Published online 27 April 2011
Nucleic Acids Research, 2011, Vol. 39, No. 13 e85
doi:10.1093/nar/gkr227

Single-molecule analysis of genome rearrangements in cancer

Jessica C. M. Pole[1], Frank McCaughan[2], Scott Newman[1], Karen D. Howarth[1], Paul H. Dear[2] and Paul A. W. Edwards[1,*]

[1]Hutchison/MRC Research Centre and Department of Pathology, University of Cambridge, Hills Road, Cambridge, CB2 0XZ and [2]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

Received January 1, 2010; Revised March 29, 2011; Accepted March 30, 2011

ABSTRACT

Rearrangements of the genome can be detected by microarray methods and massively parallel sequencing, which identify copy-number alterations and breakpoint junctions, but these techniques are poorly suited to reconstructing the long-range organization of rearranged chromosomes, for example, to distinguish between translocations and insertions. The single-DNA-molecule technique HAPPY mapping is a method for mapping normal genomes that should be able to analyse genome rearrangements, i.e. deviations from a known genome map, to assemble rearrangements into a long-range map. We applied HAPPY mapping to cancer cell lines to show that it could identify rearrangement of genomic segments, even in the presence of normal copies of the genome. We could distinguish a simple interstitial deletion from a copy-number loss at an inversion junction, and detect a known translocation. We could determine whether junctions detected by sequencing were on the same chromosome, by measuring their linkage to each other, and hence map the rearrangement. Finally, we mapped an uncharacterized reciprocal translocation in the T-47D breast cancer cell line to about 2 kb and hence cloned the translocation junctions. We conclude that HAPPY mapping is a versatile tool for determining the structure of rearrangements in the human genome.

INTRODUCTION

Genome rearrangements, such as chromosome translocations and tandem duplications, play a major role in inherited genetic disease and cancer (1,2). In particular, it has emerged that chromosome translocations and other genome rearrangements play an important role in the common cancers, such as prostate and lung cancer, just as they do in leukaemias and sarcomas (2,3), and duplications and deletions are at least as important as single-nucleotide polymorphisms in constitutional disease (1).

Full analysis of genome rearrangements requires map information, i.e. information about distances between particular sequences (markers) and how they are linked together, as illustrated in Figure 1. Few of the tools available for genomics do this. For example, array-comparative genomic hybridization (CGH) can identify unbalanced breakpoints as steps in the copy-number profile, but does not give information about how the breaks are joined together. Even the new techniques based on sequencing, such as paired-end-read strategies (4–6), identify breakpoint junctions but do not show how these junctions are joined together (Figure 1D–G). In particular, paired-end-read sequencing identifies new junctions created by genome rearrangement, but cannot tell whether these junctions are on the same rearranged chromosome or not. More generally, most available genomic tools provide only local information, which cannot readily be used to determine large-scale organisation.

HAPPY mapping is a genome mapping technique that measures linkage, and hence the physical distance, between markers, over a wide range of distances—from <1 kb to >200 kb (7–9). It was used, for example, in the construction of the framework map of human chromosome 14 (10). HAPPY mapping exploits the fact that when the genome is broken into pieces, neighbouring sequences will tend to remain together more often than distant sequences (Figure 2). A fragmented genome is diluted until samples contain a fraction of a haploid genome. These samples are then assayed for the presence

*To whom correspondence should be addressed. Tel: +44 1223 763338; Email: pawe1@cam.ac.uk
Present address:
Jessica C. M. Pole, BlueGnome Ltd, Breaks House, Mill Court, Great Shelford, Cambridge CB22 5LD, UK.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

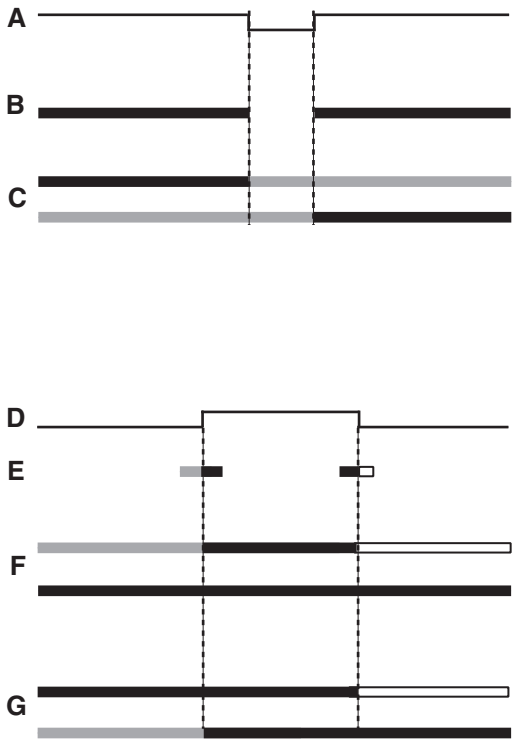


Figure 1. Linkage information is needed to determine the structure of some genome rearrangements. (**A–C**) Two rearrangements that cannot be distinguished by array-CGH. (A) Array-CGH profile that would be obtained for the black chromosome from either (B) or (C). (B) A small deletion in the black chromosome; (C) a reciprocal translocation has resulted in net loss of a small section of the black chromosome at the translocation breakpoints. (**D–G**) Two rearrangements that cannot be distinguished either by array-CGH or by finding breakpoint junctions by large-scale sequencing (5,6), but could be distinguished by HAPPY mapping, provided the breakpoints are within 1 Mb of each other. (D) Array-CGH data for the black chromosome obtained from either (F) or (G); (E) junction sequences obtained from either (F) or (G); (F) a piece of the black chromosome is inserted between pieces of the grey and white chromosomes; while in (G), there are two separate translocations of the black chromosome with respectively the grey and white chromosomes, both including the region between the dotted lines. (The grey and white 'chromosomes' could also represent other parts of the black chromosome, as in an inversion, for example).

of marker sequences. Marker sequences that are close together will be present in the same samples, while distant marker sequences will show an unrelated pattern of distribution amongst samples (Figure 2) (11). Thus, HAPPY mapping is conceptually analogous to mapping by inheritance in families (meiotic mapping) and radiation-hybrid mapping, but works by examining single DNA molecules.

HAPPY mapping has until now been used to map normal genomes [e.g. (8,10)], but it should also be applicable to detecting and analysing rearrangements of a known genome, by looking for changes in expected linkage. It could complement genome-wide methods by determining the physical relationship between rearranged sections of the genome. A chromosome translocation, for example, will create new linkage between the newly juxtaposed sequences, and weaken linkage between the sequences separated by breakage. This would enable HAPPY mapping to distinguish between situations such as those shown in Figure 1. It should also detect balanced rearrangements, such as inversions, which are not detected by array-CGH.

We demonstrate here that HAPPY mapping works well when applied to various types of rearrangement of the human genome, by applying it to cancer cell lines. In particular, we show that it is sufficiently sensitive to map rearrangements in the presence of normal as well as rearranged copies of the genome.

METHODS

HAPPY mapping method

The HAPPY mapping method (7,11) is outlined in Figure 2. DNA is fragmented to a desired size range, highly diluted and samples taken so that approximately half the samples are positive for any given marker, corresponding to about 0.7 haploid genomes per sample (due to the Poisson distribution of molecules). Typically 88 samples are taken (a 96-well plate with eight wells as negative controls), and then the presence or absence of all the marker sequences is assayed in each of the diluted samples.

In the current implementation of HAPPY mapping, polymerase chain reaction (PCR) markers are used. PCR is in two stages: first, primers for all markers are pooled and used in a multiplexed PCR, amplifying all the target sequences simultaneously in each of the 88 diluted samples. The amplification product from each sample is then diluted and divided into a number of replicate 96-well or 384-well plates, and individual markers specifically amplified using semi-nested primers, i.e. one of the primers used in the multiplex PCR, with a new nested primer. These second-round PCR products are scored for presence or absence of the marker.

As illustrated in the first example below, the nearer that two markers are in the genome, the more similar the distribution of positive PCRs over the samples will be.

In addition to providing mapping information, the data provide copy-number information: the number of positive samples for a given marker will reflect the relative copy number of the marker in the genome (Figure 2E and F). This method of measuring copy number changes has been dubbed Molecular Copy-number Counting (MCC) (9,12).

Data analysis

The linkage calculations are essentially those used in genetic linkage mapping and radiation hybrid mapping. Briefly, for any two markers, the software calculates the probability (P) of seeing the experimentally observed degree of co-segregation (i.e. the observed proportion of concordant typing results) when the assumed frequency of physical breaks between the two markers (θ) is varied from 0 (markers are adjacent, never broken apart) to 1 (markers are infinitely far apart, and hence always separated by DNA breaks). It then divides the maximum value of P by the value of P at $\theta = 1$, and takes the logarithm (base 10) of the result. This is the log of odds (LOD) score for linkage between the two markers. Linkage

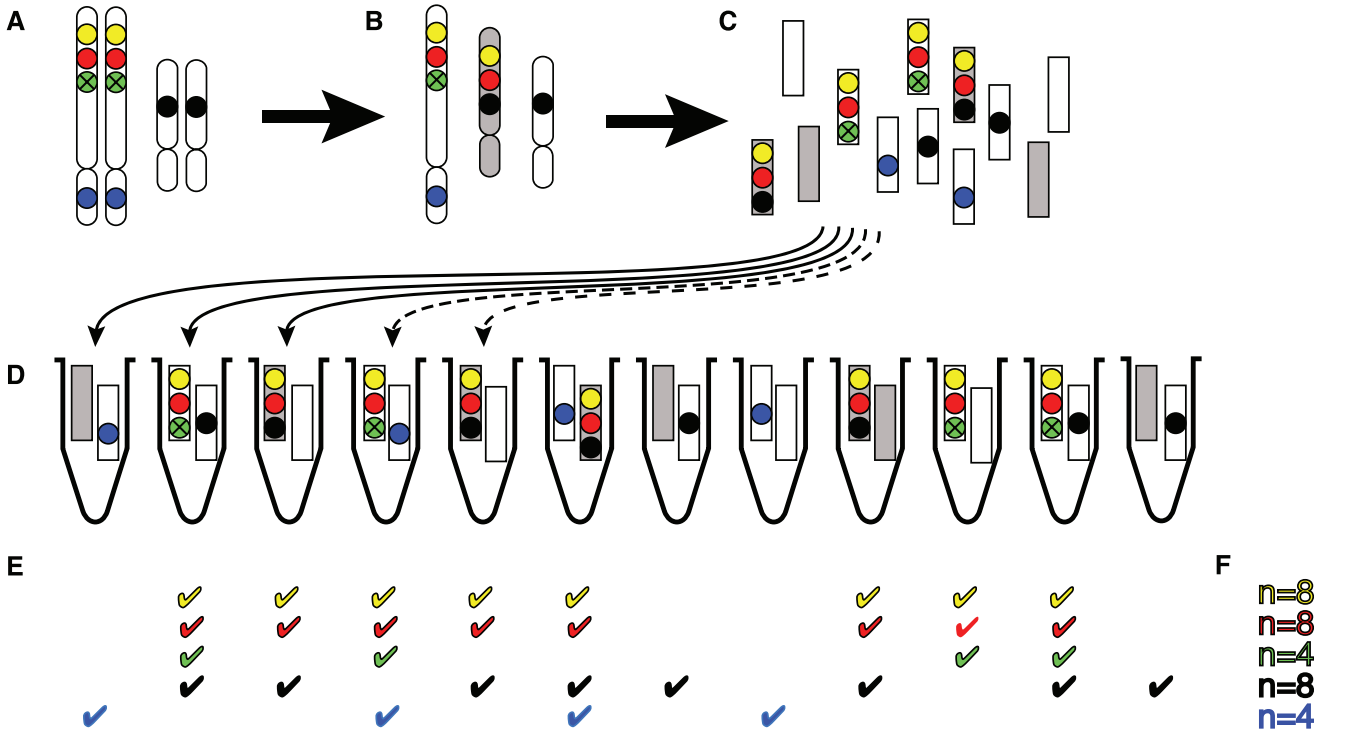


Figure 2. Combined HAPPY mapping and molecular copy-number counting gives positional information and copy number. (**A**) The circles represent marker sequences on chromosomes in a normal cell. (**B**) An unbalanced translocation leaves one copy of each normal chromosome and one copy of a hybrid chromosome (grey). (**C**) DNA is prepared and broken at random; each fragment is large enough to span several markers. (**D**) Fragmented DNA is dispensed into samples (typically 88), each containing about half a genome's worth of fragments. (**E**) Each sample is scored, usually by multiplexed PCR, for the presence (ticks) of each marker. The red and yellow markers always occur together (co-segregate), since they remain adjacent on all chromosomes; the red and green markers cosegregate often (four of the samples) but not always (four samples contain the red marker without the green), since they remain adjacent on the normal chromosome but not the hybrid chromosome; likewise, the red and black markers, which are brought together on the hybrid chromosome, occur together in about half of the samples. Other pairs of markers that are never adjacent (e.g. black and green; blue and red) occur together only occasionally, by chance. (**F**) The number of samples that score positive for any given marker also reflects the relative number of copies in the genome; in this idealized example, the green and blue markers are present at only half the copy-number of the others. This way of measuring copy number has been described previously as molecular copy-number counting (12).

calculations and graphical representations were performed using custom software (Dear,P.H., unpublished data).

The genomic copy number at each marker was calculated from the proportion (P) of the 88 samples found to be positive for the marker sequence using the Poisson equation. The average number of copies of that marker sequence in each sample (copies per sample, or CPA), is

$$\mathrm{CPA} = -\ln(1 - P)$$

CPA is proportional to the number of copies of a given marker sequence per genome.

DNA template preparation

Cell culture has been described (13). DNA was prepared by casting cells in agarose in capillaries, extruding the 'strings' of agarose and repeated extraction with 1% Li dodecyl sulphate, 10 mM Tris buffer, 1 mM ethylenediaminetetraacetic acid (EDTA), pH 8.0 (10). DNA dilutions were generally made by cutting defined lengths from the 'strings', melting them in 0.5× PCR buffer in high-performance liquid chromatography (HPLC) grade water, at 69°C for 5 min, and taking samples directly without further dilution or shearing. DNA was typically diluted to ~0.7 haploid genomes of DNA per 5-µl sample (about 0.4 pg/µl for a diploid genome). To adjust this and allow for aneuploidy of cancer cells, various DNA dilutions were tested in a preliminary PCR. Five-microlitre samples were then dispensed into 88 wells of 96-well microtitre plates, with eight wells as negative controls, without the use of a repeat pipettor to minimize shearing of the DNA, and stored at −80°C under mineral oil. Such dilutions typically show an average DNA fragment size (deduced from PCR results as in Figure 3A) of 100–200 kb, and proved suitable for the experiments reported here. Alternatively, up to 0.5–1-Mb fragments can be preserved by cutting them directly from a pulsed-field gel electrophoresis (10); this method was used for Figure 1A–C, taking nominally 500-kb fragments, because the data were taken from a larger mapping exercise. Smaller DNA could also be used for small-scale mapping, by increasing the shearing. Dilution, optional preamplification and first-round PCRs were set up in a 'clean room' to minimize risk of contamination, which in practice is not a significant problem, as confirmed by

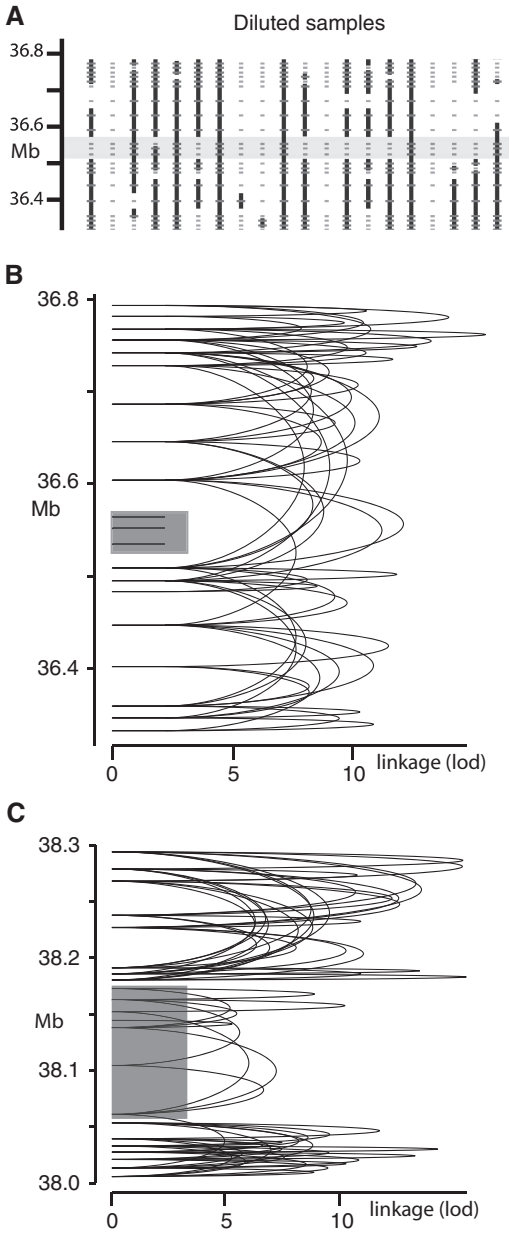


Figure 3. Distinguishing between simple deletion and loss at a rearrangement junction. (**A, B**) HAPPY mapping indicates that copy number loss at 36.5 Mb on chromosome 8 in T-47D is an interstitial deletion; (A) part of raw data. Rows are PCR markers, vertical scale is genome position. Columns are 20 (of 88) diluted DNA samples. Horizontal lines are PCR markers and the individual dashes represent PCR results; positive hits are joined by vertical black bars, representing the presence of DNA. The possible deletion is shaded. Evidence that this is a simple interstitial deletion comes from the concordance of markers that flank the copy number loss, i.e. markers are positive on both sides or negative on both sides. This is expressed (B) as linkage between markers, calculated as the log of odds (LOD) that the markers are linked, using all 88 samples. Horizontal lines represent markers, loops (arcs) represent linkage between them: the stronger the linkage, the further the loop extends to the right. For clarity, only linkage LOD > 7 is shown, and linkage to the markers within the deletion is omitted. (**C**) Copy-number loss at 38.2 Mb on chromosome 8. Linkage LOD > 5 is shown. In contrast to (A, B), this is loss at an inversion junction, and there is no linkage of this strength across the copy number loss, even though, because DNA fragments were large, linkage over flanking sequences extends >100 kb.

the negative control samples, and consistency of maps and copy-number data obtained. The number of samples could also be varied, but 88 proved a good compromise: increasing, say to 2 X 88 increases statistical power, but marker results and resolution are also limited by the performance of individual primers sets and their spacing, so it is usually better to double the number of markers rather than the number of samples per marker. However, where linkage is limited by the size of DNA fragments, the extra statistical power of using 2 X 88 samples is valuable: an example is shown in Figure 6.

Preamplification

Before marker PCR, an optional preamplification can be performed, by random primer extension (PEP-PCR) (14). This provides a number of plates with identical template, so that additional marker sets can be tested sequentially on the same samples (14). This was used in the HCC1187 translocation experiment (Figure 5). PEP-PCR was carried out in a total volume of 7 µl containing 10 µM of degenerate 15-mer PCR primer, 1 × PCR buffer II (Applied Biosystems, City California, USA), 2 mM $MgCl_2$, 200 µM each dNTP and 0.1 U/µl Amplitaq DNA polymerase (Applied Biosystems). Thermocycling conditions were 93°C for 5 min, followed by 50 cycles of 30 s at 94°C, 2 min at 37°C, a temperature ramp of 0.1°C s^{-1} up to 55°C, 4 min at 55°C. The PEP-PCR was then diluted with 200 µl of water and aliquotted 5 µl per well into replicate 96-well plates.

Two-stage PCR

PCR was semi-nested, i.e. primer sets consisted of three primers: the first PCR used a forward-external and reverse primer, the second PCR used a forward-internal and the same reverse primer. Internal amplimer length was designed to be 60–150 bp, and the position of the external primer no more than 150 bp upstream of forward-internal primer. Primers (Supplementary Table S1) were generally designed automatically using custom software (Dear,P.H., unpublished data) against the repeat-masked human reference genome sequence NCBI Build 36. Typically, primer length was 20–23 bp; melting temperature (T_m) 52–60°C [based on $T_m = 2 \times (A+T) + 4 \times (G+C)$]; with at least two guanine or cytosine bases at the 3′ end and at least one at the 5′ end; and no runs of 4 or more of the same base allowed. Some primers were designed manually using Primer3 (http://frodo.wi.mit.edu/), without repeat masking. Primers were then tested by *in silico* PCR for uniqueness (http://genome.ucsc.edu/cgi-bin/hgPcr?command=start). Primers were supplied by Eurofins MWG Operon (Ebersberg, Germany) and Sigma-Aldrich (Poole, UK). A small proportion of newly designed primer sets will fail (11), but the distinction between a failed and deleted marker can be made by testing the markers on normal DNA.

The first of the two PCRs was a multiplex PCR, using a pool of the forward-external and reverse primers for all markers. [Previous work shows that this is robust to over 1000 markers (15).] The first PCR products were diluted and replicated into multiple 96-well plates, each used for a

second round (non-multiplexed) PCR in which separate semi-nested PCRs were performed for each individual marker, using the forward-internal and reverse primers for one marker.

PCR was otherwise conventional. Phase 1 reactions were in 10 µl containing 0.15 µM of each oligo, Gold PCR buffer (Applied Biosystems), 2 mM $MgCl_2$, 200 µM each dNTP and 0.1 U/µl Taq Gold DNA polymerase (Applied Biosystems). Thermocycling conditions were a hot start at 93°C for 9 min, followed by 28 cycles of 20 s at 94°C, 30 s at 50°C and 1 min at 72°C. Products were diluted to 1000 µl with water and 5 µl replicated into fresh microtitre plates for Phase 2 PCR.

The Phase 2 PCRs used the forward-internal and reverse primer for one marker on each 88-well set of diluted phase 1 products. For convenience, the transfers of template were done robotically into 384-well microtitre plates, each plate containing the reactions for four markers. Reaction conditions for the Phase 2 PCR were 1 × PCR Gold buffer, 1.5 mM $MgCl_2$, 200 µM each dNTP, and 1 × EvaGreen dye (Biotium Inc, Hayward, CA, USA), with 1 µM of the relevant forward-internal and reverse primers, and thermocycling at 93°C for 9 min, followed by 33 cycles of 20 se 94°C, 30 s 54°C and 1 min 72°C.

Results were scored either by electrophoresis on polyacrylamide gels or by melting-curve analysis on a real-time PCR thermocycler. For electrophoresis, precast 108-well horizontal 6% polyacrylamide gels (MIRAGE gels, Genetix, Hampshire, UK; or made in house) were run for 10 min at 10 V/cm. Melting-curve analysis was in an ABI 7900HT (Applied Biosystems) with the manufacturer's SDS software and using the EvaGreen dye included in the second PCR reactions.

RESULTS

A copy-number loss that was a simple interstitial deletion

We first tested HAPPY mapping's ability to distinguish between different situations that result in local loss of copy number: loss can result from a simple interstitial deletion, or from loss of sequence at the breakpoints of a rearrangement such as a translocation or inversion (Figure 1A–C). HAPPY mapping measures linkage between markers and a simple interstitial deletion will preserve or even increase the linkage between the markers that flank it, whereas pieces of chromosome that are separated by a translocation or inversion will no longer be linked (Figure 1A–C).

We identified a copy-number loss that could be a small interstitial deletion, in the T-47D breast cancer cell line. This was an incidental finding during exploratory HAPPY mapping of the inversion described below (Figure 4) at low resolution, with DNA fragments of up to 500 kb and markers designed at 40-kb spacing. Additional markers were added, and mapping repeated.

Three consecutive PCR markers on chromosome 8, at 36.528–36.556 Mb (all genomic positions are on the reference genome NCBI Build 36.1, Hg18), gave a copy number half that of the flanking markers, at 36.504 and 36.593 Mb, corresponding to a fall in copy number from 4 to 2 [T-47D is pseudo-tetraploid (16)] (Figure 3).

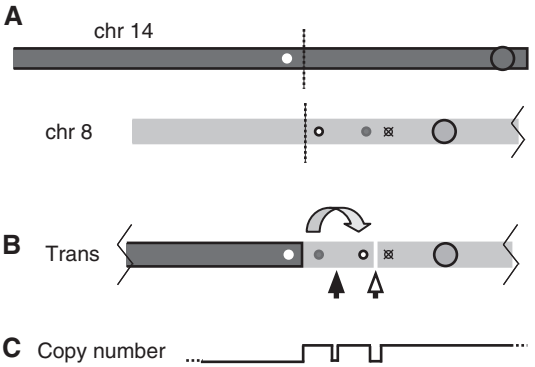


Figure 4. Complex translocation of chromosomes 8 and 14 in T-47D with inversion and losses of chromosome 8. This was characterized previously, by fluorescence *in situ* hybridization (FISH) and cloning of junction sequences (13). (**A**) Normal chromosomes 14 (long arm only) and 8 (short arm only). Dotted lines, breakpoints of translocation; small circles, key points on the chromosomes; large circles, centromeres. (**B**) Rearranged chromosome. The distal part of 14 has been translocated to 8, but, in addition, 3.5 Mb of chromosome 8 adjacent to the translocation has been inverted, and about 100 kb has been lost at the inversion junction (indicated by the gap and open arrowhead). The black arrowhead marks the further copy-number loss at 36.5 Mb, investigated in Figure 3A, which may be on this chromosome or on the normal chromosome 8. (**C**) Resulting copy-number profile, showing the two copy-number losses. There are two copies of the 8;14 translocation, two normal chromosome 8s and two normal 14 s, per cell, as T-47D is pseudo-tetraploid (13,16). [Adapted from ref. (13)].

Figure 3A shows part of the raw HAPPY mapping data, the first 20 of 88 diluted samples. The three markers within the deletion (grey shading in Figure 3A) were positive in roughly half as many diluted samples as the flanking markers. Linkage along the chromosome was reflected in the agreement between the scores for neighbouring markers, i.e. where a marker was positive in a diluted sample, the flanking markers were usually also positive. This represents the presence of one or more DNA molecules that span those markers.

Evidence that this was indeed a simple deletion came from the markers that flanked the deletion: where a sample showed absence of the deleted region, the flanking markers were usually in agreement, i.e. positive on both sides or negative on both sides. The deleted region was absent in 42 of 88 samples, and in 37 of these 42 the flanking markers were in agreement. (A few results will differ, either because a break in the diluted DNA happens at random to fall between the two markers or because of PCR errors.) This suggested that the flanking pieces of DNA were indeed joined to each other.

The best way to express the results is in terms of linkage, or, more precisely, the probability that two markers are physically linked. This is calculated from the agreement between markers using approaches developed for genetic mapping. Linkage is expressed as $\log_{10}$ of odds (LOD score), so for example, a LOD score of 7 means that the observed results are 10^7 times more likely to have arisen if the markers are linked than if they are not linked. The probabilities of linkage are shown in Figure 3B, for all possible combinations of markers, represented as loops

joining the markers, with the height of the loops indicating strength of linkage. Strong linkage was seen across the deletion, with LOD scores >7.

More precisely, the linkage shows that the breakpoints of the deletion have a high probability of being close to each other. HAPPY mapping cannot exclude that the rearrangement is more complex than a simple deletion, e.g. that there is DNA inserted in between the breakpoints.

A copy-number loss that was not a simple interstitial deletion

A contrasting example was also analysed, where a copy-number loss was not a simple interstitial deletion, but represented loss of sequence at an inversion junction. In T-47D there is copy-number loss at an inversion junction, extending over about 110 kb, at 38.1 Mb on chromosome 8 (13) (open arrowhead in Figure 4). Figure 3C shows HAPPY mapping linkage at this copy-number loss. There is no linkage above the threshold set (LOD > 5) across the junction, whereas there is clear linkage among markers on each side of the region of loss, extending over more than 100 kb. (There remains some linkage across the copy-number loss, because half the copies of chromosome 8 are normal and not affected by the inversion. The strongest LOD across the loss was 3.2, average 1.5. Similarly, there is some linkage between markers in the region of loss and flanking markers, average 1.6, range 0.2–5.05, but it is relatively low because LOD score measures overall concordance and the rearranged copies contribute discordant marker results.)

The inversion also creates a new junction between 38.1 and 34.5 Mb, and this was detected as new linkage (Supplementary Figure S1).

Detection of new linkage at a translocation junction

To show that HAPPY mapping would detect linkage between the newly juxtaposed sequences at the breakpoints of a chromosome translocation, we mapped a translocation junction in the HCC1187 breast cancer cell line that we had previously cloned and sequenced (17) (Figure 5A).

HAPPY mapping was applied to detect the new junction created by the translocation, using PCR markers spaced at approximately at 7 kb intervals around the breakpoints marked A and C in Figure 5. It showed linkage across the new junction, e.g. between marker 1a06 on chromosome 1 and marker 2a11 on chromosome 8 (Figure 5B). There was also a loss of linkage between markers separated by the translocation: e.g. on chromosome 8 the loss of linkage was between markers 2a11 and 1g06 (Figure 5B).

Copy-number counting also supported our previous findings (17) that the rearrangement is unbalanced for chromosome 1, with a change in copy number from 4 to 2 copies at the break, while the rearrangement is balanced for chromosome 8, with 3 copies throughout the marker set (Figure 5B). The 4 to 2 step could have been used on its own to map the breakpoint on chromosome 1 (12), while, as expected, the copy numbers on chromosome 8 were essentially constant at an intermediate level.

This application showed that we could detect gain of linkage between normally distant regions of the genome, despite a background of other, unrearranged copies. In addition, it confirmed that there were no additional rearrangements at this junction, such as flanking inversions or deletions.

Assembling new junctions into a genome map

As explained in Figure 1, and illustrated in Figure 6, when rearrangements are discovered by sequencing of genome fragments, particularly using the 'paired end read' strategy (2,5,6), junctions between distant parts of the normal genome are found, but assembling these new junctions into a map of the genome may not be possible without additional information. One simple example of this is where two new junctions could be either close together on the same rearranged chromosome or on two completely different chromosomes.

Constructing a correct map requires long-range information, such as HAPPY mapping can provide. HAPPY mapping could exploit junction-specific PCR markers, that is, primer sets that span a newly discovered junction. Furthermore, the linkage should then be clearer than for markers that are also detecting unrearranged copies.

To illustrate this consider Figure 6, which shows a typical example of this problem. Three rearrangement junctions that might be linked, J1–J3, have been found in the HCC1187 cell line, either by paired end read sequencing (18) or by array painting (Figure 6A) (17,19). There are four ways these junctions could be assembled into a map (Figure 6B). Two junctions, J1 and J2, map about 1.4 kb apart on chromosome 11, and could reflect a 1.4-kb fragment of chromosome 11 inserted into an 11;16 translocation junction (Figure 6B, cases i or iv). Such fragments have been named 'genomic shards' (20). However, the junctions could equally well be on separate rearranged chromosomes (Figure 6B, cases ii or iii). Similarly, junctions J2 and J3 either could be at opposite ends of a 55-kb insert of chromosome 16 into chromosome 11 (Figure 6B, iii or iv) or could be on two different chromosomes, the products of an approximately-reciprocal translocation (Figure 6B, i or ii).

We applied HAPPY mapping to map these junctions, designing junction-specific primer sets by placing primers on opposite sides of the junctions. To decide whether J2 and J3 are on the same chromosome, 55 kb apart, or on separate chromosomes and hence unlinked, we needed to determine the expected linkage over 50–60 kb, on a control region. This control region needed to be at the same copy number in HCC1187 as the junctions, i.e. one copy. A single-copy region on chromosome 13 was chosen (17), and primers designed over an interval of 90 kb. For simplicity, we used DNA diluted from stored agarose strings; as quite a lot of DNA fragments in such material would be <50 kb long (Figure 6C), we improved statistical robustness of the mapping by doubling the sample number to 2×88 samples.

Figure 6 shows the various possible arrangements of the junctions, and the control measurements of linkage versus

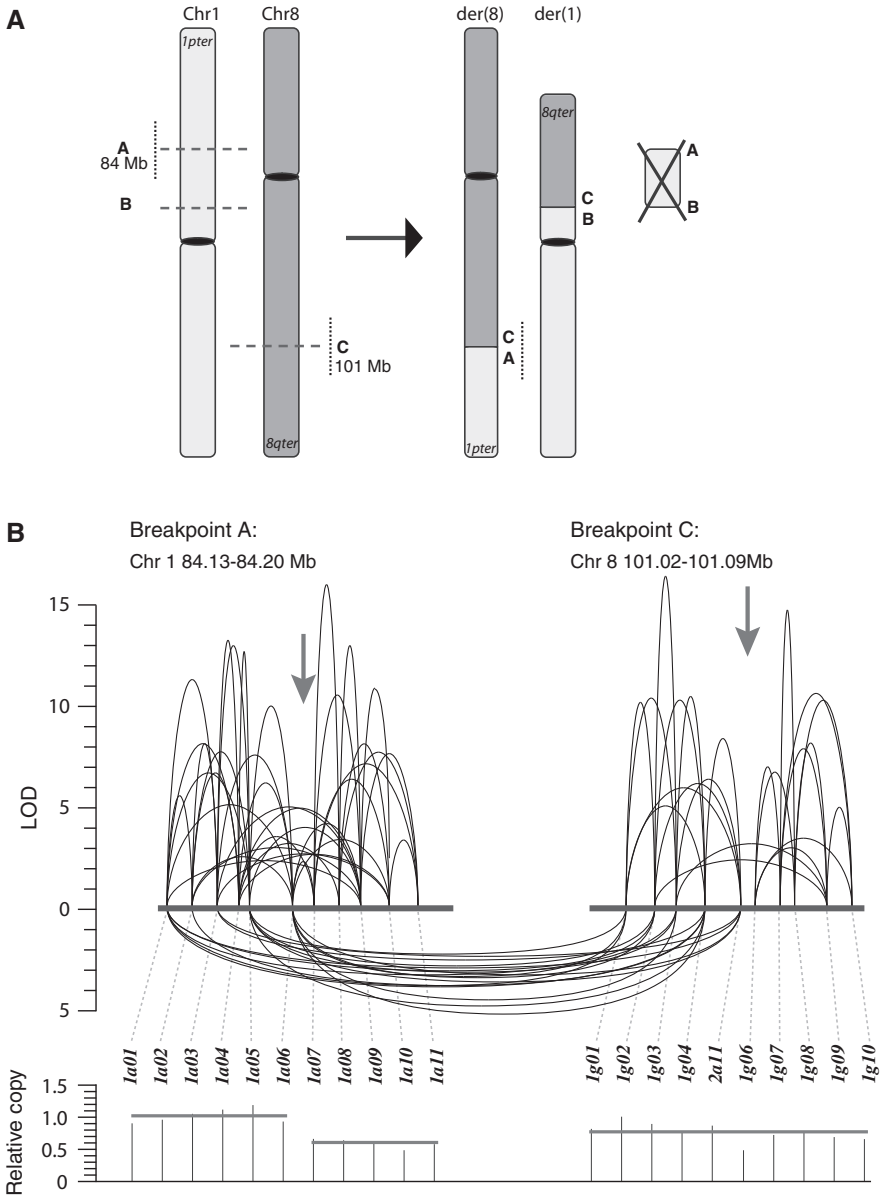
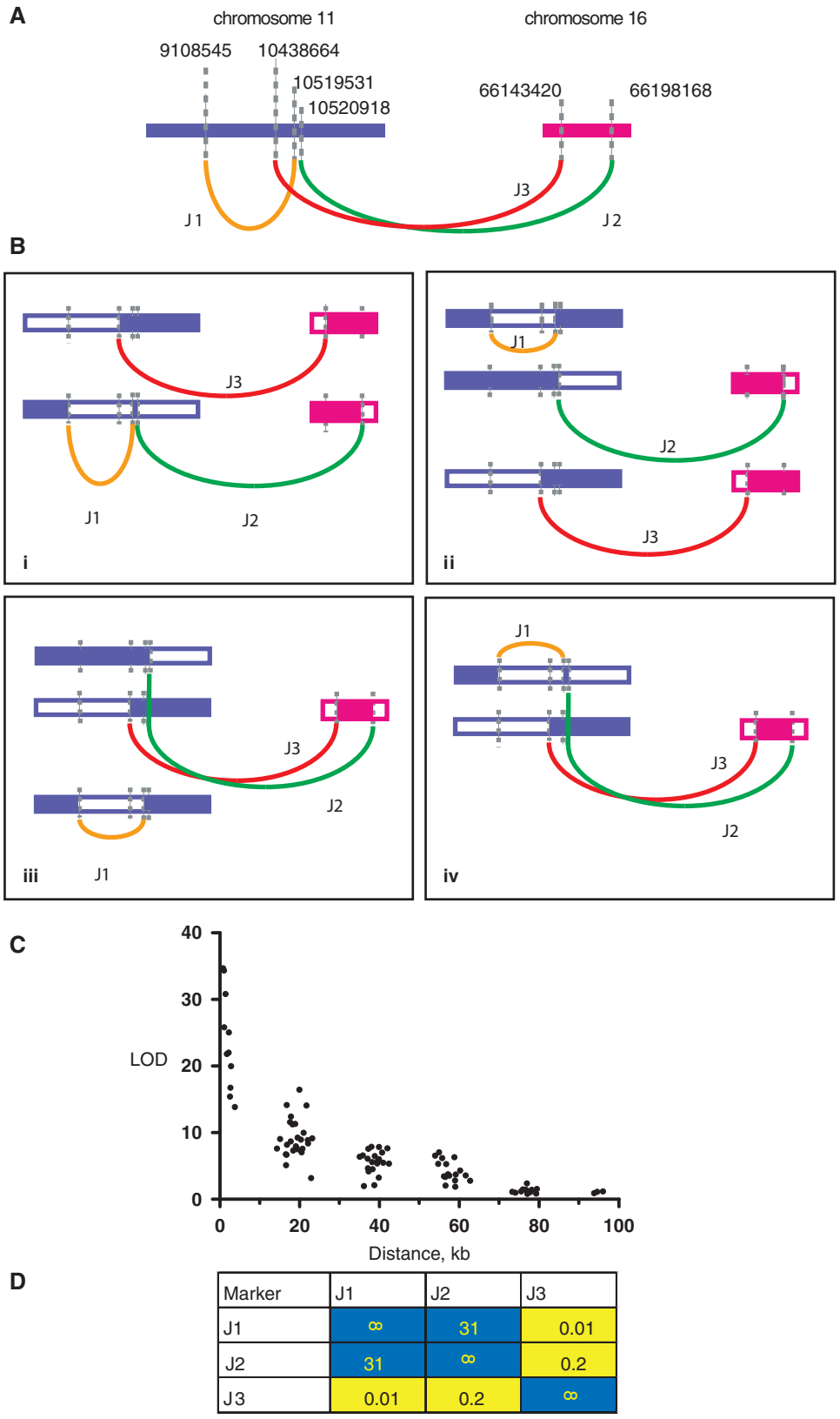


Figure 5. Detection of changed linkage at a chromosome translocation. (**A**) Schematic of translocation between chromosome 1 and 8, previously mapped using array painting (17), in breast cancer cell line HCC1187. On the left, normal chromosomes 1 and 8 with points of breakage A, B and C (dashed horizontal lines). Right, resulting products and piece of chromosome 1 that is lost. Dotted vertical lines show (not to scale) markers used for mapping. There are two copies per cell of both translocated chromosomes, and additional untranslocated copies of the breakpoint regions. In consequence, there are three copies throughout the chromosome 8 region, but a transition from four to two copies at the chromosome 1 breakpoint. (**B**) HAPPY mapping around the breakpoints. The heavy horizontal lines represent the reference genome to scale; markers are named below. The arcs indicate linkage between markers as LOD scores as before; linkages between markers normally on the same chromosome are above the line; linkages between markers on different chromosomes are below. LOD values >2 are shown. The arrows indicate the largest losses of linkage compared to that expected for the normal genome, which are likely to indicate the breakpoints. Copy-number results are shown below the marker names as vertical lines indicating relative copy number, expressed as mean number of copies per sample of the mapping panel; the horizontal lines joining these show the average values across the markers they cover.

distance on chromosome 13. Junctions J1 and J2 showed very high linkage, equivalent to an adjacent position in the genome. J2 and J3, however, were unlinked, whereas at 55-kb separation they would have been expected to show an LOD score of around 3.5 (range 1.9–7 for 17 control primer pairs). Thus, in a quite simple experiment, we have strong evidence that the junctions are arranged as in Figure 6B(i): J1 and J2 flank a 'genomic shard' inserted into a 11;16 translocation junction; while J2 and J3, counter-intuitively, do not represent flanking junctions of an insert, but are on the two separate products of a near-reciprocal translocation. Both these results are in agreement with our molecular cytogenetic analysis of this cell line: specifically, we have previously obtained a PCR product that spans the J1–J2 combined junction, and have shown that J2 and J3 are on separate products of an



D

Marker	J1	J2	J3
J1	∞	31	0.01
J2	31	∞	0.2
J3	0.01	0.2	∞

Figure 6. Establishing linkage between rearrangement junctions found by paired-end sequencing. (**A**) Three junctions, J1–J3, which were identified by paired-end sequencing or cloning, that join points on chromosomes 11 and 16. Blue and magenta bars, regions of normal chromosome 11 and 16 showing where the breakpoints of the junctions map. Arcs show the junctions, numbers are genomic positions of breakpoints in basepair. (**B**) Four possible ways in which these junctions could be joined together. The blue and magenta bars correspond to copies of the regions of chromosomes 11 and 16 shown in (A). Coloured-in parts of the bars are present in the postulated rearranged chromosome, whereas regions that are not present are not coloured in. For example, in (B)(i) (which is the correct model), in the upper part, chr. 11 is broken at 10.4 Mb and the fragment extending from

(continued)

11;16 reciprocal translocation (17,19). The molecular cytogenetic approaches used were, however, much more laborious than the HAPPY mapping approach and not suited to scaling up.

Detection and mapping of a previously uncloned reciprocal translocation junction

To apply HAPPY mapping to an unknown rearrangement, we analysed what appeared to be a reciprocal translocation between chromosomes 10 and 20, a t(10;20)(q21;q13.2) (Figure 7A), in the cell line T-47D, aiming not only to map the translocation to the point of cloning, but also to determine whether the breakpoints were joined to each other in the expected way, or whether there were additional flanking rearrangements.

HAPPY mapping was applied across the breakpoints, which we had previously mapped to about 100-kb resolution (Supplementary Figure S2). Initially, markers were spaced roughly every 5 kb, then mapping was repeated with additional markers added around the breakpoints.

Clear new linkage appeared between markers on the two chromosomes, across both translocation junctions, and linkage extended as expected to flanking markers away from the breaks (Figure 7; Supplementary Figure S3 and Supplementary Table 2). There was also the expected reduction in linkage between the markers on each chromosome that were split by the translocation. The breakpoints could be identified either from the new linkage or from the loss of linkage. The linkage also showed that there were no major additional rearrangements, and that the translocation was almost exactly reciprocal.

The breakpoints on chromosome 10 were deduced to be between markers v4-a10 (forward external primer at 57 445 590 bp NCBI Build 36.1) and v4-a12 (at 57 446 736 bp). On chromosome 20, breaks were between markers v3-b02 (at 54 177 356 bp), joined to v4-a12, and v4-b06 (at 54 179 191 bp), joined to v4-a10. Marker v3-b01, at the junction on chromosome 20, was at about half the copy number of its neighbours: a likely explanation was that it was absent from both products of the translocation.

The junctions were cloned by PCR between primers from these marker sets (Supplementary Data). The translocation was exactly reciprocal: the junction on both chromosomes was at chr20: 54 178 034 or 1 bp later, and chr10: 57 446 414. The absence of the v3-b01 marker from the translocation products was explained, as the primers span the breakpoint on chromosome 20.

DISCUSSION

HAPPY mapping as a tool to map genome rearrangements

In order to understand a genomic rearrangement it is generally necessary to construct a map of it. Recent new technologies, such as high-resolution array-CGH and massively parallel sequencing, have improved our ability to detect rearrangements, but they do not provide maps. In particular, high-throughput sequencing, which can detect rearrangements by finding sequences that span rearrangement junctions (2,4–6), does not solve this problem, because sequences can only be assembled unambiguously into longer runs if there is a unique genome order. If some copies of the genome are rearranged while others are intact, or rearranged differently, as usually occurs in human disease, there may be more than one way to assemble junctions into a complete picture that will show how genes are affected. Figures 1D–G and 6 show examples of such ambiguity, and similar situations are not uncommon in constitutional (1) or cancer (19,20,21) rearrangements.

HAPPY mapping is able to map rearrangements by measuring linkage. We showed, for example, that it was able to distinguish between two situations that result in local copy number loss: a simple interstitial deletion and loss of material at an inversion junction (Figures 1 and 3). It was also able to detect change in linkage resulting from translocation and inversion, and we were able to exploit this to clone the breakpoints of a reciprocal translocation in T-47D.

One particular strength of HAPPY mapping is its ability to detect linkage over distances >10 kb. This is particularly relevant to assembling rearrangement junctions into maps, as we illustrated in Figure 6. HAPPY mapping should also be able to reach across complex local rearrangements which quite commonly occur at the junctions of large-scale rearrangements, such as fragments of DNA copied from elsewhere the genome and inserted into the junction of a deletion or tandem repeat, so-called 'genomic shards' (1,20,21). It would also permit mapping of breakpoint junctions at repeat sequences, which tend to be invisible to sequencing-based strategies. For example, the chromosome 10 breakpoint in the reciprocal translocation we mapped is in the middle of an L1 repeat.

Strengths and limitations of the technology

The use of HAPPY mapping to map rearrangements has several strengths. Firstly, resolution can be freely adjusted from >200 kb down to below 1 kb by appropriate choice

Figure 6. Continued
the break towards the q telomere is joined at J3 to a piece of chr. 16 extending from a break at 66.14 Mb towards the q telomere; while on a separate chromosome, chr.11 up to a break at 9.1 Mb is joined to the 1.4-kb fragment of chr. 11 between J1 and J2 (10.5195–10.5209 Mb), and then joined to a piece of chr. 16 extending up to the breakpoint at 66.19 Mb. (**C**) Relationship between linkage and genomic distance, established for markers on a control region of the genome. Note that at around 50–60 kb separation, LOD scores between 1.9 and 7 were obtained between various combinations of control markers. (**D**) LOD scores obtained between the junction-specific markers. Linkage between J1 and J2 had a LOD of 31, i.e. these junctions are very close, whereas no significant linkage was found between J2 and J3, which should have shown a LOD score between 1.9 and 7 if they were 55 kb apart on the same chromosome. These scores support the arrangement shown in B(i), as expected.

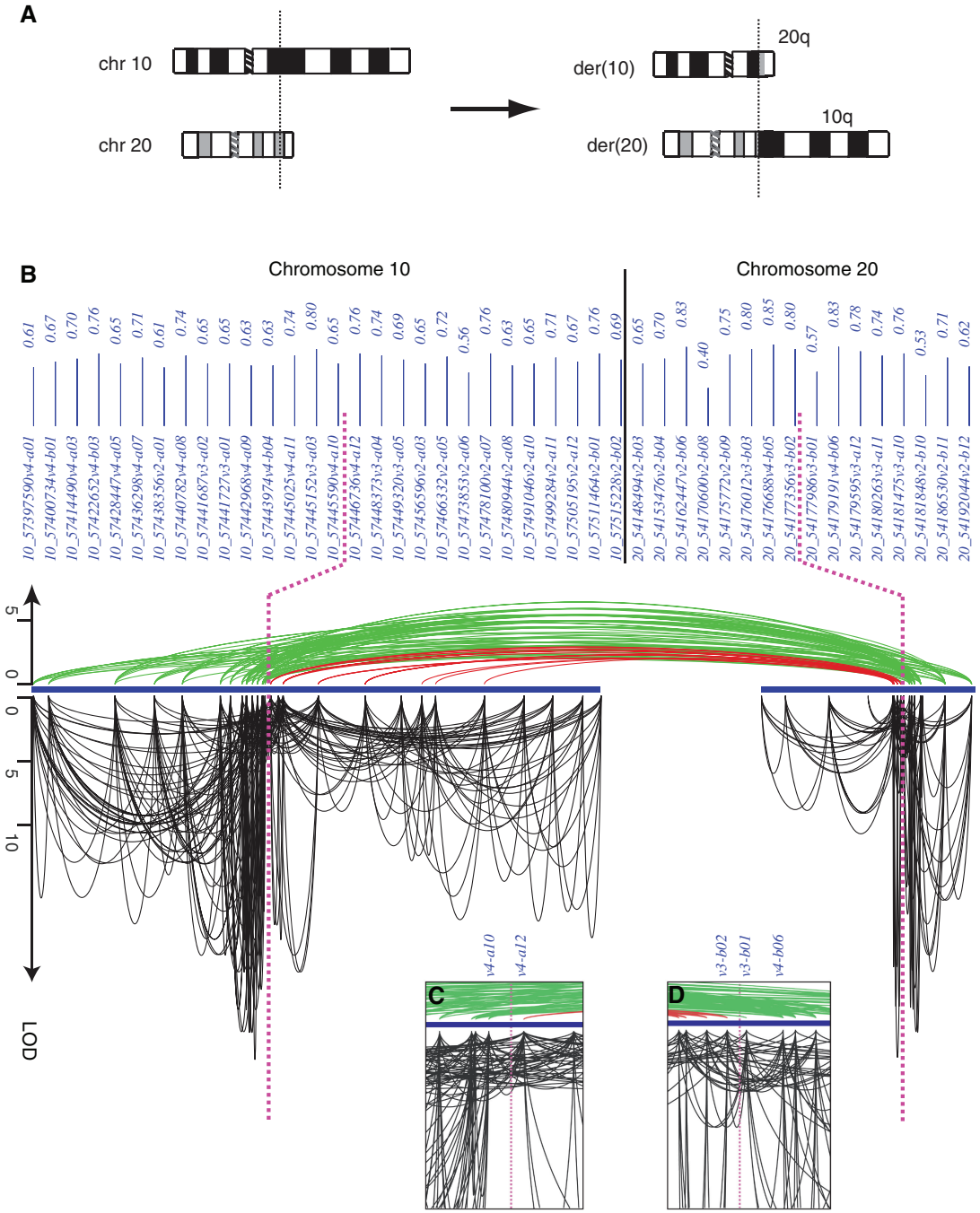


Figure 7. Mapping a reciprocal chromosome translocation t(10;20)(q21;q13.2) in the T-47D cell line. (**A**) Diagram of translocation. (**B**) HAPPY mapping shows linkage of markers on chromosome 10 (left side of diagram) to markers on chromosome 20 (right side), and loss of linkage across the breakpoints on each chromosome (dotted red lines). As in Figure 5, the heavy blue horizontal lines represent the reference genome to scale; the arcs indicate linkage between markers as LOD scores as before; linkages between markers normally on the same chromosome are below the line; linkages between markers on different chromosomes are above. Linkage arcs in green join the two parts of the derivative 10 chromosome, der(10)t(10;20)(q21;q13.2) (i.e. the product chromosome with the chromosome 10 centromere); linkage arcs in red join the parts of the reciprocal product der(20)t(10;20)(q21;q13.2). LOD values >2 are shown. Marker names include their chromosome and position of Fex primer in bp. Above the marker names are vertical lines indicating relative copy number. Insets (**C**) and (**D**): the breakpoint regions with genomic scale stretched 6-fold. Data shown was from a final mapping run in which all markers were used together, for illustration purposes.

of primer sets and the size of DNA fragments. Successive rounds of mapping can home in on rearrangements to higher and higher resolution, as in our last example. Both marker spacing and DNA fragment size can be adjusted. To show linkage between the chosen markers, the DNA has to be of sufficiently high molecular weight to span several markers. The upper limit of resolution is set by the size of DNA fragments that can be selected and

diluted without fragmentation. Fragments of up to around 1 Mb can be prepared by pulsed-field electrophoresis, taking samples that contain a fraction of a genome directly from the gel. These are best combined with markers spaced not more than 200 kb apart (10). In the mapping experiments presented here, the more convenient approach of diluting DNA from melted agarose was generally used, giving quite large DNA fragments, up to hundreds of kilobases, depending on how fresh the DNA preparation is. These show up large-scale changes in linkage, with neighbouring markers confirming each other. The only limitation of such large fragments is that the relative order of markers is less clearly distinguished, and, therefore, local rearrangements, such as small inversions, would be overlooked or poorly resolved. To analyse such local rearrangements, smaller DNA fragments should be used.

Secondly, the method is technologically simple and can be implemented without elaborate equipment. On the scale demonstrated here, the PCRs would be manageable with multi-channel pipettes. Equally, if the PCRs are set up with a robot there is very little hands-on time, and the technique can be scaled up.

Thirdly, only very small amounts of DNA are required. Each sample requires ~0.7 haploid genomes (2.3 pg). Therefore, an experiment with 100 samples would require 0.23 ng DNA, the equivalent of 35 diploid cells or fewer polyploid cancer cells (22).

Among potential limitations, the sensitivity of HAPPY mapping to analyse a rearrangement can be limited by the presence of normal (or differently rearranged) DNA. This should not be an issue for germ-line rearrangements, where equal amounts of rearranged and normal genome are present. In most of our experiments, there was an equal amount of normal and rearranged DNA, and linkage changes were clear. In cancers, however, there may be two or more other copies of the cancer genome or DNA from contaminating non-cancer cells. Where rearranged copies are less than half the total, linkage probability can be maximized by increasing the number of dilution samples scored, as in Figure 6, and by ensuring that the DNA fragments are long enough to give strong linkage over the distances involved.

This problem can however be eliminated where it is possible to use junction-specific markers, as in Figure 6, which completely ignore normal DNA. This will be the favoured approach if sequencing and paired-end sequencing become the dominant way to discover rearrangements. Junctions will be discovered, and junction-specific markers will then permit assembly of the discovered junctions into a map, as in Figure 6.

A further limitation is the cost of deploying PCR-based HAPPY mapping on a large scale, such as the whole genome, which is discussed below.

In the context of tumour biopsies, as opposed to cell lines or germ-line rearrangements, an important strength is that the technology only requires DNA, and only in small amounts. This is in contrast to cytogenetic methods, which require chromosome spreads and therefore dividing cells, and so are often restricted to cell lines (2). HAPPY mapping does require intact DNA, so cannot be applied, over any substantial distance, to DNA derived from formalin-fixed paraffin-embedded material, because such DNA is usually fragmented to <1 kb, but high-quality DNA is often available from snap-frozen tumour material, and should usually be available from patients with germ-line rearrangements.

An issue with tumour biopsies, other than leukaemia samples, will be contaminating normal DNA from non-cancer cells. There are two solutions: use of junction-specific markers as just discussed, and microdissection. Since very little DNA is required, enriching for tumour cells by microdissection is entirely feasible, and can achieve almost 100% tumour cells for many epithelial malignancies (22).

The place of HAPPY mapping in studying rearranged genomes

The tools available for interrogating genome structure can usefully be divided into those that permit genome-wide scans (e.g. cytogenetics, array–CGH, paired-end sequencing) and those that are targeted at specific genomic loci (e.g. FISH). HAPPY mapping, as described here, using PCR, belongs in the second group: it is most suited to detailed analysis of individual rearrangements that have already been discovered by genome-wide technologies.

HAPPY mapping offers the ability to determine the physical relationship between segments of rearranged DNA, on a scale from <1 kb up to about 1 Mb, using small amounts of input DNA. It provides information somewhat analogous to FISH, except that it operates up to 1 Mb, while FISH can map rearrangements in the range 100 kb to 1 Mb on interphase nuclei, and larger than several megabases on metaphase chromosome spreads (e.g. 13). As discussed above, HAPPY mapping complements array CGH and paired-end sequencing (Figure 1). Array-CGH detects breakpoints but cannot tell which breakpoints are joined to which. It also fails to detect balanced rearrangements, including inversions and reciprocal translocations. Resolution is often still limiting: current genome-wide arrays on a single slide, such as the Affymetrix SNP6 array, will often not map breakpoints well enough for immediate cloning by PCR, and can barely detect single-copy duplications of less than around 100 kb (e.g. 19). Paired-end massively parallel sequencing [including variations often called mate-pair sequencing, in which both ends of circularized DNA fragments of 2–5 kb are sequenced (5)], can identify rearrangement junctions (e.g. 18), but cannot tell how they assemble into a larger-scale map, e.g. whether two junctions are on the same chromosome or not (Figures 1 and 6). HAPPY mapping can provide this information, provided the junctions are within about 1 Mb of each other, which will normally be a large enough scale to determine how genes are affected by the rearrangement, e.g. whether two junctions represent opposite ends of a small insertion (Figure 6). Paired-end sequencing has been used to provide longer-range mapping, by end-sequencing fosmid or BAC libraries made from a rearranged genome, e.g. to study structural aberrations and identify haplotypes in human

genomes (4,23), but library construction is demanding. For tumour samples, a further major limitation of current paired-end protocols, particularly mate-pair and library construction, is the need for several micrograms of input DNA to permit fragment size selection (24,25).

Could HAPPY mapping be used to scan the whole genome for rearrangements? Scaling up to genome-wide scans by PCR is probably impractical, because of the numbers of primers and PCR reactions involved—e.g. a scan for rearrangements >30 kb would need 100 000 markers at 30-kb intervals and 10^7 PCR reactions. The statistical significance of linkage changes would also be weakened by the many samples.

However, in future, it should be possible to implement HAPPY mapping using massively parallel sequencing instead of PCR markers, and this may permit genome-wide discovery and assembly of rearrangements. Instead of using PCR to score each diluted sample of a HAPPY panel for the presence or absence of chosen marker sequences, this can be done by global amplification of each diluted sample, followed by massively parallel sequencing. Exhaustive sequencing is not necessary (and, given the imperfections of global amplification, is not possible): the genome can be divided into segments of, for example, 5 kb, and the presence of only a few sequence reads from a given segment is then enough to confirm that segment's presence in a given dilution sample. In this way, relatively light sequencing of each dilution sample is enough to 'type' it for several hundred thousand 'markers' (DNA segments) throughout the genome. The cost of sequencing can be minimized by sequencing many dilution samples together, using bar-coding to distinguish the samples [i.e. adding a sample-specific linker sequence to all the DNA fragments in a given sample (26)].

This method of using sequencing instead of PCR markers has already been applied in the *de novo* mapping of the normal genome of *Hydra* (~1.3 Gb, two-thirds the size of the human genome), and worked well (Rokhsar,D., Chapman,J., David,C., Steele,R. and Dear,P.H., unpublished data).

This approach should enable genome-wide detection of complex rearrangements at a reasonable cost. Indeed, where second-generation sequencing is already envisaged for identifying junctions, copy-number changes or mutations, this type of mapping could be performed at little additional cost, simply by starting with a panel of suitable HAPPY dilution samples instead of a single large DNA sample.

CONCLUSION

HAPPY-mapping has previously been used in the assembly of whole genomes (15,27–29). In this paper, we have demonstrated its utility for identifying the physical relationship between segments of DNA in a rearranged cancer genome. HAPPY-mapping delivers high-resolution mapping information over a range of distances (1–200 kb) that can easily be controlled through marker design and fragment size selection. It will therefore be a useful complementary technique to genome-wide techniques, such as array-CGH and paired end sequencing, which discover rearrangements but do not provide long-range mapping information. In addition, DNA requirements are minimal and the equipment needed is available in most molecular biology laboratories.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Terry Rabbitts, now of the Leeds Institute of Molecular Medicine, who helped to start this collaboration through his involvement in molecular copy-number counting; Bee Ling Ng and Nigel Carter, Sanger Institute, for chromosome sorting; and Koichi Ichimura, V. Peter Collins and Department of Pathology microarray facility for DNA microarrays.

FUNDING

The Breast Cancer Campaign; Cancer Research UK; and the Medical Research Council. Funding for open access charge: Cancer Research UK.

Conflict of interest statement. A patent for molecular copy number counting has been granted to the UK Medical Research Council. P.H.D. is named on the patent..

REFERENCES

1. Hastings,P.J., Lupski,J.R., Rosenberg,S.M. and Ira,G. (2009) Mechanisms of change in gene copy number. *Nat. Rev. Genet.*, **10**, 551–564.
2. Edwards,P.A.W. (2010) Fusion genes and chromosome translocations in the common epithelial cancers. *J. Pathol.*, **220**, 244–254.
3. Mitelman,F., Johansson,B. and Mertens,F. (2007) The impact of translocations and gene fusions on cancer causation. *Nat. Rev. Cancer*, **7**, 233–245.
4. Volik,S., Zhao,S., Chin,K., Brebner,J.H., Herndon,D.R., Tao,Q., Kowbel,D., Huang,G., Lapuk,A., Kuo,W.L. *et al.* (2003) End-sequence profiling: sequence-based analysis of aberrant genomes. *Proc Natl Acad. Sci. USA*, **100**, 7696–701.
5. Korbel,J.O., Urban,A.E., Affourtit,J.P., Godwin,B., Grubert,F., Simons,J.F., Kim,P.M., Palejev,D., Carriero,N.J., Du,L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
6. Campbell,P.J., Stephens,P.J., Pleasance,E.D., O'Meara,S., Li,H., Santarius,T., Stebbings,L.A., Leroy,C., Edkins,S., Hardy,C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.
7. Dear,P.H. and Cook,P.R. (1993) HAPPY mapping: linkage mapping using a physical analogue of meiosis. *Nucleic Acids Res.*, **21**, 13–20.
8. Glöckner,G., Eichinger,L., Szafranski,K., Pachebat,J.A., Bankier,A.T., Dear,P.H., Lehmann,R., Baumgart,C., Parra,G., Abril,J.F. *et al.* (2002) Dictyostelium Genome Sequencing Consortium. Sequence and analysis of chromosome 2 of *Dictyostelium discoideum*. *Nature*, **418**, 79–85.
9. McCaughan,F. and Dear,P.H. (2010) Single-molecule genomics. *J. Pathol.*, **2**, 297–306.

10. Dear,P.H., Bankier,A.T. and Piper,M.B. (1998) A high-resolution metric HAPPY map of human chromosome 14. *Genomics*, **48**, 232–241.
11. Dear,P.H. (1997) HAPPY mapping. In Dear,P.H. (ed.), *Genome Mapping, A Practical Approach*. IRL Press, Oxford, UK, pp. 95–123.
12. Daser,A., Thangavelu,M., Pannell,R., Forster,A., Sparrow,L., Chung,G., Dear,P.H. and Rabbitts,T.H. (2006) Interrogation of genomes by molecular copy-number counting (MCC). *Nat. Methods*, **3**, 447–453.
13. Pole,J.C., Courtay-Cahen,C., Garcia,M.J., Blood,K.A., Cooke,S.L., Alsop,A.E., Tse,D.M., Caldas,C. and Edwards,P.A. (2006) High-resolution analysis of chromosome rearrangements on 8p in breast, colon and pancreatic cancer reveals a complex pattern of loss, gain and translocation. *Oncogene*, **25**, 5693–5706.
14. Piper,M.B., Bankier,A.T. and Dear,P.H. (1998) A HAPPY map of *Cryptosporidium parvum*. *Genome Res.*, **8**, 1299–1307.
15. Eichinger,L., Pachebat,J.A., Glöckner,G., Rajandream,M.A., Sucgang,R., Berriman,M., Song,J., Olsen,R., Szafranski,K., Xu,Q. *et al.* (2005) The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, **435**, 43–57.
16. Morris,J.S., Carter,N.P., Ferguson-Smith,M.A. and Edwards,P.A.W. (1997) Cytogenetic analysis of three breast carcinoma cell lines using reverse chromosome painting. *Genes Chromosomes Cancer*, **20**, 120–139.
17. Howarth,K.D., Blood,K.A., Ng,B.L., Beavis,J.C., Chua,Y., Cooke,S.L., Raby,S., Ichimura,K., Collins,V.P., Carter,N.P. *et al.* (2008) Array painting reveals a high frequency of balanced translocations in breast cancer cell lines that break in cancer-relevant genes. *Oncogene*, **27**, 3345–3359.
18. Stephens,P.J., McBride,D.J., Lin,M.L., Varela,I., Pleasance,E.D., Simpson,J.T., Stebbings,L.A., Leroy,C., Edkins,S., Mudie,L.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.
19. Howarth,K.D., Pole,J.C.M., Beavis,J.C., Batty,E.M., Newman,S., Bignell,G.R. and Edwards,P.A.W. (2011) Large duplications at reciprocal translocation breakpoints that might be the counterpart of large deletions and could arise from stalled replication bubbles. *Genome Res.*, **21**, 525–534.
20. Bignell,G.R., Santarius,T., Pole,J.C., Butler,A.P., Perry,J., Pleasance,E., Greenman,C., Menzies,A., Taylor,S., Edkins,S. *et al.* (2007) Architectures of somatic genomic rearrangement in human cancer amplicons at sequence-level resolution. *Genome Res.*, **17**, 1296–1303.
21. Alsop,A.E., Taylor,K., Zhang,J., Gabra,H., Paige,A.J. and Edwards,P.A.W. (2008) Homozygous deletions may be markers of nearby heterozygous mutations: the complex deletion at FRA16D in the HCT116 colon cancer cell line removes exons of WWOX. *Genes Chromosomes Cancer*, **47**, 437–447.
22. McCaughan,F., Darai-Ramqvist,E., Bankier,A.T., Konfortov,B.A., Foster,N., George,P.J., Rabbitts,T.H., Kost-Alimova,M., Rabbitts,P.H. and Dear,P.H. (2008) Microdissection molecular copy-number counting (microMCC)-unlocking cancer archives with digital PCR. *J. Pathol.*, **216**, 307–316.
23. Kidd,J.M., Cheng,Z., Graves,T., Fulton,B., Wilson,R.K. and Eichler,E.E. (2008) Haplotype sorting using human fosmid clone end-sequence pairs. *Genome Res.*, **18**, 2016–2023.
24. Feldman,A.L., Dogan,A., Smith,D.I., Law,M.E., Ansell,S.M., Johnson,S.H., Porcher,J.C., Ozsan,N., Wieben,E.D., Eckloff,B.W. *et al.* (2011) Discovery of recurrent t(6;7)(p25.3;q32.3) translocations in ALK-negative anaplastic large cell lymphomas by massively parallel genomic sequencing. *Blood*, **117**, 915–919.
25. Schatz,M.C., Delcher,A.L. and Salzberg,S.L. (2010) Assembly of large genomes using second-generation sequencing. *Genome Res.*, **20**, 1165–1173.
26. Meyer,M. and Kircher,M. (2010) Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.*, 2010, pdb.prot5448.
27. Hall,N., Pain,A., Berriman,M., Churcher,C., Harris,B., Harris,D., Mungall,K., Bowman,S., Atkin,R., Baker,S. *et al.* (2002) Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13. *Nature*, **419**, 527–531.
28. Abrahamsen,M.S., Templeton,T.J., Enomoto,S., Abrahante,J.E., Zhu,G., Lancto,C.A., Deng,M., Liu,C., Widmer,G., Tzipori,S. *et al.* (2004) Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*, **304**, 441–445.
29. Ling,K.H., Rajandream,M.A., Rivailler,P., Ivens,A., Yap,S.J., Madeira,A.M., Mungall,K., Billington,K., Yee,W.Y., Bankier,A.T. *et al.* (2007) Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. *Genome Res.*, **17**, 311–319.