Nucleic Acids Research, 1993, Vol. 21, No. 1 13–20

Happy mapping: linkage mapping using a physical analogue of meiosis

Paul H.Dear+ and Peter R.Cook*
Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK

Received November 5, 1992; Revised and Accepted December 3, 1992

ABSTRACT

We have devised a simple method for ordering markers on a chromosome and determining the distances between them. It uses *hap*loid equivalents of DNA and the *poly*merase chain reaction, hence '*happy* mapping'. Our approach is analogous to classical linkage mapping; we replace its two essential elements, chromosome *breakage* and *segregation*, by *in vitro* analogues. DNA from any source is *broken* randomly by γ-irradiation or shearing. Markers are then *segregated* by diluting the resulting fragments to give aliquots containing ~1 haploid genome equivalent. Linked markers tend to be found together in an aliquot. After detecting markers using the polymerase chain reaction, map order and distance can be deduced from the frequency with which markers 'co-segregate'. We have mapped 7 markers scattered over 1.24 Mbp using only 140 aliquots. Using the 'whole-genome' chain reaction, we also show how the approach might be used to map thousands of markers scattered throughout the genome. The method is powerful because the frequency of chromosome breakage can be optimized to suit the resolution required.

INTRODUCTION

Various strategies are being applied to map complex genomes, including those involving the use of radiation to break genomes (eg refs 1–3). All suffer a number of drawbacks. Random cloning and chromosome 'walking' are impeded by repeated sequences and inability to clone some loci (4), or confused by rearranged or co-ligated inserts (5). Difficulties in cutting DNA into large fragments (6) and resolving them (7) prevent efficient restriction mapping. The infrequency of meiotic recombination also limits resolution of classical linkage mapping to, at best, one marker per 10^6 bp (8). The resolution offered by *in situ* hybridization using metaphase chromosomes is little better (9) though this can be improved when applied to interphase nuclei (10).

We have developed a simple approach for mapping that sidesteps some of these problems (Fig 1). It should prove useful in closing final gaps in maps, for mapping over distances that are inaccessible using other approaches and for generating maps of diverse groups of organisms and individuals. Our approach is analogous to classical linkage mapping where meiosis both **breaks** DNA by crossing-over and **segregates** it into aliquots containing haploid amounts of DNA (ie sperm or eggs). We replace these processes by *in vitro* analogues; we break DNA physically and then dilute and divide it into aliquots that contain ~1 haploid equivalent. As we control the frequency of breakage, we control the scale over which we map, from a few kilobases to a few megabases. Generally we introduce many more breaks into the chromosome than meiosis and so attain far higher resolution.

Cells (either diploid or haploid, obtained from any tissue or individual) are embedded in agarose and extracted to leave DNA, just as for pulsed-field electrophoresis (Fig 1). DNA is then broken at random into fragments that are ~3× longer than the average marker spacing. γ-irradiation generates large fragments for long-range mapping; shearing gives short fragments (<0.4 Mbp) for high-resolution mapping. Aliquots containing ~1 haploid equivalent are taken from the pool of fragments; most contain 0 or 1 copy of any given marker, some 2 or more. Once fragments have been divided into aliquots, no special care need be taken to preserve linkage. Now markers are amplified using the polymerase chain reaction (PCR; ref 11) and the products analyzed by gel electrophoresis. If two markers are closely linked, they tend to lie on the same DNA fragment and so are often found together in an aliquot (ie, they 'co-segregate'). Unlinked markers lie on different DNA fragments and therefore do not show this association. A table relating the association (lod score) between different pairs of markers is constructed and then marker order and distance can be calculated, much as in conventional linkage mapping.

A forerunner of this strategy used sperm as a source of haploid DNA and so required the technically-difficult manipulation of single sperm (12; see also 13). Use of diploid cells (from any tissue type from any organism) makes the technique simpler, easily automatable and accessible to anyone with a thermal cycler and electrophoresis equipment. But use of diluted DNA means that two (or more) copies of the same marker might occasionally be found together in an aliquot so that more samples must be analyzed to determine linkage. However, this disadvantage is

* To whom correspondence should be addressed

+Present address: MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

offset if we size-select fragments (eg by pulsed-field electrophoresis or flow sorting) before division into aliquots to eliminate the smallest and largest that contain no mapping information but which add considerably to the work required.

We demonstrate the approach by mapping 7 markers scattered along 1.24 Mbp of the X chromosome in the Duchenne muscular dystrophy (DMD) locus. We use only 140 two-phase PCRs and gel tracks to generate a detailed map that agrees well with the known map (14,15). Using the 'whole-genome' chain reaction (16), we also show how the approach might be used to map thousands of markers scattered throughout the genome (eg sequences encoding cDNAs; ref 17).

MATERIALS AND METHODS

Preparation of DNA

Peripheral blood lymphocytes were isolated from whole female blood using Ficoll-Paque (Pharmacia) and encapsulated in agarose microbeads (18) at 10^6 cells/ml agarose (ie 6μg/ml DNA). Encapsulated cells were washed 6× at intervals of 30min in 10mM Tris (pH 8.0), 1mM EDTA, 1% w/v lithium dodecyl sulphate and stored at 4°C (no change in linkage was detected after storage for 3 weeks). As some cells fail to be encapsulated and others are then dislodged from the bead surface, we assume the DNA concentration is now 3μg/ml packed beads.

γ-irradiation and pulsed-field electrophoresis

Microbeads were γ-irradiated (40J/kg; ref 19). Then 25μl (packed vol) beads were mixed with 1.5ml agarose (Sigma type VII; 0.5% in 0.5× TBE) at 40°C, loaded into a single well (8cm wide) in a gel (1% BioRad ultra-pure agarose; 15×15×0.2cm; 0.5× TBE; size standards, *S.cerevisiae* chromosomes from Promega) and subjected to electrophoresis (LKB Pulsaphore with hexagonal electrode array; 0.5× TBE; 15°C; 150V; pulse-time 200s) for 20h. Marker lanes were excised, stained with ethidium (0.6μg/ml) and viewed under uv light. The remainder of the gel (which contained too little DNA to be detected by ethidium-staining) was equilibrated with 10mM Tris (pH 8.0), 1mM EDTA (TE), stored at 4°C and 500nl agarose plugs (aliquots) containing DNA excised using a glass capillary.

Shearing and sonication of DNA

10μl beads (packed vol) were mixed with 5ml agarose (Sigma type VII in TE) at 40°C, melted at 70°C for 10 min in a 15ml tube, the tube inverted twice to disperse and shear DNA and the solution poured between glass plates 0.8mm apart and allowed to set. Agarose plugs were excised as above. Alternatively, 5μl beads (packed vol) were suspended in 5ml TE, melted at 70°C, sonicated to shear DNA to <0.01 Mbp and 1μl aliquots taken.

Polymerase chain reactions

All PCRs (5 or 10μl, overlaid with 20μl mineral oil; Sigma) were conducted in tubes or 96-well plates (Techne) in 20mM Tris-HCl (pH 8.3 at 20°C), 50mM KCl and 200μM of each dNTP supplemented with primers, $MgCl_2$ and Taq polymerase as indicated. Table I gives primer sequences and their derivation. Primers (19−21mers) were designed to have ~50% GC with 2 G/C at the 3′ end and 1 at the 5′ end. Of all primers designed using these criteria, ~90% primed successfully using a 2-phase PCR (see below). Taq polymerase from Cetus and Boehringer was used for first- and second-phase PCRs respectively. A typical 2-phase PCR used in Fig 2A was as follows. After a first-phase PCR [5μl; 2.0mM $MgCl_2$; 1μM each oligo in primer set 9-EXT (Table I); 0.25u Taq polymerase; initial denaturation at 93°C for 5min, followed by 25 cycles at 95°C for 20s, 50°C for 30s and 72°C for 60s], 45μl water was added and 2μl of the mixture used for a second-phase PCR [10μl; 4mM $MgCl_2$; 1μM each oligo in primer set 9-INT (Table I); 1u Taq polymerase; initial denaturation at 93°C for 5min, followed by 33 cycles of 94°C for 10s, 54°C for 30s, 72°C for 60s]. Reaction products were then mixed with 10μl sample buffer (25mM EDTA, 10mg/ml Ficoll 400, 0.1mg/ml bromophenol blue in TBE), and 5μl loaded onto a gel (3.5% agarose; 1×TBE containing 0.6μg/ml ethidium bromide; 20V/cm for 42min; ref 20).

The detection efficiencies of all primers were initially tested on serial dilutions of sheared human DNA (~9, 3 and 1 pg; ie ~3, 1 and 0.3 haploid equivalents). Dilutions with ~9 pg should give intense bands (copies of each marker) after amplification in an ethidium-stained gel (as above). With ~3 pg dilutions, bands remain equally intense, but occur in only ~65% of samples (Poisson distribution). With ~1 pg, bands remain as intense but occur in ~30% of samples. Tightly-linked markers D31-D32 provide a further estimate of detection efficiency: assuming they are never separated by shearing, any occurrence of one amplimer without the other in a sheared sample must be due either to failure to amplify or to contamination, which can be measured using negative controls. In fact, D31 was found without D32, or *vice versa*, in ~8% of aliquots containing 3pg of sheared DNA (0.05−0.4 Mbp). Given the observed contamination rate of ~1.5% and assuming that failure of amplification of D31 and D32 are independent events, this is the result expected from a detection efficiency of ~94% for either marker.

Having established that markers could be detected efficiently, the DNA content of aliquots was estimated as follows. Fragments from a known number of cells were assumed to be uniformly distributed throughout the gel after pulsed-field electrophoresis or throughout the solution after shearing or sonication. Then appropriate volumes expected to contain ~1 haploid equivalent (ie ~3 pg) were taken; then each amplimer should occur in ~65% of aliquots (Poisson distribution). If they do not, it can be assumed that there is more or less than 1 haploid equivalent present per aliquot and the sample volume should be increased correspondingly. Note that both theory (12) and practice show that the amount of work required is relatively insensitive to DNA content/aliquot between 0.5−1.3 haploid equivalents, provided that each aliquot in a set contains the same amount of DNA. Again, tightly-linked markers (D31 and D32) served as a positive control for linkage; they should occur together in ~65% of aliquots.

Whole-genome PCRs were modified from Zhang *et al.* (ref 16; 5μl; 50 cycles of 92°C×1min, 37°C×2min, 55°C×4min; transition from 37°C to 55°C at 0.1°C/sec) using 2.5mM $MgCl_2$, 7.5μM primer N_{15} (a mixture of all 15-mers), 0.5u Taq polymerase (Cetus); then 15μl water were added and 1μl of this mixture used for the two-phase PCR as above.

Contamination control

Rigorous precautions were taken to exclude foreign DNA (especially PCR products) from samples, particularly prior to the second-phase PCR. First-phase PCRs were prepared using dedicated equipment in a laminar-flow hood in a room remote from that used for second-phase PCRs and gel analysis. Gel tanks used to resolve large fragments were decontaminated (1M HCl;

2hr). Disposable plastics were free of contamination provided they were not stored in the PCR product-analysis area. The frequency of markers detected in negative controls was <2% (whole-genome PCR gave contaminations of <4%). [This has now improved; in a recent series of ~1300 nested PCR reactions in a different laboratory, there was only 1 contaminant band.]

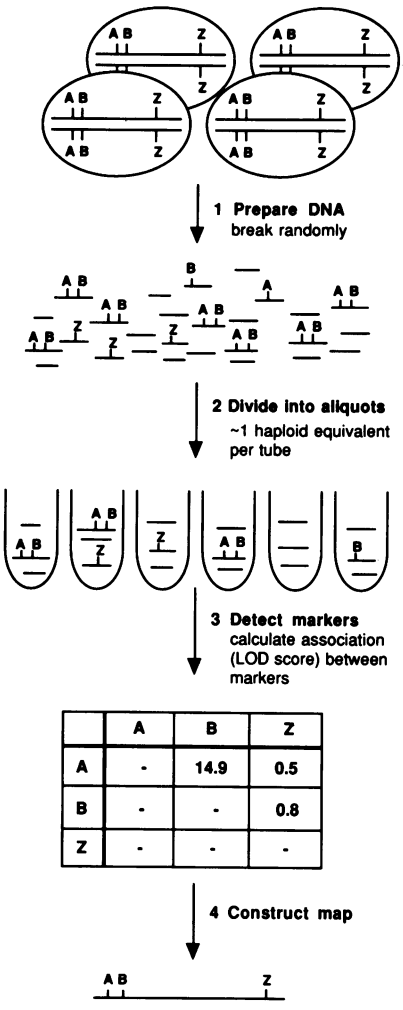
Analysis

Data entry and analysis were performed using programs written by P.H.D. in Turbo Basic (Borland). A data entry program creates a file recording the presence or absence of each marker in each aliquot. From the frequency of occurrence of all markers, we determine the mean DNA content per aliquot, assuming a Poisson distribution of markers. Then, for each possible pair of markers, AB, the number of aliquots containing both (AB), one (Ab or aB) or neither (ab) of the markers is determined. [We use the convention of upper- or lower-case letters to denote presence or absence of a marker.] These numbers are n(AB), n(Ab), n(aB) and n(ab) respectively. Lod and θ values are then determined as follows.

At each value of T (probability of breakage between markers A and B) from 0 to 1 at intervals of 0.01, the concentration of fragments (mean number per aliquot) carrying both markers or either marker alone is calculated, based on the mean DNA content per aliquot (above). Then the probabilities of an aliquot containing both, either or neither of the markers [p(AB), p(Ab) or p(aB), p(ab) respectively] are calculated assuming a Poisson distribution of fragments amongst aliquots. From these probabilities, the total probability of obtaining the observed set of results is calculated as $n(AB)^{p(AB)}*n(Ab)^{p(Ab)}*n(aB)^{p(aB)}*n(ab)^{p(ab)}$; the logarithm of this probability, L_T, is then taken. The value of L at T=1 (ie L_1) is then subtracted from all values of L_T. Then, the maximum value of L_T is the log-odds (lod) for linkage between A and B, and the value of T at which it occurs is the best estimate of θ, the probability of breakage between them.

For each possible order, the minimum number of breaks required to give observed results was calculated. In classical linkage mapping, the likeliest order is that requiring fewest breaks; here, this is only approximately true due to experimental errors (contamination, failure to amplify) and the possibility of two or more marker copies in an aliquot. For each of the likeliest 100 orders found by this method, the marker spacing (in terms of θ) which best fit the pairwise linkage data was found, and the corresponding likelihood of the map calculated.

The mapping function relating θ to distance was determined from first principles. In the case where DNA from which the aliquots are taken is of a single size (S), distance (D) is directly proportional to θ up to a value of $\theta=1$, at which D=S. However, two-dimensional pulsed-field gels showed that DNA at a given point in the gel after electrophoresis in one-dimension contained a proportion of smaller fragments arising from breakage during electrophoresis (not shown). Therefore we model the distribution of fragment sizes by assuming (simplistically) that a given region of the gel contains not only fragments of the expected size (based on comparison with size standards), but also a range of smaller fragments whose concentration is proportional to their size. The mapping function is then calculated as the mean of the functions for all fragment sizes, weighted according to their respective concentrations.



	A	B	Z
A	-	14.9	0.5
B	-	-	0.8
Z	-	-	-

Figure 1. Mapping strategy illustrated using three markers, A and B (linked) and Z (remote) on the two chromosomes of a diploid cell. (1) DNA is prepared from many cells and broken into large fragments. (2) Fragments are divided into aliquots (typically 100–200) containing ~1 haploid genome equivalent. (3) Which markers are present in each aliquot is determined (using PCR). A table can then be constructed, in which 'lod scores' reflect the tendency of pairs of markers to occur together in the aliquots. Markers A and B tend to occur together, giving a high lod score (in this example, 14.9). Marker Z does not co-occur with A or B more frequently then expected by chance, so lod scores between Z and A, or Z and B, are low. (4) A map can be deduced from the lod scores.

RESULTS

Markers

In principle, a marker can be any single-copy sequence that can be amplified using the PCR to give copies that can be identified, either directly on a gel or indirectly using some other procedure (eg hybridization). Each marker is defined here using 2 sets of primer pairs, one nested within the other. After a 'first-phase' amplification using the exterior pair, the marker is amplified to levels detectable after ethidium-staining using a 'second-phase' PCR with the nested pair. 7 markers from the DMD locus on the X chromosome were analyzed (DMD markers are prefixed with D, followed by the exon number; D0 lies upstream of exon 1), plus two others, which are 0.06 Mbp apart on the unlinked β-globin locus on chromosome 11 (ie BGL and EGL). Table I lists primer sequences, regions of complementarity within the

Table I. Marker and primers.

MARKER		PRIMER			
NAME	BP	SEQUENCE		POSN	NOTE
SET 9-EXT					
D0	-	(L)	AACAAGCCTCTATTCAGAACC	-777	5
		(R)	GTCTGAATAAGAGAAGCAGCAGC	-516	5
D3	-	(L)	AGGCTTTAGTTTTCAAAAGGGG	-89	1,8
		(R)	TACGAGGTTGCTTTACTAAGG	229	1,8
D21	-	(L)	GAACAAAGAGCCAAGGTATCC	-194	1,8
		(R)	CTGTAGCTCTTTTTTTCTCTCTGG	158	1,2,8
D31	-	(L)	AGGAAGCTACATGGTAGAGG	-107	1,8
		(R)	CAAGTTGTCCAATATAGACTGG	166	1,8
D32	-	(L)	CAAGATGTCTCCATGAAGTTTCG	14	1,8
		(R)	GAAAGTCAAGGGGCACCTGC	203	1,8
D45	-	(L)	AAACATGGAACATCCTTGTGGGGAC	344	4,8
		(R)	CATTCCTATTAGATCTGTCGCCCTAC	176	4,8
D48	-	(L)	CAGGTTTCCAGAGCTTTACC	-3	3,8
		(R)	CCAACCTCCTGTAATACTCC	353	3,8
BGL	-	(L)	GAAGAGCCAAGGACAGGTAC	62010	7
		(R)	CAACTTCATCCACGTTCACC	62258	7
EGL	-	(L)	CATAGTCCAAGCATGAGCAG	-10996	6
		(R)	GTCACATTCTGTCTCAGGCA	-10705	6
SET 9-INT					
D0	201	(L)	GGTGTTTTAAGAATTGGC	-723	5
		(R)	CAGGCAGCTGAATTCTGAAGG	-542	5
D3	137	(L)	AATGTATGCATTGGAAGTGTGC	-56	1,8
		(R)	TGTCAGGCCTTCGAGGAGG	63	1,8
D21	262	(L)	GGGTATTAAGCTAAACTTGCC	-106	1,8
		(R)	CTGCACATCAGAAAAGACTTGC	135	1,2,8
D31	246	(L)	GAGGAGAGTTTCTGAATTTCG	-82	1,8
		(R)	GTATAATGCCCAACGAAAACACG	142	1,8
D32	150	(L)	CCAGCCAATTTTGAGCAGCG	50	1,2,8
		(R)	TTTGCCACCAGAAATACATACC	178	1,8
D45	380	(L)	GAGCTAACCGAGAGGGTGC	-223	3,8
		(R)	GCTGTTTGCAGACCTCCTG	139	3,8
D48	301	(L)	GCTCAAATAAAAGACCTTGGG	40	3,8
		(R)	CTGTGCCTATTGTGGTTATCC	320	3,8
BGL	214	(L)	GGCTGTCATCACTTAGACC	62030	7
		(R)	AGTAACGGCAGACTTCTCC	62225	7
EGL	127	(L)	GAGTCATGCTGAGGCTTAGG	-10869	6
		(R)	AGTCAGGTGGTCAGCTTCTC	-10762	6

Markers: name and amplimer length (internal primers). Primers: L and R denote 5′ and 3′ ends relative to published sequence; the position is that in the published sequence (see notes) unless otherwise indicated. Notes: (1) Koenig, M. (unpublished). (2) Ref 23. (3) Ref 24. (4) Primer sequences taken directly from ref 24. (5) Ref 25. (6) Ref 26. (7) See EMBL accession number J00179 for a list of refs. (8) Position given is relative to first base of corresponding exon.

two loci, and size of copies resulting from amplifications using the nested primer pair; a map of the DMD locus is given in Fig 4A.

Mapping the DMD locus

Computer simulations demonstrated the approach was theoretically feasible and indicated how tolerant it was of experimental variables like DNA content per aliquot, contamination and amplification efficiency (12; unpublished). The following illustrates the reproducibility and power of the method in practice.

Human female white blood cells were embedded in agarose and extracted to leave DNA. After γ-irradiation (40J/kg), the resulting DNA fragments were resolved on a pulsed-field gel and plugs of agarose containing ~1 haploid equivalent of DNA of ~1.75 Mbp were excised using a glass capillary. Each plug (aliquot) was then amplified in a first phase (25 cycles) using 9 primer-pairs (defining the 9 markers) simultaneously, followed

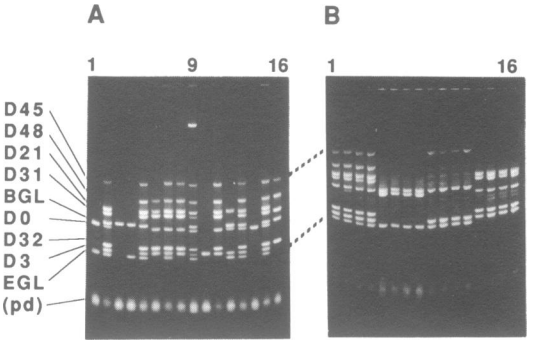


Figure 2. Marker detection after amplifying aliquots containing ~1 haploid equivalent either (A) directly or (B) after a whole-genome PCR. (**A**). 15 aliquots (~500nl from a pulsed-field gel) containing ~1 genome equivalent of ~1.75 Mbp DNA were amplified using 2-phase PCRs with the nested primer sets 9-EXT and 9-INT. Samples from each amplification were analyzed by gel electrophoresis and the gel stained and photographed. Lane 9: marker track containing each amplimer, plus λDNA. Positions of amplimers of each marker and primer dimers (pd) are indicated. (**B**). 4 aliquots like those used in (A), each containing ~1 genome equivalent of ~1.75 Mbp DNA, were amplified using a whole-genome PCR. After adding 15μl water, 4×1μl sub-aliquots from each were subjected to a 2-phase PCR and the products analyzed as in (A). The photograph therefore shows 4 sets (lanes 1–4, 4–8, 9–12, 13–16) of amplimers from each of the 4 original aliquots.

Table II. Lod scores between different pairs of markers derived using 140 aliquots of ~1.75 Mbp DNA. θ values are given in brackets.

	D0	D3	D21	D31	D32	D45	D48	BGL	EGL
D0	—	9.2 (0.37)	2.1 (0.65)	2.0 (0.66)	1.8 (0.68)	0.2 (0.89)	0.2 (0.89)	0.7 (0.77)	0.6 (0.80)
D3		—	8.2 (0.40)	4.7 (0.52)	4.5 (0.54)	1.7 (0.69)	1.4 (0.71)	0.8 (0.76)	0.1 (0.91)
D21			—	11.7 (0.33)	11.3 (0.34)	3.6 (0.58)	3.7 (0.57)	0.5 (0.80)	0.9 (0.76)
D31				—	45.3 (0.02)	4.6 (0.54)	5.4 (0.51)	0.1 (0.93)	0.0 (1.0)
D32					—	5.7 (0.50)	5.2 (0.52)	0.0 (0.95)	0.0 (1.0)
D45						—	18.9 (0.24)	0.0 (1.0)	0.0 (1.0)
D48							—	0.0 (0.99)	0.0 (1.0)
BGL								—	13.9 (0.28)
EGL									—

by a second-phase (33 cycles) with 9 'nested' primer pairs. The products of each reaction were then run on gels. Fig 2A shows an ethidium-stained gel with 15 lanes of the 140 involved in this experiment (lane 9 contains markers). Copies of some markers tend to be found together in a lane more frequently than others. For example, the top two bands (ie amplimers of the closely-linked DMD markers D45 and D48) are either found together (in 8 lanes) or absent together (in 5 lanes); in only 2 lanes is one present without the other, suggesting the two markers must be closely linked.

Analysis of the 'cosegregation' frequencies of all 9 markers in the 140 aliquots allowed map distance and order to be calculated (Materials and methods). Results can be expressed as the likelihood that two markers are linked (ie as a logarithm of the odds of linkage or lod); lods between all pairwise combinations of markers are listed in Table II. These lod scores are

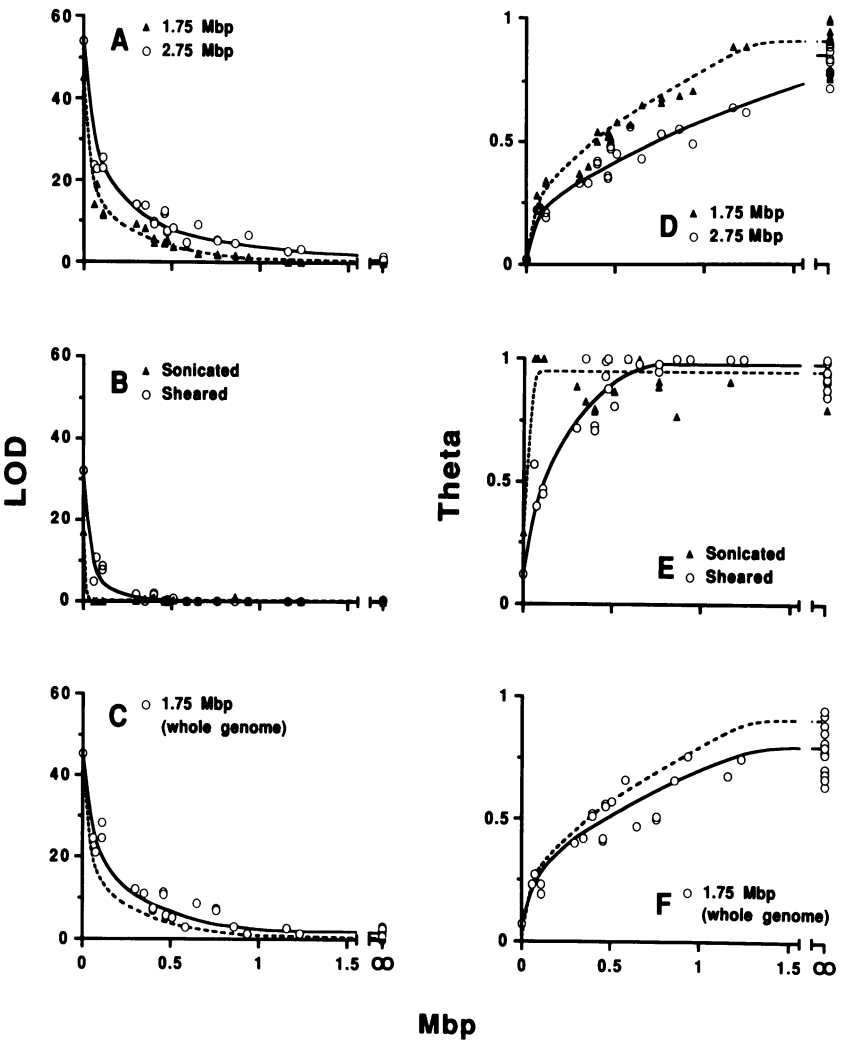


Figure 3. How DNA fragment size affects the relationship of (A–C) lod score or (D–F) θ with known inter-marker distance (Mbp). DNA was broken into different sizes, divided into aliquots containing ~1 haploid equivalent, markers amplified either (A,B,D,E) directly or (C,F) after whole-genome PCRs and lod scores or θ values between them plotted against inter-marker distance. Many of the 14 data points (between DMD and globin markers) at ∞ are superimposed. (**A,D**). 1.75 and 2.75 Mbp fragments derived by running γ-irradiated DNA (40J/kg) on a pulsed-field gel. Lods are highest at the shortest distances, falling to a background lod at ~0.7 Mbp (1.75 Mbp fragments) or >1.24 Mbp (2.75 Mbp fragments). θ increases with distance, reaching a maximum at ~1 Mbp (1.75 Mbp fragments) or >1.24 Mbp (2.75 Mbp fragments). (**B,E**). Sonicated (ie <0.01 Mbp) and sheared (ie 0.05–0.4 Mbp) DNA; no linkage is seen beyond ~0.01 and ~0.3 Mbp respectively. (**C,F**). 1.75 Mbp fragments (as in A), analyzed after whole-genome PCR. Dashed lines, taken from A and D (ie without whole genome PCR), are included for comparison.

analogous to those used in classical linkage mapping (21). For example, tightly-linked D31 and D32 show a lod of 45.3; in contrast, the lod between distant D0 and D48 is 0.2, similar to that between unlinked markers (eg D0 and EGL; 0.6).

Data in Table II have been re-plotted in Fig 3A (triangles) to illustrate how lod relates to the known inter-marker distance. Linkage can be detected up to ~0.7 Mbp, beyond which lods are not significantly above the background 'noise' of linkage seen between unlinked markers. In theory, the lod between unlinked markers should be zero, reflecting no association. In practice, various factors conspire to give a lod >0 between unlinked markers: thus, variation in amount of DNA in an aliquot, or in amplification efficiency (eg, due to a 'cold-spot' in the thermal cycler) tend to increase or decrease detection of all markers simultaneously, giving apparent associations. [The mean lod between all pairwise combinations of DMD and unlinked globin markers is 0.26.]

The lod between a pair of markers reflects the likelihood that they are linked. A second variable, θ, is an estimate of the probability of breakage between them, and hence reflects the distance between them. It is analogous to the recombination frequency in classical linkage analysis. θ is derived in parallel with the lod (Materials and methods), and is shown in Table II (brackets) and re-plotted in Fig 3D (triangles). At a distance of ~0.7 Mbp, θ reaches the level of the background 'noise' seen between unlinked markers, beyond which inter-marker distance cannot be reliably estimated. In theory the maximum value of θ should be 1, reflecting complete breakage between distant or unlinked markers. Again, experimental variability causing the 'background' lod of >0 also reduces the maximum value of θ

Table III. Rank order and relative likelihoods of different marker orders determined using various fragment sizes. For each candidate order, the optimal marker spacing and associated likelihood was determined. The likelihoods of the six most probable orders are expressed relative to that of the likeliest order in each experiment.

SIZE (Mbp)	RANK	MARKER ORDER	RELATIVE LIKELIHOOD
1.75	1	D0 D3 D21 D31 D32 D48 D45	1.0
	2	D0 D3 D21 D32 D31 D48 D45	0.93
	3	**D0 D3 D21 D31 D32 D45 D48**	0.80
	4	D0 D3 D21 D32 D31 D45 D48	0.71
	5	D0 D3 D31 D32 D21 D48 D45	1.1×10^{-4}
	6	D0 D3 D32 D31 D21 D48 D45	1.1×10^{-4}
2.75	1	D0 D3 D21 D32 D31 D45 D48	1.0
	2	**D0 D3 D21 D31 D32 D45 D48**	0.69
	3	D0 D3 D21 D32 D31 D48 D45	7.9×10^{-3}
	4	D0 D3 D21 D31 D32 D48 D45	5.4×10^{-3}
	5	D0 D3 D31 D32 D21 D45 D48	9.7×10^{-4}
	6	D0 D3 D32 D31 D21 D45 D48	6.6×10^{-4}
1.75*	1	D0 D3 D21 D32 D31 D45 D48	1.0
	2	**D0 D3 D21 D31 D32 D45 D48**	0.24
	3	D0 D3 D31 D32 D21 D45 D48	1.4×10^{-2}
	4	D0 D3 D21 D32 D31 D48 D45	4.0×10^{-3}
	5	D0 D3 D32 D31 D21 D45 D48	3.0×10^{-3}
	6	D0 D3 D21 D31 D32 D48 D45	1.0×10^{-3}

*: determined using whole-genome PCR. The correct order is given in bold.

to <1. [The mean value of θ between all pairwise combinations of DMD and unlinked globin markers is 0.91.]

Various possible marker orders are then derived; the 6 most probable are listed in Table III. The top 4 in rank order were about equally likely, much more so than any of the others; they differ only in order of the most closely-spaced markers (ie D31−D32 and D45−D48). This highlights the general problems of ordering closely-spaced markers when their neighbours are relatively distant; this is exacerbated when they lie at the end of a group of markers, with linkage data available from only one side.

In classical linkage mapping, the recombination fraction is related to distance by a mapping function that has been refined by the experience of mapping many different loci. As we do not yet have any experience, we derive our mapping function, relating θ to distance, from first principles, using the known fragment size (Materials and methods). The resulting map is gratifyingly similar to the known map (Fig 4A,B); the total map length is slightly shorter and smaller distances tend to be overestimated (eg D45−D48), largely due to our simplified mapping function. [The order of the closely-linked markers in square braces can only be resolved by analyzing >140 aliquots.] In future, a mapping function refined by experience should give distortion-free maps, as in classical linkage mapping.

In practice we need to know when to stop analyzing more aliquots. As the number increases (in blocks of 70 aliquots/microtitre plate, allowing for controls), the number of map orders diminishes, until the only remaining ambiguities involve the ordering of very tightly-linked markers (eg D31−D32). For example, after analyzing 70 aliquots, there were 35 orders with probabilities >0.01 relative to the likeliest. [*A priori*, there are 7!/2=2520 different map orders for the 7 DMD markers.]

Effect of fragment size

Computer simulations showed that using longer fragments allowed linkage to be demonstrated more accurately over greater distances. Therefore the experiment described above (involving ~1.75 Mbp fragments) was repeated using ~2.75 Mbp fragments (derived from the same gel) to give analogous lod scores and θ values (Fig 3A,D; circles), rank orders (Table III) and map (Fig 4C). The longer fragments allow linkage over greater distances to be detected more reliably: at the greatest inter-marker distance tested (D0−D48 is 1.24 Mbp), the lod (3.26) was greater than the highest 'background' lod (1.53) between unlinked markers (Fig 3A; circles).

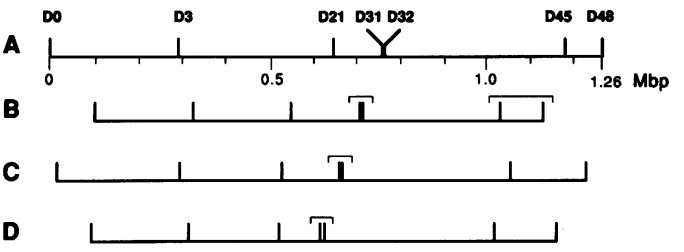


Figure 4. (A) Known map of DMD locus, plus maps (drawn to the same scale) derived by happy mapping using fragments of (B) ~1.75 Mbp, (C) ~2.75 Mbp and (D) ~1.75 Mbp using whole-genome PCR. Marker order in each case was the same, except where a square brace indicates a pair of markers (eg D31, D32) could be interchanged.

The analogous experiment involving smaller ~0.05−0.4 Mbp fragments is illustrated in Fig 3B. Here fragments were not resolved using a pulsed-field gel; rather, DNA was prepared by a conventional method which shears it, and then simply diluted to give aliquots containing ~1 haploid equivalent. Although the correct marker order was deduced using only 140 aliquots as before, this correct map was only marginally more likely than many others (ie there were 192 marker orders with probabilities of >0.1 relative to the likeliest); therefore, no map is presented. However, linkage between markers over distances up to ~0.2 Mbp could be unambiguously detected by this simple procedure (eg Fig 3B,E; circles).

Sonication destroys linkage

As might be expected, sonicating DNA into <0.01 Mbp fragments destroyed all linkages except those between D31 and D32, which are <5 kbp apart (Fig 3B,E, triangles).

Mapping hundreds of markers

Two approaches can be used to map >>9 markers simultaneously. First, the number of primer pairs used during the critical first-phase amplification can be increased; we find no reduction in detection efficiency as the number increases from 1 to 10, so the use of more is feasible. As products of the first-phase PCR may be split and sub-sets of markers amplified in the second phase, the number of markers resolvable by electrophoresis need not be limiting. However, there must be an upper limit to the number of markers which can be amplified from one set of aliquots in this way. Further markers must be mapped from additional sets, giving independent maps that must be unified by many more experiments.

The second approach—a modification of the basic strategy—circumvents problems associated with unifying separate maps. DNA is divided into aliquots as before, and then **all** DNA in each aliquot—rather than just the markers—is amplified by a 'whole-genome' PCR, using low-stringency conditions and a mixture of all possible 15-mers (16). Each amplified aliquot now contains many copies of all sequences originally present, and can be sub-divided into sub-aliquots, each of which can be analyzed

for several markers using specific primers. Results from sub-aliquots can be pooled during analysis, giving the same result as if all markers had been analyzed simultaneously in the original aliquots. Because the **same** primary aliquot is screened for **all** markers, a unified map of all the markers can be constructed.

This approach, then, involves addition of a whole-genome PCR before stage 3 in Fig 1. Results are illustrated using 180 primary aliquots containing ~1.75 Mbp fragments taken from the same pulsed-field gel used for Fig 3A. After whole-genome PCRs (50 cycles), each aliquot was sub-divided into 20, giving 20 identical sets of sub-aliquots. One set was used to map the same markers used above, giving lod scores and θ values (Fig 3C,F; circles), rank orders (Table III) and map (Fig 4D). Results are essentially similar to those obtained without the whole-genome PCR. Now the remaining 19 sets of sub-aliquots can be used to map a further 19 sets of different markers; linkage information on all these markers can then be included in one map.

The reliability of whole-genome amplification is illustrated in Fig 2B. From each of 4 aliquots amplified by whole-genome PCR, 4 sub-aliquots were taken and DMD and β-globin markers amplified with specific primers. In all cases but one, the four sub-aliquots are identical (as expected), with the exception of an additional band in lane 2, representing a contaminant. Use of three—rather than two—phases of PCR generates higher backgrounds against which authentic marker bands can nevertheless be distinguished. The procedure is also more susceptible to contamination, increasing the number of aliquots that must be analyzed (in this case we used 180, rather than 140). The disadvantages are, of course, offset by the advantage of constructing a unified map.

In separate experiments, we have shown that amplified products of the whole-genome PCR can be sub-divided into 1000 sub-aliquots without reducing the detection efficiency, making it possible to analyze 1000 sets of markers simultaneously (ie ~10,000 markers). We have not yet determined the maximum number of sub-aliquots that can be derived from any one primary aliquot, but if the random primer used for the whole-genome PCR contains a unique sequence at its 5′ end, then resulting amplimers can be further amplified efficiently using a primer complementary to the unique sequence under stringent conditions. This should give many millions of copies of all markers in the original aliquots, permitting sub-division into many tens of thousands of sub-aliquots. Sets of these aliquots could become a central resource akin to a clone library or a family panel used for classical linkage studies. Sets could be sent to individual researchers so that they could screen markers that were of particular interest, using their own primers. Then co-segregation frequencies would be pooled centrally, allowing a composite map of all the markers analyzed to be built up efficiently.

DISCUSSION

The approach has a number of advantages. First, it is general, easily automated and applicable to diploid or haploid cells from an individual of any organism. Second, by varying the frequency of breakage (by shearing, γ-irradiation), loci spaced from a few kilobases to several megabases apart can be mapped. In practice, the upper limit of resolution afforded by pulsed-field gels (currently ~10 Mbp), coupled with the problem of DNA breakage during electrophoresis (discussed in Materials and methods), limits our approach to mapping up to ~5 Mbp. However, preliminary experiments suggest the approach can be extended to tens of Mbp using chromosomes broken by shearing and separated into aliquots containing ~1 haploid equivalent using a flow sorter (not shown). We anticipate that the approach will be widely used to map between 0.1–2 Mbp, a range that includes most YAC inserts and that is relatively inaccessible to other approaches. Note that although the increasing size of YAC inserts makes it easier to build contigs, it makes it correspondingly harder to map markers within the YAC. [Restriction mapping YACs of this size remains very difficult in practice; there are, for example, few such maps. And once a restriction map has been generated, markers must still be placed on the map.] Another use would be in closing final gaps in maps, for example across unclonable regions in contig maps. Third, the approach is applicable to both non-polymorphic markers (eg cDNAs) and most polymorphic markers (RFLPs, microsatellites). With RFLPs, we design primers against non-polymorphic flanking regions; with microsatellites, we recognize length-variants as different alleles of the same markers. Hence, we can use a wide variety of physical and genetic markers. Fourth, the approach does not rely on cloning with its associated problems (eg non-clonable sequences, chimerism).

Disadvantages include the need to avoid contamination by foreign DNA during amplification of single molecules and—in common with all physical mapping strategies (but in contrast to classical linkage mapping)—the inability to map phenotypes directly. Our strategy also yields information on marker position, rather than a set of ordered clones.

In its present form, the method could be used to map at high density even the largest human chromosome. A single set of ~200 aliquots including positive and negative PCR controls, amplified by whole-genome PCR, would be split into 1000 replicate sets of sub-aliquots, the sets sent to separate research groups for marker screening, and results pooled centrally. If each set were analyzed for 5 markers (half the number we have shown to be feasible), a map of 5000 markers would result. Alternatively, all analysis could be performed by a single group, entailing 200 whole-genome PCRs and $1000 \times 200 = 200{,}000$ nested PCRs and gel tracks. Even without automation, an individual can perform and analyze $4 \times 96 = 384$ PCRs per day, making this feasible for a group of 3 people within 1–2 years. The average marker spacing along a 150 Mbp chromosome would be 30 kbp. If markers were scattered randomly, no inter-marker gaps greater than 250 kbp would be expected; as we can detect linkage over >1.2 Mbp, even a non-random scattering of markers should leave few, if any, gaps to be closed by other methods. This amount of effort compares favourably with that of Chumakov *et al.* (22), who required ~355 PCRs per marker to screen a YAC library. Our strategy can also be used to map ~10,000 markers throughout the human genome, for example sequences encoding cDNAs (eg ref 17).

Until now, mapping strategies have been sufficiently labour-intensive that efforts have inevitably concentrated on the human genome. We hope that HAPPY mapping adds another powerful and easy method to our existing repertoire, making it possible to map genomes of different organisms for which current approaches are uneconomical.

ACKNOWLEDGEMENTS

This project was supported by the Wellcome Trust, the human genome mapping project of the Medical Research Council and the Cancer Research Campaign. We particularly thank Dr

M.Koenig for supplying us with unpublished DMD sequences and Prof. J.Edwards, Dr. N.Carter, Dr. K.Kolble, D.de Vos and M.Eagle for their help.

REFERENCES

1. Cox, D.R., Burmeister, M., Price, E.R., Kim, S. and Myers, R.M. (1990) Science *250*, 245–250.
2. NIH/CEPH collaborative mapping group. (1992) Science *258*, 67–86.
3. Bellanne-Chantelot, C. *et al.* (1992) Cell *70*, 1059–1068.
4. Wyman, A.R., Wolfe, L.B. and Botstein, D. (1985) Proc. Natl. Acad. Sci. USA *82*, 2880–2884.
5. Little, P. (1992) Nature *359*, 367–368.
6. Drmanac, R., Petrovic, N., Glisin, V. and Crkvenjakov, R. (1986) Nucl. Acids Res. *14*, 4691–4692.
7. Schwartz, D.C. and Cantor, C.R. (1984) Cell *37*, 67–75.
8. White, R. and Lalouel, J.-M. (1988) Ann. Rev. Genet. *22*, 259–279.
9. Lawrence, J.B., Singer, R.H. and McNeil, J.A. (1990) Science *249*, 928–932.
10. Engh, G. van den, Sachs, R. and Trask, B.J. (1992) Science *257*, 1410–1412.
11. Li, H., Gyllensten, U.B., Cui, X., Saiki, R.K., Erlich, H.A. and Arnheim, N. (1988) Nature *335*, 414–417.
12. Dear, P.H. and Cook, P.R. (1989) Nucl. Acids Res. *17*, 6795–6807.
13. Cui, X., Li., H., Goradia, T.M., Lange, K., Kazazian, H.H., Galas, D., Arnheim, N. (1989) Proc. Natl. Acad. Sci. USA *86*, 9389–9393.
14. Koenig, M. *et al.* (1989) Am. J. Hum. Genet. *45*, 498–506.
15. Den Dunnen, J.T., Grootsholten, P.M., Bakker, E., Blonden, L.A.J., Ginjar, H.B., Wapenaar, M.C., Paassen, H.M.B., van Broeckhoven, C., Pearson, P.L. and van Ommen, G.J.B. (1989) Am. J. Hum. Genet. *45*, 835–847.
16. Zhang, L., Cui, X., Schmitt, K, Hubert, R., Nividi, W. and Arnheim, N. (1992) Proc. Natl. Acad. Sci. USA *89*, 5847–5851.
17. Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.N., Utterback, T.R., Nagle, J.W., Fields, C. and Venter, C. (1992) Nature *355*, 632–634.
18. Cook, P.R. (1984) EMBO J. *3*, 1837–1842.
19. Cook, P.R. and Brazell, I.A. (1976) J. Cell Sci. *22*, 287–302.
20. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) 'Molecular cloning'. 2nd edition. Cold Spring Harbor Laboratory Press.
21. Ott, J. (1986) In 'Human genetic diseases; a practical approach', ed Davies, K.E.. IRL Press, Oxford.
22. Chumakov, I. *et al.* (1992) Nature *359*, 380–387.
23. Koenig, M., Monaco, A.P. and Kunkel, L.M. (1988) Cell *53*, 219–228.
24. Chamberlain, J.S., Gibbs, R.A., Ranier, J.E., Nguyen, P.N. and Caskey, C.T. (1988) Nucl. Acids Res. *16*, 11141–11156.
25. Klamut, H.J., Gangopadhyay, S.B., Worton, R.G. and Ray, P.N. (1990) Mol. Cell. Biol. *10*, 193–205.
26. Li, Q., Powers, P.A. and Smithies, O. (1985) J. Biol. Chem. *260*, 14901–14910.