Jk	.375	2	English	201	[29]*
P	.375	2	English	1166	[30]*
Fy	.368	2	English	1166	[30]*
Hapt	.367	2	English	218	[31]
GLO	.363	2	S.W. German	169	[32]
Pg	.359	2	Caucasian	931	[33]
CDA	.352	2	Caucasian	189	[34]
Bf	.342	3	Caucasian	158	[35]
Gc	.325	2	English	49	[36]*
PGM_3	.308	2	English	583	[37]*
PGM_1	.286	6	English	2115	[38]*
C3	.277	2	Norwegian	400	[39]
Amy_2	.171	2	Caucasian-U.S.	673	[40]
ESD	.162	2	European	867	[41]
ADA	.089	2	English	1353	[42]*
GALT	.087	2	S.W. German	195	[43]
UMPK	.086	3	Caucasian	386	[44]
KELL	.084	2	British	8767	[45]*
AK_1	.082	2	British-N. Ireland	1887	[46]*
Lu	.058	2	English	1166	[30]*
6PGD	.042	3	English-London	4558	[47]*
PEPA	.002	3	British	2283	[48]*

* Data compiled from: Mourant AC, Kopec AC, Domaniewska-Sobczak: *The Distribution of the Human Blood Groups and Other Polymorphisms*, 2nd ed. London, Oxford Univ. Press, 1976

well-known principles: multigenerational pedigrees are more useful than nuclear families, and large families give considerably more information than smaller families. The expected LOD score also varies with the recombination fraction. Figure 3 (from [22]) shows that an expected LOD score of 0.1 spans both the range of reasonable sibship sizes (2 to 8) and recombination fractions .3 to .0001. To establish a LOD score of 3.0, we would need 30 of these units; this translates into 300 individuals to establish a linkage at $\theta = .3$, and as few as 21 individuals to establish linkage in three 5-child families where tight linkage between the index and marker loci exists. These numbers can be decreased further if the nuclear families are linked in a multigenerational pedigree. The advantages of large genealogies can also be exploited further if the genealogy of untested family members is known and principles of sequential sampling are employed to determine which members should be studied. One excellent source of large pedigrees of known genealogical structure is the computerized data base created by the genetics group at the University of Utah, a file that contains seven generations of descendants of the Utah Mormon pioneers, who, because of large family sizes, often have descendants numbering in the thousands [50, 51]. This data base is also useful in the search for new markers in pedigrees where the known markers have been assayed. The numbers cited above (15 to 300 individuals) must be multiplied by a function of the reciprocal of the PIC of the marker locus used. For this reason, we are initiating our search for RFLPs in large Mormon pedigrees with genetic defects, which are already being analyzed for known chromosomal markers.

The above discussion emphasizes the interplay between sample size, degree of polymorphism of each marker locus, and recombination distance between the loci in question. In summary, we expect that a set of about 150 markers, each 0.2 Morgans apart, will be required, and that if their average information content (PIC) is .5, several

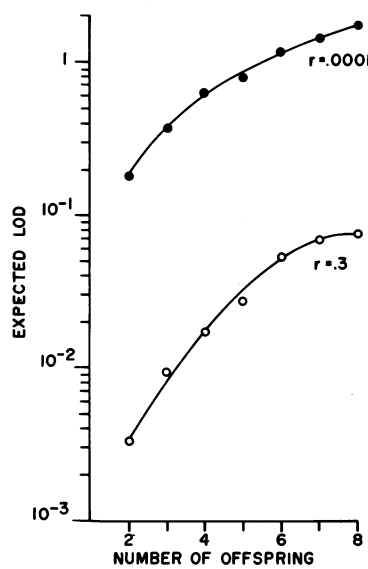

FIG. 3.—Expected LOD score for two recombination fractions (r) as a function of sibship size

hundred individuals, at most, will be necessary to establish linkage to any given index locus. During the process of mapping RFLPs, we need to make more extensive pedigree studies using RFLPs which are less polymorphic. If the number of subjects available for study is limited, a larger set of markers could allow one to reduce the number of individuals required. An expected LOD score of 3 will result from about 6,000 analyses (individuals × markers) over a wide range of marker densities [52].

The most efficient procedure in establishing RFLPs will be to study a small set of large pedigrees which have been genotyped for all known polymorphic markers (red cell antigens, isoenzymes, HLA, and chromosome-banding polymorphisms). All probes showing usable PIC values will be studied further to establish segregation and define linkage groups. RFLP markers will also be studied further for linkage with known markers. As the study develops, a set of RFLPs denoted R_{ijk} will emerge, where i denotes the linkage group (chromosome, if known); j, the locus within the chromosome; and k, the specific fragment at the locus. Many fragments at a locus made with different restriction enzymes will be investigated. In addition, neighboring clones will be studied to expand polymorphism to an amount sufficient for linkage studies.

Before RFLPs become a useful tool for mapping disease loci, the number of members of set R will be reduced. Some R_{ijk} will be unnecessary for initial screening, and efficiency dictates that a primary set, R', be established, eliminating some of the j sites, and for each site, some of the fragments. If linkage is not detected, the secondary set, R'', containing additional fragments at R' loci and additional loci, will be sequentially examined until linkage is established.

Once linkage is suspected, neighboring loci will be explored until the closest locus is found. One may expand the set of fragments at this locus by seeking neighboring fragments to increase polymorphism.

For some counseling purposes, we wish to find the closest RFLP to the disease locus which has unique alleles in both parents. Another possible approach is to try to find two RFLP loci which flank the disease locus on the genetic map. Such flanking markers will greatly increase the certainty of genotype determinations used as the basis for counseling.

Estimates of the Frequency of Polymorphism in DNA

The above considerations lead us to the question of how much polymorphism to expect in human DNA. Although little is known about the frequency of DNA polymorphism in humans, several sources of information can be used to provide order of magnitude estimates, including determinations of base sequence divergence based on melting temperature depressions of reannealed DNA, and measurements of amino acid sequence divergence among proteins.

Estimates from the frequency of protein polymorphism. The frequency of polymorphism among 71 protein studies in European populations [53] was 0.28, suggesting that the minimum DNA sequence polymorphism must be at least 0.28 base changes per 1,000 base pairs, the size of the average gene. Assuming that visible electromorphs represent about one-third of the nucleic acid sequence changes and that many third-base changes are not reflected in corresponding amino acid changes, we can

estimate the frequency of DNA sequence polymorphism to be about 0.001 per base pair among protein-coding sequences. This sets a lower limit for the amount of base-sequence polymorphism, since sequences found as mRNA have been shown to diverge more slowly than the bulk of the single-copy DNA [54].

Estimates from the depression of the melting temperature of reassociated DNA. Divergence of nucleic acid sequence between related species has also been determined by measurement of the relative depression of the denaturation temperature of DNA that has been melted and reassociated with DNA from a heterologous species. However, only one attempt has been made to determine the extent of polymorphism within a mammalian species by heteroannealing DNA from separate individuals of the same species [55]. This study indicates that the DNA-sequence polymorphism due to single-base changes between an individual wild mouse and an inbred C57B1 mouse is low. In fact, no melting temperature depression was observed, from which one can conclude that the resultant heterozygosity was probably less than .5% (Britten, personal communication, 1979), corresponding to a DNA polymorphism frequency of 1.1%. This conversion from degree of heterozygosity to degree of polymorphism is approximate and conservative; approximate in that it uses a formula [56] which assumes no selection, and conservative in that we have defined polymorphism as requiring a minor allele with a frequency of at least 10%, as less polymorphic loci are not of interest for our purposes. This suggests an upper limit to the extent of DNA polymorphism in mammals due to base-pair changes.

RFLP calculation. Assuming an average of the above estimates of polymorphism for human DNA due to base-pair changes, we can calculate the probability that a restriction fragment will be polymorphic: If we let p be the fraction of nucleotide sites showing significant polymorphism (frequency of the minor allele greater than 10%); n, the number of nucleotides in the restriction enzyme site; and S, the probability that a given restriction site is not polymorphic, then $S = (1 - p)^n$. Taking into account such parameters as average restriction fragment length, AT to GC ratio, and appearance of new sites as well as disappearance of old sites, Upholt [57] was able to show that f, the fraction of restriction fragments showing no polymorphism, will be approximately equal to S^2 $(2 - S)$. If we assume a nucleotide site polymorphism of 1.1%, the upper limit suggested by the mouse study described above, then the probability that a restriction fragment defined by a four-base enzyme would also be polymorphic is 12.3%. Using the lower limit for polymorphism suggested by the protein electromorph studies, 0.23, the probability of restriction fragment polymorphism becomes 0.34%. The corresponding probabilities for fragments defined by restriction enzymes which recognize six-base sequences are 17.7% and 0.50%. Even though the probability per fragment of polymorphism is somewhat less for the fragments defined by the four-base enzymes, they should be more useful in revealing polymorphism due to single-base changes in DNA, since there are approximately eight times as many base fragments produced per unit length of DNA.

Polymorphism due to insertions and deletions. Another possible source of polymorphism other than single-base-pair changes is the insertion or deletion of blocks of DNA. These DNA polymorphisms would be underrepresented in the above estimates, but will serve to increase the amount of DNA polymorphism. In yeast and

Drosophila, RFLPs have been discovered which are due to such alterations. In these species, the mobile DNA segments are moderately repetitive [14, 58]. The frequency in humans of DNA polymorphism from this source cannot yet even be estimated. Insertion and deletion of such mobile DNA segments could, however, also be a major mechanism for the generation of restriction fragment length variation in humans. The change in mobility of a restriction fragment due to insertion or deletion is unpredictable in advance. Insertions would increase the length of a restriction fragment if the inserted segment does not contain a restriction site or, possibly but not necessarily, decrease the fragment length if it does contain a restriction site. Similarly, deletions completely internal to the restriction fragment will decrease its length, whereas deletions, including one of the restriction sites, could either increase or decrease the fragment length depending on the size of the deletion and location of the next nearest restriction site. Deletion of the entire probed sequence would not be revealed and would have a recessive phenotype. The most diagnostic feature of insertions and deletions would be their pleiotropic effect on some restriction fragments generated by several restriction enzymes.

If mobile, repetitive DNA segments were found in human DNA, these could serve as a probe for loci likely to show polymorphism. The mobile sequence could be cloned and used to screen the human genomic library. Isolation and analysis of the sister DNA segment clones should reveal the presence of adjacent, single-copy sequences. These then could be used as probes of DNA from individuals to determine whether the repetitious segment is mobile from the new locus, thereby defining a new polymorphic site. Although the number of fragments examined is the important parameter in discovering RFLPs due to single-base changes, the length of DNA screened is the most important factor in the case of insertions and deletions, since many different restriction fragments, defined by different enzymes, will reveal the same insertion or deletion.

SOURCE OF RECOMBINANT DNA PROBES DEFINING RFLP LOCI

To find polymorphisms, we must be able to identify specific and unique regions of the genome. The recent development of recombinant DNA technology permits the cloning in *E. coli* of individual segments of the human genome. These DNAs can be labeled with radioisotopes to high specific activities and used as hybridization probes of DNA from individuals, being digested with various restriction fragments from specific regions.

Two libraries of human recombinant DNA have been constructed by two quite different methods, each with properties relevant to this purpose. One library was constructed from the DNA of the entire human genome [17], using a bacteriophage vector. The inserted segments in these recombinant phage range in size from 15 to 20 kb. A second library of human DNA sequences was constructed from messenger RNA (L. Villa-Komaroff, personal communication, 1978). To prepare this human library, messenger RNA was enzymatically copied into DNA (cDNA) inserted into a plasmid vector. These segments of messenger sequences are each about 500 base pairs long. The phage library contains segments representing the DNA sequences of the entire human genome, and the cDNA library reflects those human sequences transcribed into messenger RNA, providing an opportunity to characterize each set respecting polymorphism.

The phage clones with their large inserts will reveal many more restriction fragments than the much smaller inserts of the cDNA clones. Most of the long segments, however, are likely to contain interspersed repetitive sequences [59] which would hybridize to fragments from many different regions of the genome. The cDNA clones should mostly reflect gene sequences present in only one copy (or a few, see below) per haploid genome, but many represent DNA sequences which are more highly conserved and reveal fewer polymorphisms.

The liability of the large phage probes is that they generally will contain moderately repetitious sequences interspersed within single copy. There is evidence that the human genome, like most eukaryotic genomes, is organized with moderately repetitious sequences interspersed within the single-copy sequences [59]. Since the mid-repeat DNAs will hybridize to many fragments, presumably from many parts of the genome, the pattern of the Southern transfer becomes too complex for ready interpretation. However, the available evidence suggests that not all the single-copy human DNA contains short-period interspersed mid-repeat sequences; a fraction of the single-copy DNA should contain quite long (more than 10 kb) segments of DNA with no interspersed mid-repeat material. Preliminary results indicate that a reasonable proportion (1%–3%) of the phage clones contains no mid-repeat sequences, and the clones are usable as probes without further alteration [60]. These were discovered by screening the phage library with radiolabeled mid-repeat human DNA, selecting clones which do not hybridize.

Figure 4 shows that a locus can be substantially expanded by using the initial recombinant DNA sequence to probe the phage library for adjacent sequences, from which a new composite probe can be constructed. The procedure is called "walking" and consists of identifying successive, adjacent clones of recombinant DNA, orienting them by their restriction enzyme site pattern with the primary clone, and excising the mid-repeat sequences as indicated above. Additional steps can be taken, the new sequences each having their associated polymorphisms, until the desired degree of polymorphism is achieved. The combined set of subcloned, single-copy-only fragments from a specified longer primary sequence, containing no mid-repeats, constitutes a usable probe for that region.

REFINING MODELS OF INHERITANCE USING THE RFLP MAP

The resolution of genetic and environmental components of disease has evolved in the last decade into a well-formed discipline of genetic epidemiology. While the methods of analysis are still widely disputed [61, 62], many of the problems at least are agreed upon. A description of the genetic epidemiology of a disease must involve unraveling the model for the underlying genetic predisposition (endogenous factors); understanding the environmental contributions (exogenous factors); including the effects of chance, development, and age; and, finally, understanding the variability of expression of the phenotype.

In principle, linked marker loci can allow one to establish, with high certainty, the genotype of an individual and, consequently, assess much more precisely the contribution of modifying factors such as secondary genes, penetrance, and environment. Tightly linked markers (such as HLA) have been used to clarify genetic models

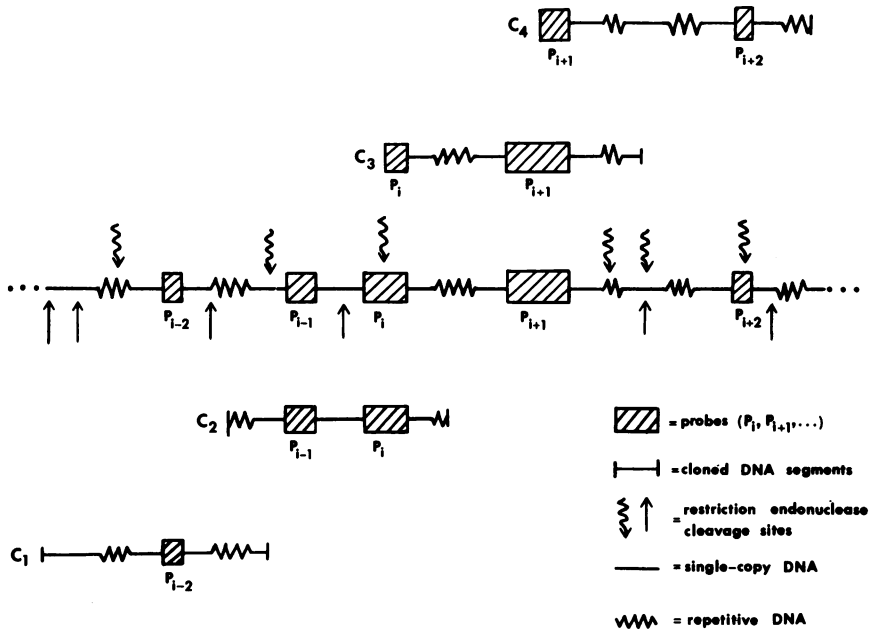

FIG. 4.—Expansion of probe p_i with neighboring probes p_{i-2}, p_{i-1}, p_{i+1}, and p_{i+2}. Probes are diagrammed in an intact DNA region with restriction cleavage sites indicated and also in homologous cloned DNA segments, C_1, C_2, C_3, and C_4, which may contain parts of some probes.

and improve recurrence risks [5]. This analysis simultaneously established the "recessive" genetic model and demonstrated linkage to HLA. Furthermore, studies of additional pedigrees failed to demonstrate heterogeneity in the etiology of the disease and suggested a disease frequency approximately 40 times greater than previously estimated [63]. The analysis also showed that one large, informative pedigree gave twice the information available in nine smaller ones.

However, since not all disease loci can be expected to be found in the HLA region, and since other markers, only slightly polymorphic, would not have allowed the discrimination necessary to resolve the genetic system, this approach has been severely limited. The intrinsic liabilities to the marker systems are compounded by heterogeneity, late age of onset, variable expressivity, environmental interactions, and other complications found associated with many important diseases. In fact, known markers have failed to provide linkage to important disease loci (such as Huntington disease). Because of these difficulties, only 2%–3% of known Mendelian disorders have been mapped.

As an alternate approach, substantial progress has been and is being made in determining the chromosomal location of many genes by the method of segregating human chromosomes from human:mouse fusion hybrids (see [64] for review). However, to map disease loci, this method depends upon knowledge of the biochemical defects at the cellular level. Since the biochemical basis for most genetic disease remains obscure, their loci are refractory to localization by the fusion approach or its variations.

DISCUSSION

The application of a set of probes for DNA polymorphism to DNA available to us from large pedigrees should provide a new horizon in human genetics. Recent applications of these methods, specifically to hemoglobinopathies [18, 19], have already provided major results. Once established, DNA polymorphisms will be useful for in utero assays, as has been shown for hemoglobin [18, 65]. This is important since, even when successful, current linkage studies provide markers which are generally unusable for preventive medicine because they cannot be assayed in utero, or because of only marginally informative polymorphism. With linkage based on DNA markers, parents whose pedigrees might indicate the possibility of their carrying a deleterious allele could determine prior to pregnancy whether or not they actually carry the allele and, consequently, whether amniocentesis might be necessary. Analogously, individuals potentially at risk for heritable disorders, including some forms of cancer, of which expression occurs late in life, ought to be able to determine in advance of obvious symptoms whether they are at risk for the disease. Moreover, when a specific linkage is found, further refinements should be able to increase the available polymorphism and tighten the linkage until a valid tool for preventive medicine is established.

A large set of RFLPs will have many other uses in human population genetics. A large set of truly neutral markers will reopen questions of genetic distance between geographical isolates and may shed more light on population structure and selection as competing mechanisms of allelic variation and linkage disequilibrium. Multilocus theory will have abundant data to test its predictions.

Human cytogenetic analysis will also benefit from this new technology. Often birth defects phenotypically resemble known chromosomal deletions but show no missing bands on the suspect chromosome. With a map of RFLPs, one could look for areas where the alleles of one parent were entirely missing and infer a deletion too small to be detected by standard techniques. Similarly, trisomies would produce unique patterns which could reveal small translocations.

ACKNOWLEDGMENTS

We are indebted to numerous individuals for valuable suggestions during the development of this project. We wish to thank the graduate students and faculty who participated in the Alta genetics retreat, April, 1978, for their part in the discussions which originated this project. Specifically, we would like to acknowledge Maurice S. Fox, Raymond F. Gesteland, D. Timothy Bishop, Arlene Wyman, and Tim Helentjaris.

REFERENCES

1. ELSTON RC, STEWART J: A general model for the genetic analysis of pedigree data. *Hum Hered* 21:523–542, 1971
2. CANNINGS C, SKOLNICK MH, DE NEVERS K, SRIDHARAN R: Calculation of risk factors and likelihoods for familial diseases. *Comput Biomed Res* 9:393–407, 1976
3. CANNINGS C, THOMPSON EA, SKOLNICK M: Probability functions on complex pedigrees. *Adv Appl Prob* 10:26–61, 1978
4. RENWICK JH: Progress in mapping human autosomes. *Br Med Bull* 25:65–73, 1969
5. KRAVITZ K, SKOLNICK M, CANNINGS C, ET AL.: Genetic linkage between hereditary hemochromatosis and HLA. *Am J Hum Genet* 31:601–619, 1979

6. NATHANS D, SMITH H: Restriction endonucleases in the analysis and restructuring of DNA molecules. *Annu Rev Biochem* 44:273–293, 1975
7. SOUTHERN EM: Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* 98:503–517, 1975
8. JEFFREYS A, FLAVELL R: A physical map of the DNA regions flanking the rabbit β-globin gene. *Cell* 12:429–439, 1977
9. GRODZICKER T, WILLIAMS J, SHARP P, SAMBROOK J: Physical mapping of temperature-sensitive mutations of adenoviruses. *Cold Spring Harbor Symp Quant Biol* 39:439–446, 1974
10. HUTCHINSON C III, NEWBOLD J, POTTER S, EDGELL M: Maternal inheritance of mammalian mitochondrial DNA. *Nature* 251:536, 1974
11. POTTER S, NEWBOLD J, HUTCHINSON C III, EDGELL M: Specific cleavage analysis of mammalian mitochondrial DNA. *Proc Natl Acad Sci USA* 72:4496–4500, 1975
12. PETES TK, BOTSTEIN D: Simple Mendelian inheritance of the reiterated ribosomal DNA of yeast. *Proc Natl Acad Sci USA* 74:5091–5095, 1977
13. GOODMAN H, OLSON M, HALL B: Nucleotide sequence of a mutant eukaryotic gene: the yeast tyrosine inserting ochre suppressor SUP4-0. *Proc Natl Acad Sci USA* 74:5453–5457, 1977
14. CAMERON JR, LOH EY, DAVIS RW: Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell* 16:739–751, 1979
15. OLSON MV, HALL BD, CAMERON JR, DAVIS RW: Cloning of the yeast tyrosine transfer RNA genes in *Bacteriophage lambda*. *J Mol Biol* 127:285–295, 1979
16. MANIATIS T, HARDISON R, LACY E, ET AL.: The isolation of structural genes from libraries of eucaryotic DNA. *Cell* 15:687–701, 1978
17. KAN Y, DOZY A: Antenatal diagnosis of sickle-cell anaemia by DNA analysis of amniotic-fluid cells. *Lancet* 2:910–912, 1978
18. JEFFREYS AF: DNA sequence variants in the Gγ-, Aγ-, σ- and β-globin genes of man. *Cell* 18:1–10, 1979
19. TUAN D, BIRO PB, DE RIEL JK, LAZARUS H, FORGET BG: Restriction endonuclease mapping of the human γ globin gene loci. *Nucleic Acids Res* 6:2519–2544, 1979
20. BIRD AP: Use of restriction enzymes to study eukaryotic DNA methylation. II. The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J Mol Biol* 118:49–60, 1978
21. MORTON NE: Segregation and linkage, in *Methodology in Human Genetics*, edited by BURDETTE WJ, San Francisco, Holden Day, 1962, pp 17–52
22. THOMPSON EA, KRAVITZ K, HILL J, SKOLNICK M: Linkage and the power of a pedigree structure, in *Genetic Epidemiology*, edited by MORTON NE, CHUNG CS, New York, Academic Press, 1978, pp 247–253
23. THOMAS G, BODMER WF, BODMER J: The HLA system as a model for studying the interaction between selection, migration and linkage, in *Population Genetics and Ecology*, edited by KARLIN S, NEVO E, New York, Academic Press, 1976, pp 465–497
24. CLEGHORN TE: *MNSs* gene frequencies in English blood donors. *Nature* 187:701, 1960
25. RACE RR, TAYLOR GL, CAPPELL DF, MCFARLANE MN: Recognition of a further common Rh genotype in man. *Nature* 153:52–53, 1944
26. ROBSON EB, HARRIS H: Further studies on the genetics of placental alkaline phosphatase. *Ann Hum Genet* 30:219–232, 1967
27. DAIGER SP, LABROWE ML, PARSONS M, WANG L, CAVALLI-SFORZA LL: Detection of genetic variation with radioactive ligands. III. Genetic polymorphism of transcobalamin II in human plasma. *Am J Hum Genet* 30:202–214, 1978
28. CHEN S-H, GIBLETT ER, ANDERSON J, FOSSEN BL: Genetics of glutamic pyruvate transaminase: its inheritance, common and rare variants, population distribution and differences in catalytic activity. *Ann Hum Genet* 35:401–409, 1972
29. RACE RR, HOLT HA, THOMPSON JS: The inheritance and distribution of the Duffy blood groups. *Heredity* 5:103–110, 1951

30. MOURANT AE, KOPEĆ AC, DOMANIEWSKA-SOBCZAK: *The Distribution of Human Blood Groups and Other Polymorphisms*, 1st ed. London, Oxford Univ. Press, 1954
31. GIBLETT ER: *Genetic Marker Human Blood*, 1st ed. Philadelphia, FA Davis, 1969
32. KÖMPF J, BISSPORT S, GUSSMAN S, RITTER H: Polymorphism of red cell glyoxalase I (EC: 4,4,1,5). A new genetic marker in man. *Hum Genet* 27:141–143, 1975
33. SAMLOFF IM, TOWNES PL: Pepsinogens: genetic polymorphism in man. *Science* 168:144–145, 1970
34. TENG Y-S, ANDERSON JE, GIBLETT ER: Cytidine deaminase: a new genetic polymorphism demonstrated in human granulocytes. *Am J Hum Genet* 27:492–497, 1975
35. ALPER CA, BOENISCH T, WATSON L: Genetic polymorphism in human glycine-rich beta-glycoprotein. *J Exp Med* 135:68, 1972
36. HIRSCHFELD J: The Gc-system. *Prog Allergy* 6:155–186, 1962
37. HOPKINSON DA, HARRIS H: A third phosphoglucomutase locus in man. *Ann Hum Genet* 31:359–367, 1968
38. HOPKINSON DA, HARRIS H: Rare phosphoglucomutase phenotypes. *Ann Hum Genet* 30:167–181, 1966
39. TEISBERG P: High voltage agarose gel electrophoresis in the study of C_3 polymorphism. *Vox Sang* 19:47–56, 1970
40. MERRIT DA, RIVAS M, BAXLER D, NEWELL R: Salivary and pancreatic amylase: electrophoretic characterizations and genetic studies. *Am J Hum Genet* 25:510–520, 1973
41. HOPKINSON DA, MESTRINER MA, CORTNER J, HARRIS H: Esterase D: a new human polymorphism. *Ann Hum Genet* 37:119–137, 1973
42. HOPKINSON DA, COOK PJL, HARRIS H: Further data on the adenosine deaminase (ADA) polymorphism and a report of a new phenotype. *Ann Hum Genet* 32:361–367, 1969
43. BISSPORT S, KÖMPF J: Red cell galactose-1-*p*-uridyl transferase—formal genetics and linkage relations. *Humangenetik* 18:93–94, 1973
44. GIBLETT ER, ANDERSON J, CHEN S-H, TENG Y-S, COHEN F: Uridine monophosphate kinase: a new genetic polymorphism with possible clinical implications. *Am J Hum Genet* 26:627–635, 1974
45. CLEGHORN TE: The occurrence of certain rare blood group factors in Britain. PhD thesis, England, Univ. of Sheffield, 1961
46. RAPLEY S, ROBSON EB, HARRIS H, MAYNARD SMITH S: Data on the incidence, segregation, and linkage relations of the adenylate kinase (AK) polymorphism. *Ann Hum Genet* 31:237–242, 1967
47. CARTER ND, FILDES RA, FITCH LI, PARR CW: Genetically determined electrophoresis variations of human phosphogluconate dehydrogenase. *Acta Genet Stat Med (Basel)* 18:109–122, 1968
48. LEWIS WHP, HARRIS H: Human red cell peptidases. *Nature* 215:351–355, 1967
49. CHAKRABORTY R, FUERST PA, FERRELL RE: Potential information in family studies of linkage, in *Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Disease*, edited by SING CF, SKOLNICK M, New York, Alan R. Liss, 1979, pp 297–303
50. SKOLNICK M: Prospects for population oncogenetics, in *Genetics of Human Cancer*, edited by MULVIHILL JJ, MILLER RW, FRAUMENI FJ JR, New York, Raven Press, 1977, pp 19–25
51. WILLIAMS R, SKOLNICK M, CARMELLI D, ET AL.: Utah pedigree studies: design and preliminary data for premature male CHD deaths, in *The Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Heart Disease*, edited by SING CF, SKOLNICK M, New York, Alan R. Liss, 1979, pp 711–732
52. BISHOP DT, SKOLNICK M: Numerical considerations for mapping DNA polymorphic markers, in *Proceedings Conference on Human Health Data from Defined Populations, Banbury Center, Cold Spring Harbor Laboratory, October, 1979*, Banbury Report No. 4, Cold Spring Harbor Laboratory. In press, 1980
53. HARRIS H, HOPKINSON D: Average heterozygosity per locus in man: an estimate based on

the incidence of enzyme polymorphisms. *Ann Hum Genet* 36:9–20, 1972
54. ROSBASH M, CAMPO M, GUMMERSON K: Conservation of cytoplasmic poly(A)-containing RNA in mouse and rats. *Nature* 258:682–686, 1975
55. BRITTEN R, CETTA A, DAVIDSON E: The single-copy DNA sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus*. *Cell* 10:509–519, 1977
56. LEWONTIN RC: *The Genetic Basis of Evolutionary Change*. New York, Columbia Univ. Press, 1974
57. UPHOLT W: Estimation of DNA sequence divergence from comparison of restriction endonuclease digests. *Nucleic Acids Res* 4:1257–1265, 1977
58. RUBIN GM: Isolation of a telomeric DNA sequence from *Drosophila melanogaster*. *Cold Spring Harbor Symp Quant Biol* 42:1041–1046, 1978
59. SCHMID C, DEININGER P: Sequence organization of the human genome. *Cell* 6:345–358, 1975
60. WYMAN A, WHITE RL: Restriction fragment length polymorphism in human DNA. In preparation
61. MORTON NE, CHUNG CS, EDS.: *Genetic Epidemiology*. New York, Academic Press, 1978
62. SING CF, SKOLNICK M, EDS.: *Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Disease*. New York, Academic Press, 1978
63. CARTWRIGHT GE, EDWARDS CQ, KRAVITZ K, ET AL.: Hereditary hemochromatosis: phenotypic expressions of the disease. *N Engl J Med* 301:175–179, 1979
64. RUDDLE F, MCKUSICK V: The status of the gene map of the human chromosomes. *Science* 196:390–405, 1977
65. ORKIN S, ALTER B, ALTAY C, ET AL.: Application of endonuclease mapping to the analysis and prenatal diagnosis of thalassemias caused by globin-gene deletion. *N Engl J Med* 299:166–171, 1978