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 haploid equivalents of DNA and the polymerase 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 $\gamma$-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 $\approx 3 \times$ longer than the average marker spacing. $\gamma$-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
2-phase PCR used in Fig 2A was as follows. After a first-phase was used for first- and second-phase PCRs respectively. A typical PCR (see below). Taq polymerase from Cetus and Boehringer

Primers (19-21mers) were designed to have ~50% GC with and Taq polymerase as HTC1 (pH 8.3 at 20°C, 50mM KC1 and 200μM of each dNTP

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, MgCl2 and Taq polymerase as indicated. Table I gives primer sequences and their derivation.

Having established that markers could be detected efficiently, the DNA content of aliquots was 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 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 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 samples (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 MgCl2, 0.75μM primer N15 (a mixture of all 15-mers), 0.5μl 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;
or Z and B, are low. (4) A map can be deduced from the lod scores. or B more frequently then expected by chance, so lod scores between Z and A,
to occur together in the aliquots. Markers A and B tend to occur together, giving
by P.H.D. in Turbo Basic (Borland). A data entry program
be constructed, in which 'lod scores' reflect the tendency of pairs of markers
markers are present in each aliquot is determined (using PCR). A table can then
Figure 1. Mapping strategy illustrated using three markers, A and B (linked)
markers, AB, the number of aliquots containing both (AB), one
markers) and Z (remote) on the two chromosomes of a diploid cell. (1) DNA is prepared
Poisson distribution of markers. Then, for each possible pair of
in each aliquot. From the frequency of occurrence of all markers,
the mean DNA content per aliquot, assuming a
probability of obtaining the observed set of results is calculated
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
the X chromosome were analyzed (DMD markers are prefixed
levels detectable after ethidium-staining using a 'second-phase'
primers, one nested within the other. After a 'first-phase'
be amplified using the PCR to give copies that can be identified,
concentrations.
the mapping function relating θ to distance was determined
from first principles. In the case where DNA from which the
the aliquots are taken is of a single size (S), distance (D) is directly
proportional to θ up to a value of θ=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.
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
(e.g 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
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 that 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 (401/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 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 aliquots used in this experiment. Lanes 1–8 contain markers. The photograph therefore shows 4 sets (lanes 1–4, 4–8, 9–12, 13–16) of amplimers from each of the 4 original aliquots.

Markers: name and amplifier length (internal primers). Primers: L and R denote 5' and 3' ends relative to published sequence; the position is that in the published genome equivalent of ~1 Mbp DNA. The 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 (401/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 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 aliquots used in this experiment. Lanes 1–8 contain markers. The photograph therefore shows 4 sets (lanes 1–4, 4–8, 9–12, 13–16) of amplimers from each of the 4 original aliquots.

Markers: name and amplifier length (internal primers). Primers: L and R denote 5' and 3' ends relative to published sequence; the position is that in the published genome equivalent of ~1 Mbp DNA. The 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 (401/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 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 aliquots used in this experiment. Lanes 1–8 contain markers. The photograph therefore shows 4 sets (lanes 1–4, 4–8, 9–12, 13–16) of amplimers from each of the 4 original aliquots.

Markers: name and amplifier length (internal primers). Primers: L and R denote 5' and 3' ends relative to published sequence; the position is that in the published genome equivalent of ~1 Mbp DNA. The 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 (401/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 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 aliquots used in this experiment. Lanes 1–8 contain markers. The photograph therefore shows 4 sets (lanes 1–4, 4–8, 9–12, 13–16) of amplimers from each of the 4 original aliquots.

Markers: name and amplifier length (internal primers). Primers: L and R denote 5' and 3' ends relative to published sequence; the position is that in the published genome equivalent of ~1 Mbp DNA. The 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 (401/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 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 aliquots used in this experiment. Lanes 1–8 contain markers. The photograph therefore shows 4 sets (lanes 1–4, 4–8, 9–12, 13–16) of amplimers from each of the 4 original aliquots.

Markers: name and amplifier length (internal primers). Primers: L and R denote 5' and 3' ends relative to published sequence; the position is that in the published genome equivalent of ~1 Mbp DNA. The 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 (401/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 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 aliquots used in this experiment. Lanes 1–8 contain markers. The photograph therefore shows 4 sets (lanes 1–4, 4–8, 9–12, 13–16) of amplimers from each of the 4 original aliquots.
Figure 3. How DNA fragment size affects the relationship of (A–C) lod score or (D–F) \( \theta \) 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 \( \theta \) values between them plotted against inter-marker distance. Many of the 14 data points (between DMD and globin markers) at \( \infty \) are superimposed. (A,D). 1.75 and 2.75 Mbp fragments derived by running \( \gamma \)-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). (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.

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, \( \theta \), 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. \( \theta \) 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, \( \theta \) 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 \( \theta \) 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 \( \theta \).
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.

<table>
<thead>
<tr>
<th>Size (Kb)</th>
<th>Rank</th>
<th>Marker Order</th>
<th>Relative Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.75</td>
<td>1</td>
<td>D0 D3 D21 D31 D32 D48 D49</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>D0 D3 D21 D31 D32 D48 D49</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>D0 D3 D21 D32 D48 D49</td>
<td>1.1 x 10^-4</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>D0 D3 D32 D48 D49</td>
<td>1.1 x 10^-4</td>
</tr>
<tr>
<td>2.75</td>
<td>1</td>
<td>D0 D3 D21 D31 D32 D48 D49</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>7.9 x 10^-2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>5.4 x 10^-2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>9.7 x 10^-3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>D0 D3 D32 D48 D49</td>
<td>6.6 x 10^-4</td>
</tr>
<tr>
<td>1.75</td>
<td>1</td>
<td>D0 D3 D21 D31 D32 D48 D49</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>1.4 x 10^-2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>4.0 x 10^-2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>D0 D3 D31 D32 D48 D49</td>
<td>1.0 x 10^-3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>D0 D3 D32 D48 D49</td>
<td>1.0 x 10^-3</td>
</tr>
</tbody>
</table>

*: 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 underestimated (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).

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 YA-
C 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 1000×200=200,000
nested PCRs and gel tracks. Even without automation, an
individual can perform and analyze 4×96=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 would 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