Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 7353–7357, October 1986
Genetics

Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms

(genetic heterogeneity/human linkage mapping/lod score)

ERIC S. LANDER[†‡§] AND DAVID BOTSTEIN[‡]

[†]Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142; [‡]Massachusetts Institute of Technology, Cambridge, MA 02139; and [§]Harvard University, Cambridge, MA 02138

Contributed by David Botstein, April 28, 1986

ABSTRACT **Simple single-gene disorders in humans can be genetically mapped by using traditional methods of linkage analysis and increasingly abundant restriction fragment length polymorphisms (RFLPs). Many human diseases and traits, however, can be expected to be genetically heterogeneous (i.e., caused by any one of several genes), and traditional linkage analysis is much less effective in such circumstances. We present two methods, *interval mapping* and *simultaneous search*, designed to exploit the full power of a linkage map of the DNA markers. For the simplest situations, only ⅓ as many affected families are needed to map a heterogeneous trait by using these methods. Only ⅕–1/50 as many are needed to detect that genetic heterogeneity is present.**

Since the idea was proposed (1), the use of DNA restriction fragment length polymorphisms (RFLPs) as genetic markers in human linkage studies has become common. A number of diseases that display simple Mendelian inheritance, but whose molecular etiology is unknown, have been genetically shown to be closely linked to RFLP loci and thereby localized to specific chromosomal regions. These include the autosomal dominant diseases Huntington disease (2) and polycystic kidney disease (3), the autosomal recessive cystic fibrosis (4–7), and the chromosome X-linked recessive Duchenne muscular dystrophy (8). In each case, randomly chosen RFLP probes were tested one at a time for linkage to the disease, using traditional methods (9, 10).

Many, perhaps most, human diseases and biologically interesting traits, however, show more complex modes of transmission, including genetic heterogeneity, variable penetrance, polygenic inheritance, and altogether noninherited forms. Potential examples range from familial cancers and ataxia–telangiectasia to genetic forms of alcoholism and psychological disorders, if indeed any exist. Success in the case of simple Mendelian inheritance raises the hope that the RFLP approach can be used to elucidate these traits as well. Unfortunately, traditional single-marker methods are inefficient for analyzing such complicated patterns of inheritance.

Our purpose here is to explore an alternative approach: using a complete RFLP linkage map of the human genome. The construction of such a map is feasible (11, 12) and is already well underway (13, 14). By exploiting the full power of such a complete RFLP linkage map, we believe we may significantly reduce the number of families required for studying complex traits.

In this paper, we consider genetic heterogeneity. If a phenotype can be caused by mutations at any of several loci, it is said to be *genetically heterogeneous*. In well-studied organisms, such as the bacterium *Escherichia coli*, the yeast *Saccharomyces cerevisiae*, the nematode *Caenorhabditis elegans*, and the fruit fly *Drosophila melanogaster*, many phenotypes are genetically heterogeneous. Humans will likely be no different: *in vitro* complementation of cell lines suggests that xeroderma pigmentosum may be caused by mutations in as many as nine loci and ataxia–telangiectasia in as many as five (15).

To see why genetic heterogeneity confounds single-marker linkage analysis, consider a marker at a recombination fraction θ from a locus responsible for a fraction α of all occurrences of a heterogeneous trait. In the overall population, the chance that the marker will fail to cosegregate with the trait through a meiosis is the "apparent" recombination fraction $\theta' = \theta\alpha + \frac{1}{2}(1 - \alpha)$. Linkage will appear to be loose if α is small, even though θ may be small. Detecting loose linkage requires many more observations than for tight linkage, since loose linkage more closely resembles the null hypothesis of nonlinkage. Moreover, even once linkage has been detected, there remains the thorny problem of disentangling the similar effects of high θ and low α to obtain accurate estimates of these quantities: this is necessary both for locating the linked trait-causing gene and for testing whether the trait is actually heterogeneous ($\alpha < 1$). The only distinction between close linkage to a heterogeneous trait and correspondingly more distant linkage to a homogeneous trait is that in the former case apparent crossovers will be preferentially clustered in a certain fraction α of the families examined (16, 17). Detecting this clustering can require many, large pedigrees.

Neglecting to take account of even a modest degree of heterogeneity can result in missing a linkage entirely. For example, a trait-causing locus that accounts for 60% of all cases could lie within 1% of a marker and yet still be "excluded" by linkage analysis from a region of about 20% recombination fraction around the marker, if we were to assume (as is often done) that the trait is homogeneous.

We explore here two strategies designed to exploit the full power of an RFLP map to overcome these obstacles:

(i) Interval mapping. With a map, we may test whether a putative locus lies in an interval of known size between two adjacent markers. This is a more demanding hypothesis—and thus one easier to test—than whether the locus is linked to a single marker at an unknown distance.

(ii) Simultaneous search. Mapping just one of the loci causing a heterogeneous trait is inherently inefficient: the "signal" in cases due to the locus is swamped by the "noise" due to families segregating an unlinked locus. If we instead examine the several trait-causing loci simultaneously, we can extract a stronger, clearer "signal": in every family, at least one of the loci will appear to cosegregate with the disease.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: RFLP, restriction fragment length polymorphism; cM, centimorgan.

DEFINITIONS AND ASSUMPTIONS

Mathematical Preliminaries. Comparing hypotheses in human genetics is typically done as follows (9, 10). Given a pedigree, let X be the set of outcomes for the segregation in the pedigree of the markers of interest (say, a disease and an RFLP to be tested for linkage) and let f_1 and f_2 be alternative probability distributions on X (corresponding, say, to linkage at 10% and nonlinkage). Each observation $x \in X$ yields odds ratio $f_1(x)/f_2(x)$ in favor of f_1 over f_2. We make observations until the product of the odds ratios or, more conveniently, the sum of the $\log_{10}$ of the odds ratios, called the lod score, exceeds a predetermined threshold T (typically 3, corresponding to 1000:1 odds). Thus, if f_1 is in fact the correct distribution on X, the expected contribution to the lod score (*Elod*) from a single observation is

$$E(f_1, f_2) = \sum_{x \in X} f_1(x) \log[f_1(x)/f_2(x)].$$

In mathematics, this is well known as the relative entropy or Kullback–Liebler distance (18). The number of observations needed so that the expected aggregate lod score exceeds T is $T/E(f_1, f_2)$.

Assumptions About an RFLP Map. The human genome is about 3300 centimorgans long [1 centimorgan (cM) = 1% recombination]. For simplicity, we assume below the availability of a linkage map of "perfect" RFLPs, evenly spaced, with each RFLP so highly polymorphic that it is rarely found homozygous. Given 65 such RFLPs, defining intervals of size ≈52 cM, every region of the genome is within 20% recombination[¶] of an RFLP. Given 150, the intervals are ≈22 cM, and every locus is within 10%.

Since some 1000 RFLPs have already been discovered (11, 12), including an increasing number of highly polymorphic ones, the eventual availability of such maps seems ensured. (In the interim, we may compensate for uneven spacing and incomplete polymorphism by increasing the number of RFLPs, families, or both: several nearby modestly polymorphic RFLPs can be thought of as equivalent to a single highly polymorphic one.)

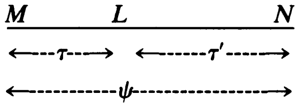
INTERVAL MAPPING

We begin with a situation appropriate to mapping a dominant trait by using three-generation pedigrees: grandparents, parents, and n children. In this case, we can study n fully informative, phase-known meioses, one per child, provided the trait is fully penetrant.[‖]

Detecting Linkage: Using Unmapped Markers. Following Smith (19), it is traditional to consider hypothesis $H_{\theta,\alpha}$: that the recombination fraction is θ between a single marker under study and a putative linked locus accounting for a fraction α of occurrences of the trait (with the rest due to an unlinked locus or a nongenetic cause).

The possible outcomes for the segregation of the trait and the marker are observing 0, 1, 2, . . ., n crossovers. Assuming homogeneity ($\alpha = 1$) and linkage at θ, the probability of i crossovers is $p_{n,\theta}(i) = \binom{n}{i} \theta^i (1-\theta)^{n-i}$. Under heterogeneity, the probability is just the weighted average for linkage at θ and nonlinkage: $p_{n,\theta,\alpha} = \alpha p_{n,\theta} + (1-\alpha) p_{n,1/2}$. If $H_{\theta,\alpha}$ is the correct hypothesis, then the expected lod score for $H_{\theta,\alpha}$ over $H_{1/2,0}$ (nonlinkage) is $E(p_{n,\theta,\alpha}; p_{n,1/2,0})$. Table 1 shows the number of families needed to attain a lod score of 3, assuming recombination fractions of 10% and 20% from the nearest RFLP.[**]

Detecting Linkage: Using Interval Mapping. Let M and N be adjacent RFLP loci at a known recombination distance ψ. Using the power of the map, we may test the hypothesis $H'_{\tau,\alpha}$: a locus L accounting for a fraction α of occurrences of the trait lies in the interval, at a recombination fraction τ from M.



[Note that, given a map function, ψ and τ determine τ'. With the Haldane function, $\psi = \tau(1-\tau') + \tau'(1-\tau)$.]

A single phase-known meiosis results in one of four events: the allele inherited at L may be coinherited with in-phase alleles at both M and N, with probability $p_{MN} = (1-\tau)(1-\tau')$; at M alone with probability $p_M = (1-\tau)\tau'$; at N alone with probability $p_N = \tau(1-\tau')$; at neither with probability $p_0 = \tau\tau'$. The outcomes for a pedigree with n such phase-known meioses are identified by the numbers i, j, k, l of events of each type observed. Under homogeneity, the probability of such an outcome is

$$p'_{\mathrm{n},\theta}(i, j, k, l) = \binom{n}{i, j, k, l} p^i_{MN} p^j_M p^k_N p^l_0.$$

Under heterogeneity, the probability is again the weighted average $p'_{n,\theta,\alpha} = \alpha p'_{n,\theta} + (1-\alpha) p'_{n,1/2}$. The expected lod score and family resources required are computed as before.

The worst case occurs when L is midway between M and N; if so, call the recombination fraction to either marker θ. We may profitably rewrite the expected lod score in this case as

$$E(p'_{n,\theta,\alpha}; p'_{n,1/2,0}) = \sum_{m=0}^{n} \binom{n}{m} \psi^{n-m}(1-\psi)^m E(p_{m,\gamma,\alpha}; p_{m,1/2,0}), \qquad [1]$$

where $\gamma = \theta^2/[\theta^2 + (1-\theta)^2]$. There is a simple interpretation of Eq. 1: Chromosomes recombinant between M and N contribute zero expected information; only nonrecombinant meioses matter. The chance that there will be m such meioses is $\binom{n}{m}\psi^{n-m}(1-\psi)^m$. Given such a meiosis, the chance that L will fail to cosegregate with the flanking markers is γ, the (conditional) probability of a double crossover.

In short, the flanking markers are mathematically equivalent to a "virtual RFLP" at a recombination fraction $\gamma \approx 0$. The closer linkage of this virtual RFLP more than offsets the fact that only nonrecombinant meioses contribute information. Table 1 shows the resources needed to map a trait lying midway between two flanking RFLPs in 22-cM and 52-cM RFLP linkage maps (i.e., at 10% or 20% from the flanking markers).

95% Certainty of Success. In planning a linkage study, it is prudent to collect more than just the average number of families needed to obtain a lod score of 3. Table 1 therefore shows the number required to ensure a 95% certainty of

[¶] Here, as throughout, we assume the Haldane map function, corresponding to no crossover interference. Positive interference, for example at the Kosambi level, makes interval mapping slightly more efficient (data not shown).

[‖] If penetrance is incomplete, unaffected children are of uncertain genotype. Thus, they add little to the analysis. In this case, n should be the number of affected children.

[**] More extensive tables for the case of single unmapped markers appear in ref. 16. To facilitate comparison, we have employed the same recombination fractions and thresholds for detecting both linkage and heterogeneity.

Table 1. Interval mapping: Numbers of families needed to detect linkage or heterogeneity with and without a map for dominant and recessive traits

			22-cM RFLP map (markers at 10%)						52-cM RFLP map (markers at 20%)					
			Linkage				Heterogeneity		Linkage				Heterogeneity	
			Average		95% success		Average		Average		95% success		Average	
Trait	n	α	Sing.	Map	Sing.	Map	Sing.	Map	Sing.	Map	Sing.	Map	Sing.	Map
Dominant	2	1.0	9	7	18	10	NA	NA	18	11	39	20	NA	NA
		0.9	12	9	25	17	368	21	22	14	50	29	2253	92
		0.7	20	15	45	33	131	5	37	24	87	53	559	16
		0.5	38	30	90	69	131	2	72	46	171	109	476	7
		0.3	101	79	245	191	228	2	194	125	466	300	756	4
		0.1	854	669	2046	1604	1370	1	1672	1070	3986	2558	4355	3
	3	1.0	6	4	12	7	NA	NA	12	7	26	14	NA	NA
		0.9	8	6	17	11	130	12	15	9	34	19	738	51
		0.7	13	9	30	21	48	3	24	15	58	35	197	10
		0.5	23	18	58	43	47	2	46	28	113	70	167	4
		0.3	59	44	148	111	76	1	119	74	295	183	255	3
		0.1	454	339	1114	835	410	1	975	588	2357	1433	1369	2
	4	1.0	5	3	9	5	NA	NA	9	5	19	10	NA	NA
		0.9	6	4	13	8	69	8	11	7	26	15	367	33
		0.7	9	7	23	16	26	2	18	11	44	26	103	7
		0.5	16	12	43	31	25	1	33	20	84	51	88	3
		0.3	39	28	104	75	39	1	83	50	212	128	130	2
		0.1	275	197	699	503	188	1	643	367	1583	917	651	1
Recessive	2	1.0	16	9	33	15	NA	NA	52	22	120	45	NA	NA
		0.9	19	12	43	23	1564	41	65	27	150	60	*	320
		0.7	32	20	75	44	410	8	107	47	252	107	5464	45
		0.5	62	39	149	92	359	4	208	92	495	216	4111	18
		0.3	168	106	405	254	579	2	573	252	1362	601	6070	10
		0.1	1444	905	3445	2165	3360	2	5063	2209	*	5251	*	7
	3	1.0	7	4	15	7	NA	NA	21	9	50	19	NA	NA
		0.9	8	5	20	10	511	19	25	11	62	25	6680	134
		0.7	13	8	33	19	115	4	41	18	101	43	1192	20
		0.5	24	14	62	37	85	2	76	33	189	83	780	8
		0.3	57	35	150	91	112	1	196	84	487	213	973	5
		0.1	415	247	1042	627	478	1	1595	655	3861	1611	4334	3
	4	1.0	4	3	9	4	NA	NA	12	5	29	11	NA	NA
		0.9	5	3	12	6	137	9	14	6	36	15	1913	66
		0.7	8	5	21	12	40	2	23	10	58	26	379	11
		0.5	13	8	37	22	32	1	41	18	107	48	257	5
		0.3	30	18	85	51	42	1	101	42	263	115	311	3
		0.1	181	102	494	284	146	1	742	286	1859	745	1193	2

Sing. = single-marker methods; Map = interval mapping; n = number of children if dominant; n = number of affected children if recessive (see text); * = >10,000; NA = not applicable.

detecting linkage, should it be present. (The actual lod score is approximately normally distributed about the *Elod* with standard deviation calculable from the distributions given above.) It is striking to note that, for a homogeneous disease, a sample large enough to ensure a 95% chance of success when interval mapping is used affords less than 50% chance when traditional single-marker methods are used.

Detecting Heterogeneity: Using Unmapped Markers. Homogeneity is typically tested (10, 19) by comparing the maximum likelihood of the sample when we allow heterogeneity to the maximum likelihood when we insist on homogeneity. If the hypothesis $H_{\theta,\alpha}$ is correct, the maximum likelihood allowing for heterogeneity occurs at (θ, α), while the maximum likelihood under homogeneity occurs at some point $(\theta^*, 1)$. In the case of a dominant trait, it is intuitively plausible—and simple calculus confirms—that θ^* is just the "apparent" recombination frequency: $\theta^* = \theta\alpha + \frac{1}{2}(1 - \alpha)$, independent of n. Thus, to compare the hypotheses of heterogeneity and homogeneity, we are interested in the expected lod score $E(p_{n,\theta,\alpha}; p_{n,\theta^*,1})$. The number of families needed to attain 10:1 odds in favor of $H_{\theta,\alpha}$ relative to $H_{\theta^*,1}$ appears in Table 1.[††]

Detecting Heterogeneity: Using Interval Mapping. We adopt a similar approach in the case of mapped markers. Allowing for heterogeneity, the maximum likelihood occurs at the true values (τ, α). If we insist on homogeneity, however, the maximum likelihood occurs at the midpoint of the interval—except if α is very close to 1, in which case the maximum occurs between the midpoint and the true location. The *Elod* is thus $E(p'_{n,\tau,\alpha}; p'_{n,\theta,1})$, except for α close to 1 ($\alpha > 0.9$, in our

[††] As a working rule, odds of 10:1 seem appropriate for "indication" of heterogeneity. We would propose a higher threshold for "proof" of heterogeneity—at least 50:1.

Table 2. Simultaneous search: Number of families needed for simultaneous search of several equally frequent loci causing a heterogeneous trait, and comparison with nonsimultaneous mapping of individual component loci

	22-cM RFLP map (markers at 10%)									
n \ k	1		2		3		4		5	
	7	7	12	19	17	39	23	68	29	107
2		7		30		65		113		174
	7	10	17	38	32	83	50	144	72	222
	5	5	7	12	10	21	12	33	15	49
3		5		18		37		62		93
	5	7	10	24	18	49	26	82	36	124
	4	4	5	8	6	14	8	21	9	29
4		4		12		24		39		58
	4	5	7	17	12	33	17	54	22	80
	3	3	4	6	5	10	6	15	6	20
5		3		9		17		27		39
	3	4	6	13	9	25	12	39	15	56

	52-cM RFLP map (markers at 20%)									
n \ k	1		2		3		4		5	
	11	11	20	35	30	76	40	135	50	216
2		11		47		102		179		276
	11	18	29	72	55	158	88	277	125	428
	8	8	12	21	17	42	22	72	27	111
3		8		29		61		104		158
	8	12	18	46	32	98	49	169	69	258
	6	6	9	15	12	28	15	45	17	68
4		6		21		42		69		103
	6	9	13	34	21	69	32	117	44	176
	5	5	7	11	9	20	10	32	12	46
5		5		16		31		50		73
	5	8	10	26	16	53	23	87	30	129

	k	
	a	*b*
n		*c*
	e	*d*

Box *k*, *n* applies to the situation of a dominant trait caused (equally often) by any of *k* loci given families with *n* fully informative meioses. Clockwise from the top left, the entries are the numbers of families needed on average to obtain 1000:1 odds in favor of the following: *a*, the correct set of *k* loci (as a set), when suspected *a priori* (ensemble test); *b*, any one of *k* component loci, when simultaneous search with RFLP map is used (specific component test); *c*, any one of *k* component loci without simultaneous search, but with RFLP map; *d*, any one of *k* component loci when single unmapped markers are used; *e*, the correct set of *k* loci (as a set), when located by simultaneous search (ensemble test with increased threshold).

cases). The number of families needed to attain 10:1 odds in favor of the correct hypothesis relative to homogeneity is given in Table 1, for a locus located at the midpoint.[‡‡]

Recessive Traits. Most of our analysis is easily adapted to the situation appropriate for mapping a recessive trait.[§§] Since unaffected children are of uncertain genotype, they contribute little to mapping the trait. So, let *n* be the number of *affected* children in the family. The phase of a recessive disease allele is usually in doubt, but the phase of the RFLPs can be determined from the DNA of grandparents, unaffected sibs, or both. Table 1 shows the resources required for studying a recessive trait of unknown phase, with RFLPs of known phase.

SIMULTANEOUS SEARCH

Following the entire set of trait-causing loci is potentially more efficient than following a single locus, as we noted in the introduction. To study *k* intervals simultaneously, we can compare different probability distributions on the Cartesian product $X_1 \times \ldots \times X_k$ of joint outcomes, where X_i is the set of outcomes for the *i*th interval. The appropriate probability distributions to use depend on the precise question we wish to ask. Three come to mind:

Testing a Specific Ensemble of Loci. Suppose that intervals 1, 2, . . ., *k* are suspected on *a priori* grounds of each containing a trait-causing locus. (To concoct an example, if we were studying familial cancers the intervals might contain known oncogenes. If an initial study failed to implicate any single oncogene as the cause, we would then want to test whether oncogenes as a class were to blame.)

For simplicity, consider the hypothesis $H_{1,2,\ldots,k}$: each of intervals 1, 2, . . ., *k* contains at its midpoint a trait-causing locus responsible for $1/k$ of the occurrences of the trait, and the alternative hypothesis H_0: *none* of the intervals are linked to a trait-causing locus. We refer to comparing these hypotheses as the *ensemble test*.

The probability distribution on the set of possible outcomes under $H_{1,2,\ldots,k}$ is simply the weighted average of the *k* product distributions obtained when we assume that the trait maps to interval *i* and is unlinked to the remaining intervals (for $i = 1, 2, \ldots, k$). Under H_0, it is the product distribution associated with nonlinkage to all the intervals. The expected lod scores for the ensemble test are given in Table 2 for the case of a trait caused by *k* equally frequent dominantly acting genes.

Testing a Specific Component Locus. A high lod score on the ensemble test is an indication that some or all of the loci in the set are involved. Proving that any particular component locus is involved and estimating the frequency with which it is a cause of the trait require studying a richer set of alternative hypotheses. A complete analysis would continue by varying the fraction of all cases attributed to each of the loci (between 0 and 1) and finding a point maximum likelihood and a confidence region around it. (The mathematics is the same as in the previous paragraph, although increasingly efficient algorithms and computing power are required.)

To illustrate what is required to test a specific component locus, let us compare hypothesis $H_{1,2,\ldots,k}$ with hypothesis $H_{1,2,\ldots,k-1,k+1}$—i.e., that the trait is due (equally often) to loci in intervals 1, 2, . . ., $k - 1$, and $k + 1$, where $k + 1$ is a locus unlinked to *k*. (Equally well, $k + 1$ could refer to a nongenetic cause—say, a virus—accounting for $1/k$ of the cases.)

Table 2 shows the number of families needed to attain a lod score of 3 in favor of the former hypothesis, when it is correct. We refer to this as a *specific component test*. For comparison, the table also lists the resources that would be required to prove linkage to a locus with $\alpha = 1/k$ if we do not use the simultaneous method.

[‡‡] A 95% certainty of success at proving heterogeneity requires considerably more families: 3–4 times as many in the case of a RFLP map and 5–6 times in the case of unmapped markers (data not shown). This reflects the fact that heterogeneity may masquerade as homogeneity more easily than vice versa.

[§§] The main differences are (*i*) even assuming homogeneity, meioses within a family are not statistically independent and (*ii*) in detecting heterogeneity, the maximum likelihood value θ^* must be determined by numerical approximation.

Searching for a Set of Loci. Usually, we will have no *a priori* beliefs about where the trait-causing loci will be. In this case, we may try all sets of k intervals in turn, in the manner just outlined. The only difference is that the threshold T for acceptance must be raised to account for an increased likelihood of false positives.

A simple Bayesian argument suggests an appropriate threshold. The *a priori* odds that a gene and a marker will be linked (at an effectively detectable distance) are about 50:1 against. Roughly, this is why Morton (9) prescribed that the data from a traditional linkage study must yield a lod of 3, or 1000:1 odds in favor of linkage, to be accepted: so that the *a posteriori* odds of linkage are 20:1 in favor. In the case of simultaneously mapping k loci of similar frequency, a comparable approximation is to take the *a priori* odds against linkage to a set of k intervals to be $\binom{50}{k}$:1 against. Hence, we might take $\log_{10} [20 \binom{50}{k}]$ as a sensible threshold. That is, lod $>$ 3, 4.4, 5.6, 6.7, and 7.6 for k = 1, 2, 3, 4, and 5, respectively.[¶¶]

Using the proposed thresholds, Table 2 shows the number of families needed to prove that a set of k loci found in this manner adequately accounts for occurrences of a trait. Such a finding would be strong evidence that the trait in question actually is genetic—which for certain disorders would be the most important discovery. It would also justify more extensive studies to prove or disprove the involvement of each of the k loci, by using the method above.

An Illustration. Consider a heterogeneous dominant trait caused (equally often) by mutations at any of three loci and suppose that we have available a 22-cM RFLP linkage map and families with n = 3 informative meioses.

With *a priori* reasons to suspect the correct three loci, we require 10 families on average to obtain 1000:1 odds for our guess in the ensemble test. (Without a prior hypothesis, 18 families would be needed to obtain 1000:1 odds in favor of a triple of loci obtained by searching the data.) Showing that a particular one of the three loci is in fact involved requires 21 families.

By contrast, showing involvement of a locus with $\alpha = 0.33$ without simultaneous mapping would require 37 families if we were to use flanking RFLP markers and 49 if we were to insist on using single unmapped markers.

DISCUSSION

The mathematical methods we describe here to exploit the full information contained in an RFLP map considerably reduce the number of families needed to study a human trait. The main conclusions that emerge from the data are as follows:

(i) Interval mapping is more efficient than using single markers. With a 52-cM map, one needs about 40% fewer families to detect linkage to a dominant trait and about 60% for a recessive trait. With a 22-cM map, the reductions are about 25% and 40%, respectively. In addition to these savings in the number of families needed to find the correct region, fewer DNA probings are required to reject *unlinked* regions when an RFLP map is used (data not shown).

(ii) Simultaneous search is more efficient than studying single loci alone. By exploiting the completeness of the RFLP map, we require about one-third as many families for typical tasks as if we used single unmapped markers. (Relative to using flanked intervals but nonsimultaneous methods, the savings are twofold.)

(iii) Heterogeneity is far easier to detect by using a map. Between 1/5th and 1/50th as many families are required for typical problems. In practice, detecting heterogeneity without an RFLP map may be somewhere between impractical and impossible. Heterogeneity is so much easier to detect with interval mapping, because it gives rise to outcomes that can be rationalized only as rare double crossovers if we insist on homogeneity.

The savings in the number of families needed is not simply a matter of economics. With genetic markers no longer a major constraint, finding enough families with multiple affected members will pose the greatest limitation to the study of human heredity. Reducing these requirements by using information in a more powerful way may bring some complex human traits within the realm of molecular analysis.

We thank Helen Donis-Keller, Luca Cavalli-Sforza, Mary-Claire King, and Newton Morton for helpful comments on the manuscript, the Whitehead Institute for computing facilities, and Michael Candan for technical assistance. This work was partially supported by grants from the System Development Foundation to E.S.L. and from the National Institute of General Medical Sciences to D.B.

[¶¶] In fact, these thresholds are slightly too conservative, because there is statistical dependency between the scores for the $\binom{50}{k}$ different sets of k loci. We shall address such mathematical issues elsewhere.

1. Botstein, D., White, R. L., Skolnick, M. H. & Davis, R. W. (1980) *Am. J. Hum. Genet.* **32,** 314–331.
2. Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. C., Sakaguchi, A. Y., Young, A. B., Shoulson, I., Bonilla, E. & Martin, J. B. (1983) *Nature (London)* **306,** 324–328.
3. Reeder, S. T., Breuning, M. H., Davies, K. E., Nicolls, R. D., Jarman, A. P., Higgs, D. R., Pearson, P. L. & Weatherall, D. J. (1985) *Nature (London)* **317,** 542–544.
4. Tsui, L. C., Buchwald, M., Barker, D., Braman, J. C., Knowlton, R. G., Schumm, J. W., Eiberg, H., Mohr, J., Kennedy, D., Plesvic, N., Zsiga, M., Markiewica, D., Akots, G., Brown, V., Helms, C., Gravius, T., Parker, C., Rediker, K. & Donis-Keller, H. (1985) *Science* **230,** 1054–1057.
5. Knowlton, R. G., Cohen-Haguenauer, O., Van Cong, N., Frezal, J., Brown, V. A., Braman, J. C., Schumm, J. W., Tsui, L. C., Buchwald, M. & Donis-Keller, H. (1985) *Nature (London)* **318,** 380–382.
6. Wainwright, B. J., Scambler, P. J., Schmidtke, J., Watson, E. A., Law, H. Y., Farrall, M., Cooke, H. J., Eiberg, H. & Williamson, R. (1985) *Nature (London)* **318,** 384–385.
7. White, R., Woodward, S., Leppert, M., O'Connell, P., Hoff, M., Herbst, J., Lalouel, J. M., Dean, M. & Vande Woude, G. (1985) *Nature (London)* **318,** 382–384.
8. Davies, K. E., Pearson, P. L., Harper, P. S., Murray, J. M., O'Brien, T., Sarfrazi, M. & Williamson, R. (1983) *Nucleic Acids Res.* **11,** 2303–2312.
9. Morton, N. (1955) *Am. J. Hum. Genet.* **7,** 277–318.
10. Ott, J. (1985) *Analysis of Human Genetic Linkage* (Johns Hopkins Press, Baltimore).
11. Braman, J., Barker, D., Schumm, J., Knowlton, R. & Donis-Keller, H. (1985) *Cytogenet. Cell Genet.* **40,** 589 (abstr.).
12. Schumm, J., Knowlton, R., Braman, J., Barker, D., Vovis, G., Akots, G., Brown, V., Gravius, T., Helms, C., Hsiao, K., Rediker, K., Thurston, J., Botstein, D. & Donis-Keller, H. (1985) *Cytogenet. Cell Genet.* **40,** 739 (abstr.).
13. White, R., Leppert, M., Bishop, D. T., Barker, D., Berkowitz, J., Brown, C., Callahan, P., Holm, R. & Serominski, L. (1985) *Nature (London)* **313,** 101–105.
14. Drayna, D., Davies, K., Hartley, D., Mandel, J. L., Camerino, G., Williamson, R. & White, R. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 2836–2841.
15. Jaspers, N. G. J., Painter, R. B., Paterson, M. C., Kidson, C. & Inoue, T. (1985) in *Ataxia-Telangiectasia*, eds. Gatti, R. A. & Swift, M. (Liss, New York), pp. 147–162.
16. Cavalli-Sforza, L. & King, M.-C. (1986) *Am. J. Hum. Genet.* **38,** 599–616.
17. Ott, J. (1983) *Ann. Hum. Genet.* **47,** 311–320.
18. Kullback, S. (1959) *Information Theory and Statistics* (Wiley, New York).
19. Smith, C. A. B. (1963) *Ann. Hum. Genet.* **27,** 175–182.