Genome-wide analysis of DNA copy-number changes using cDNA microarrays

Jonathan R. Pollack[1], Charles M. Perou[2], Ash A. Alizadeh[3], Michael B. Eisen[2], Alexander Pergamenschikov[2], Cheryl F. Williams[2], Stefanie S. Jeffrey[4], David Botstein[2] & Patrick O. Brown[1,3]

Gene amplifications and deletions frequently contribute to tumorigenesis. Characterization of these DNA copy-number changes is important for both the basic understanding of cancer and its diagnosis. Comparative genomic hybridization (CGH) was developed to survey DNA copy-number variations across a whole genome[1]. With CGH, differentially labelled test and reference genomic DNAs are co-hybridized to normal metaphase chromosomes, and fluorescence ratios along the length of chromosomes provide a cytogenetic representation of DNA copy-number variation. CGH, however, has a limited (~20 Mb) mapping resolution, and higher-resolution techniques, such as fluorescence *in situ* hybridization (FISH), are prohibitively labour-intensive on a genomic scale. Array-based CGH, in which fluorescence ratios at arrayed DNA elements provide a locus-by-locus measure of DNA copy-number variation, represents another means of achieving increased mapping resolution[2–4]. Published array CGH methods have relied on large genomic clone (for example BAC) array targets and have covered only a small fraction of the human genome. cDNAs representing over 30,000 radiation-hybrid (RH)–mapped human genes[5,6] provide an alternative and readily available genomic resource for mapping DNA copy-number changes. Although cDNA microarrays have been used extensively to characterize variation in human gene expression[7–9], human genomic DNA is a far more complex mixture than the mRNA representation of human cells. Therefore, analysis of DNA copy-number variation using cDNA microarrays would require a sensitivity of detection an order of magnitude greater than has been routinely reported[7]. We describe here a cDNA microarray-based CGH method, and its application to DNA copy-number variation analysis in breast cancer cell lines and tumours. Using this assay, we were able to identify gene amplifications and deletions genome-wide and with high resolution, and compare alterations in DNA copy number and gene expression.

We first tested the feasibility of cDNA microarray-based CGH (Fig. 1*a*) by analysing genomic DNAs from tumour cell lines with known gene amplifications or deletions. BT474 is a human breast cancer cell line in which *ERBB2* is amplified[10]. We labelled genomic DNA from BT474 cells and genomic DNA from normal female human leukocytes with Cy5 (pseudocoloured red) and

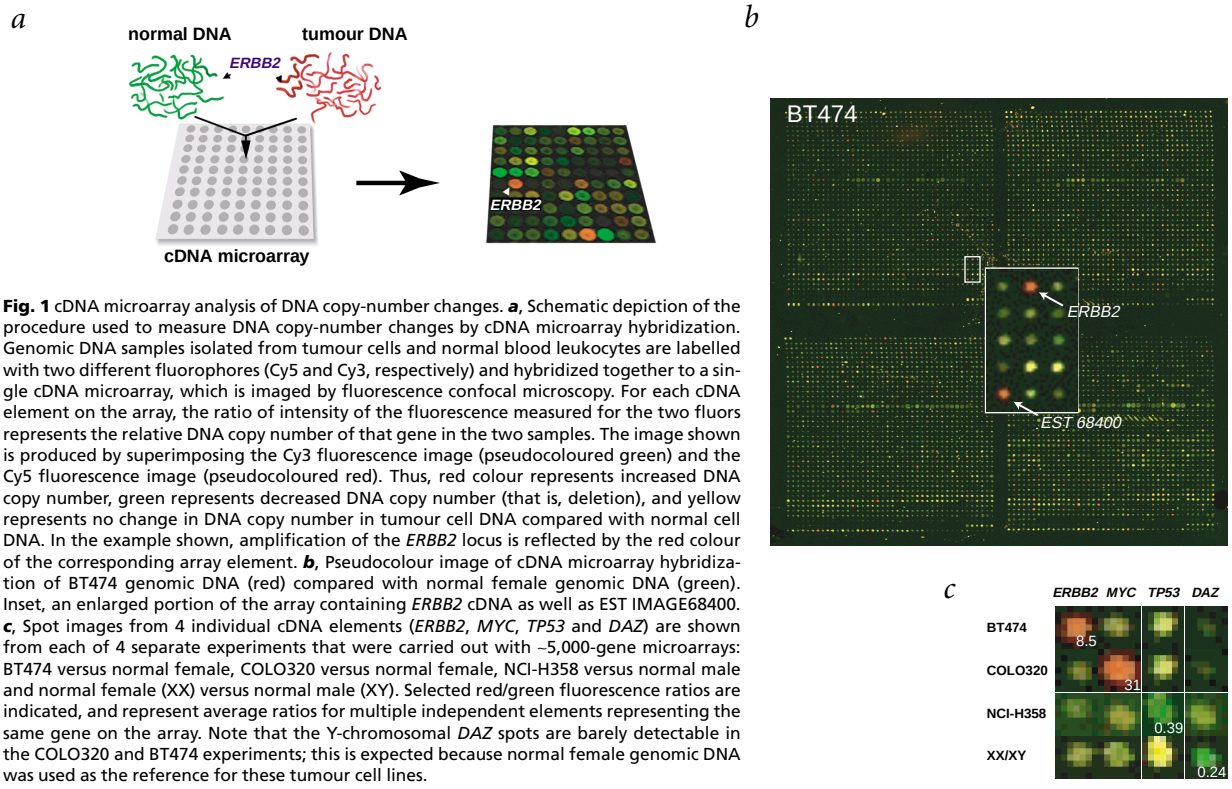

Fig. 1 cDNA microarray analysis of DNA copy-number changes. ***a***, Schematic depiction of the procedure used to measure DNA copy-number changes by cDNA microarray hybridization. Genomic DNA samples isolated from tumour cells and normal blood leukocytes are labelled with two different fluorophores (Cy5 and Cy3, respectively) and hybridized together to a single cDNA microarray, which is imaged by fluorescence confocal microscopy. For each cDNA element on the array, the ratio of intensity of the fluorescence measured for the two fluors represents the relative DNA copy number of that gene in the two samples. The image shown is produced by superimposing the Cy3 fluorescence image (pseudocoloured green) and the Cy5 fluorescence image (pseudocoloured red). Thus, red colour represents increased DNA copy number, green represents decreased DNA copy number (that is, deletion), and yellow represents no change in DNA copy number in tumour cell DNA compared with normal cell DNA. In the example shown, amplification of the *ERBB2* locus is reflected by the red colour of the corresponding array element. ***b***, Pseudocolour image of cDNA microarray hybridization of BT474 genomic DNA (red) compared with normal female genomic DNA (green). Inset, an enlarged portion of the array containing *ERBB2* cDNA as well as EST IMAGE68400. ***c***, Spot images from 4 individual cDNA elements (*ERBB2*, *MYC*, *TP53* and *DAZ*) are shown from each of 4 separate experiments that were carried out with ~5,000-gene microarrays: BT474 versus normal female, COLO320 versus normal female, NCI-H358 versus normal male and normal female (XX) versus normal male (XY). Selected red/green fluorescence ratios are indicated, and represent average ratios for multiple independent elements representing the same gene on the array. Note that the Y-chromosomal *DAZ* spots are barely detectable in the COLO320 and BT474 experiments; this is expected because normal female genomic DNA was used as the reference for these tumour cell lines.

[1]*Howard Hughes Medical Institute, Departments of* [2]*Genetics,* [3]*Biochemistry and* [4]*Surgery, Stanford Medical Center, Stanford, California 94305, USA. Correspondence should be addressed to P.O.B. (e-mail: pbrown@cmgm.stanford.edu).*

Table 1 • cDNA microarray analysis of model DNA copy-number variation

Test DNA	Reference DNA	Gene element	Mean fluorescence ratio (±1s.d.)[a] selected gene element[b]		Mean fluorescence ratio (±1s.d.)[a] all ~5,000 array elements	Test/reference DNA copy-number ratio selected gene element
BT474	normal female	*ERBB2*	8.5 (6.9–11)	n=4	1.0 (0.76–1.4)	10–15[c]
COLO320	normal female	*MYC*	31 (28–35)	n=4	1.0 (0.81–1.3)	30–50[d]
NCI-H358	normal male	*TP53*	0.39 (0.29–0.52)	n=4	1.0 (0.78–1.3)	0[e]
		TP53ΔAlu[f]	0.06 (0.05–0.08)	n=2		
normal female	normal male	*DAZ*[g]	0.24 (0.20–0.28)	n=2	1.0 (0.91–1.2)	0[h]
Turner (45,XO)	normal female	*F8C*[i]	0.54 (0.47–0.63)	n=3	1.0 (0.89–1.2)	0.5
		MCF2[i]	0.69 (0.64–0.74)	n=3		

[a]Fluorescence ratios are reported as test/reference. Selected gene element spot images appear in Fig. 1*c*. Means and standard deviations were calculated in log space. [b]Calculated from multiple (n) independent cDNA elements representing the same gene present on the array. [c]Fold-amplification estimated in the breast cancer cell line BT474 by Southern-blot analysis (ref. 11, and data not shown). [d]Fold-amplification estimated in the colon cancer cell line COLO320 by Southern-blot analysis (ref. 28, and data not shown). [e]*TP53* is homozygously deleted in the lung cancer cell line NCI-H358 (ref. 29). [f]Deletion of the *Alu* repeat in the 3′ UTR of the *TP53* target improved performance, consistent with incomplete blocking of repetitive sequences during hybridization. *Alu* repeats are present in the 3′ UTRs of approximately 5% of arrayed cDNAs. [g]Y-chromosomal gene. [h]The Y-chromosomal gene *DAZ* has an autosomal homologue (*DAZL*, which has ~90% nucleotide identity to *DAZ*, including the 3′ UTR; ref. 30), which will cause overestimation of the DNA copy-number ratio by hybridization. [i]X-chromosomal gene.

Cy3 (pseudocoloured green), respectively, and co-hybridized the labelled DNAs to a cDNA microarray containing approximately 5,000 human genes. Following hybridization, we scanned the microarray to produce a pseudocolour image (Fig. 1*b*). The average red/green fluorescence ratio of 4 independent cDNA elements representing *ERBB2* on the array was 8.5 (Fig. 1*c* and Table 1), closely approximating (but slightly underestimating) the 10–15:1 ratio determined by Southern-blot analysis (ref. 11, and data not shown). Similar analyses demonstrated our ability to detect high-level amplification of *MYC*, homozygous deletion of *TP53*, 'deletion' of the Y-chromosomal gene *DAZ* in a comparison of male and female genomic DNA, and single-copy 'deletion' of the X-chromosomal genes *F8C* and *MCF2* in a comparison of normal female with Turner syndrome genomic DNAs (Fig. 1*c* and Table 1).

To define the performance of the assay quantitatively, we hybridized genomic DNAs from cell lines containing varying numbers of X chromosomes to simulate varying levels of gene amplification and deletion for each of the 160 X-chromosomal genes present on the approximately 5,000-gene array. When we compared two samples of normal female genomic DNA (Fig. 2*a*), the red/green fluorescence ratios measured for both autosomal and X-chromosomal genes were tightly distributed around a mean value of 1. In contrast, when we compared genomic DNA

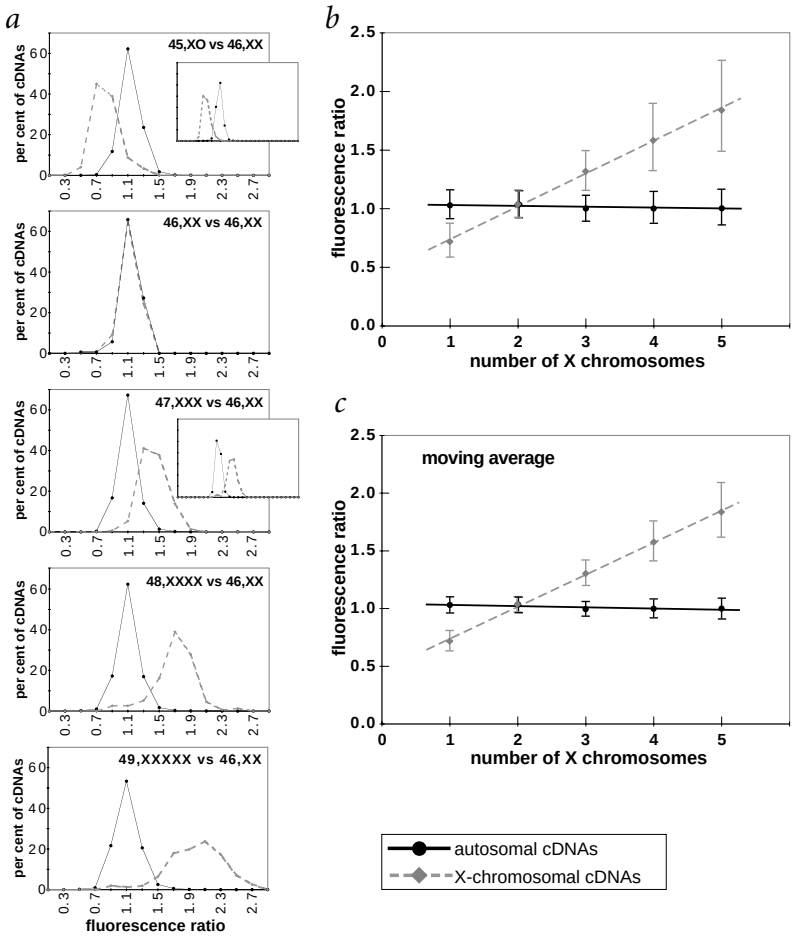

Fig. 2 Measuring X-chromosomal DNA copy-number variation. ***a***, Genomic DNA samples from 45,XO, 46,XX, 47,XXX, 48,XXXX and 49,XXXXX cell lines were separately labelled with Cy5 (red) and compared with 46,XX DNA labelled with Cy3 (green) using a microarray containing 3,920 autosomal cDNAs (representing 3,725 different genes) and 160 X-chromosomal cDNAs (~4%, representing 145 different genes); chromosomal assignments were determined from FISH or RH mapping databases. The graphs show the distribution of red/green fluorescence ratios for the autosomal cDNAs (solid line) and X-chromosomal cDNAs (dashed line), plotted as percentage of cDNAs on the ordinate versus red/green fluorescence ratio (binned by intervals of 0.2, upper boundary of bin indicated) on the abscissa. The leftward tailing in the distribution of X-chromosomal fluorescence ratios in the bottom panels is due in part to a small number of cDNAs that have been incorrectly assigned to UniGene clusters mapping to the X chromosome, or have significant homology to autosomal DNA sequences (data not shown). Insets in the first and third panels show the corresponding profiles determined using a moving average of fluorescence ratios, calculated for sets of three adjacent genes along the chromosome (as determined by RH map position). ***b***, Plot of mean (±1s.d.) fluorescence ratios of autosomal cDNAs (black circles) and X-chromosomal cDNAs (grey diamonds) from each experiment against number of X chromosomes. Mean (±1s.d.) fluorescence ratios of X-chromosomal cDNAs were as follows: XO versus XX, 0.72 (0.59–0.88); XX versus XX, 1.04 (0.93, 1.16); XXX versus XX, 1.31 (1.15–1.50); XXXX versus XX, 1.58 (1.32–1.90); XXXXX versus XX, 1.84 (1.49–2.27). Lines (solid line for autosomal and dashed line for X-chromosomal mean fluorescence ratios) were fitted using standard regression analysis. ***c***, Plot of mean (±1s.d.) moving average fluorescence ratios of autosomal cDNAs (black circles) and X-chromosomal cDNAs (grey diamonds) from each experiment against number of X chromosomes. Moving averages were calculated as described above.

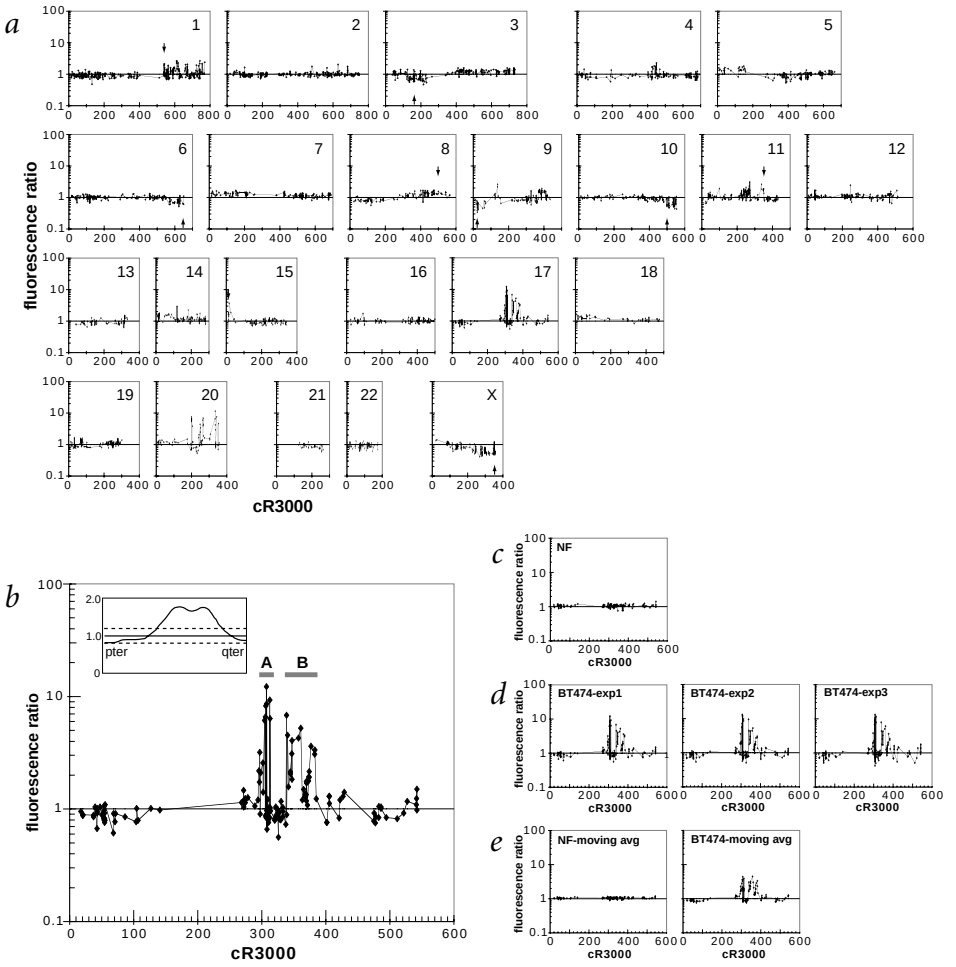

Fig. 3 Genome-wide mapping of DNA copy-number variation for breast cancer cell line BT474. ***a***, BT474 DNA copy-number profile across all chromosomes, derived from 3360 Genebridge4 RH-mapped cDNAs (representing 3,195 different genes) on the microarray depicted in Fig. 1*b*. For each chromosome, red/green fluorescence ratios of gene spots (ordinate, $\log_{10}$ scale) are plotted against chromosome map position as the cR3,000 position (abscissa) derived from radiation hybrid mapping using the Gene-bridge4 RH panel. Fluorescence ratios greater than one indicate DNA amplifications; ratios less than one indicate DNA deletion. Individual data points across a chromosome are connected by lines to facilitate viewing. Arrows indicate previously unreported sites of DNA amplification or deletion confirmed by quantitative PCR analysis of DNA copy number. Note, complete DNA copy-number profiles for normal female, Turner female (an example of chromosomal deletion) and breast cancer cell line MCF7 are available (http://genome-www.stanford.edu/aCGH/). ***b***, Enlarged view of chromosome 17 DNA copy-number profile for BT474, derived from 158 cDNAs (representing 147 different genes) RH-mapped (Gene-bridge4 RH panel) to chromosome 17. The two broad regions of amplification referred to in the text are labelled 'A' and 'B'. Inset, the same profile derived from CGH on metaphase chromosomes, redrawn from ref. 15. ***c***, Chromosome 17 DNA copy-number profile for normal female genomic DNA. ***d***, Chromosome 17 DNA copy-number profiles for BT474 determined from three independent experiments. ***e***, Moving average DNA copy-number profiles of chromosome 17 for normal female genomic DNA and BT474. Moving averages were calculated as described in Fig. 2. Complete moving average profiles for normal female genomic DNA and BT474 are available (http://genome-www.stanford.edu/aCGH/).

from a 45, XO (Turner syndrome) cell line (red) with normal female (46, XX) genomic DNA (green), the distribution of fluorescence ratios for X-chromosomal genes was shifted leftward (mean 0.72; Fig. 2*a*), reflecting the single-copy loss of X-chromosomal genes in the XO sample. Likewise, when we compared genomic DNAs from 47,XXX 48,XXXX and 49,XXXXX cell lines (red) with normal female (46, XX) genomic DNA (green), the distributions of fluorescence ratios for X-chromosomal genes were shifted rightward (means 1.31, 1.58 and 1.84, respectively; Fig. 2*a*), reflecting X-chromosomal DNA copy-number gain. The mean fluorescence ratios for X-chromosomal genes obtained in the different experiments fitted tightly to a line (Fig. 2*b*), with a regression correlation of 0.99, demonstrating that fluorescence ratios were linearly proportional to DNA copy number in this range of low-level gene amplification or single-copy deletion (in the case of XO versus XX). The slope of the line, 0.28, underestimated the true slope of 0.5, probably due primarily to cross-hybridization between some X-chromosomal genes and homologous sequences on autosomes.

Detection of single-copy deletions is important for the identification of tumour-suppressor genes. In the comparison of genomic DNA from an XO cell line with that from normal female cells, which models single-copy DNA deletion for X-chromosomal genes, we estimate that each individual array element provided approximately 85% sensitivity (15% false negatives) and approximately 85% specificity (15% false positives) for detection of single-copy gene 'deletion' (using a decision threshold at the point where the distributions of fluorescence ratios for autosomal and X-chromosomal genes crossed, at ~1 s.d. apart from their means). By using a 'moving average' of fluorescence ratios[4], it is possible to increase the accuracy of measurements with little sacrifice of mapping resolution (because of the large number of genes arrayed). A moving average analysis, calculated for sets of 3 adjacent genes along the chromosome (determined by RH map position), increased our estimates of sensitivity and specificity for detection of single-copy deletion or gain to approximately 98% (the distributions of fluorescence ratios for autosomal and X-chromosomal genes crossed at ~2 s.d. apart from their means; Fig. 2*a*,*c*). Note that with the moving average there would be a loss of sensitivity in detecting regions of amplification and deletion that are small relative to the local density of genes represented on the array. Of course, the decision thresholds and size of the moving average window could be adjusted to optimize the performance characteristics (for example, sensitivity, specificity and resolution) most desirable for any particular biological application of the assay.

The cDNA microarrays used in our experiments contained 3,360 cDNAs (representing 3,195 different genes) whose DNA sequences had been RH-mapped[5,6] using the Genebridge4 RH panel[12]. In the comparison of BT474 (red) genomic DNA and

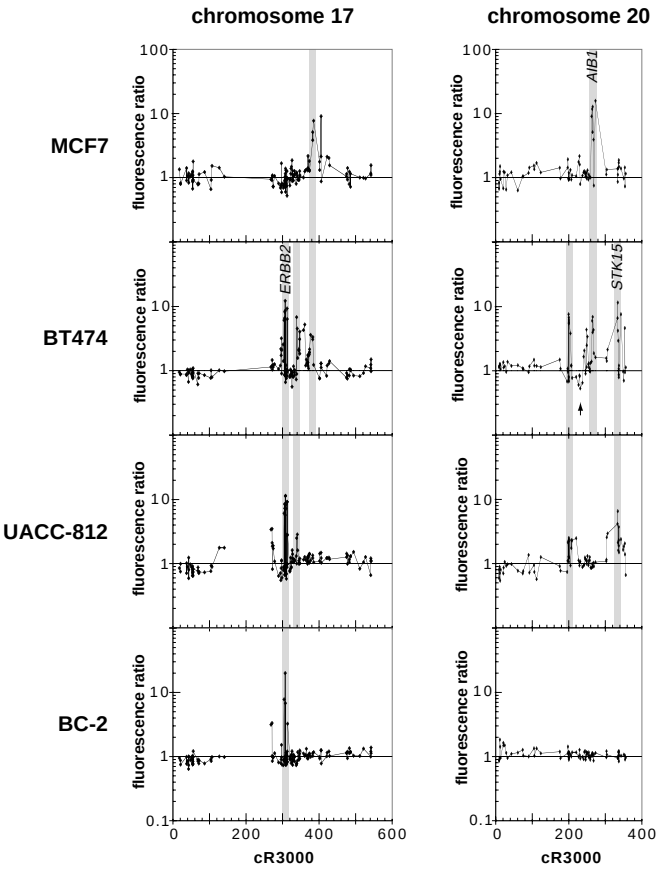

Fig. 4 High-resolution analysis of recurrent amplicons in breast cancer. DNA copy-number profiles for chromosomes 17 (derived from 158 cDNAs, representing 147 different genes) and chromosome 20 (derived from 87 cDNAs, representing 82 different genes) generated for breast cancer cell lines BT474, MCF7 and UACC-812, and primary breast tumour BC-2 are depicted. For each sample, red/green fluorescence ratios (ordinate, $\log_{10}$ scale) are plotted against the cR3,000 position derived from radiation hybrid mapping (abscissa). Individual data points across a chromosome are connected by lines to facilitate viewing. Recurrent regions of DNA amplification (with greater than fivefold amplification in at least one of the samples) are highlighted in grey. Selected genes within amplicons are identified. The arrow in the BT474 chromosome 20 profile indicates a region of deletion previously identified by BAC array CGH (ref. 3).

normal female (green) genomic DNA (Fig. 1*b*, pseudocolour image), we plotted fluorescence ratios for each RH-mapped element on the array according to their RH map location in the genome (Fig. 3*a*). Known amplifications[13] (using conventional metaphase CGH) along chromosomes 17q and 20q were apparent, as were many previously unreported regions of DNA amplification and deletion. For example, we observed amplifications along chromosomes 1q, 8q, 11q and 15q, and deletions along chromosomes 3p, 6q, 9p, 10q and Xq. These copy-number aberrations were confirmed by quantitative PCR (ref. 14) for a single gene in each region (Fig. 3*a*, arrows; data available at http://genome-www.stanford.edu/aCGH/).

An enlarged view of the fluorescence ratio plot for BT474 across chromosome 17 (Fig. 3*b*) revealed regions of amplification at approximately 17q12–q21 (*ERBB2* amplicon) and 17q22–q24, which correspond to the bimodal peak identified[15] by metaphase CGH analysis (Fig. 3*b*). The cDNA microarray measurements of DNA copy number were reproducible (Fig. 3*d*), and displayed both greater dynamic range and higher resolution than has been reported for conventional metaphase CGH. The maximum copy-number ratio measured by cDNA microarray hybridization was 12, compared with less than 2 for the published metaphase CGH analysis. Moreover, the two regions of amplification (labelled A and B in Fig. 3*b*) were more clearly resolved by cDNA microarray CGH, and the distal 17q amplicon (B) was itself further resolved into at least two discrete regions of amplification, revealing a complexity of amplicon structure in this region previously unappreciated by conventional CGH and FISH analyses[15]. With the microarrays used in these experiments, the mapping resolution was not limited by the number of genes arrayed, but by the effective resolution of the RH mapping panel, which at approximately 1 Mb (for the Genebridge4 RH panel; estimate from http://www.ncbi.nlm.nih.gov/genemap98/) represents an approximately 20-fold higher mapping resolution than attainable by metaphase CGH (ref. 16). The use of moving average ratios (Fig. 3*e*) decreased the 'noisiness' of the profile with little loss of mapping resolution (the shape of the profile was unaltered).

The identification of recurrent regions of DNA amplification in tumours has facilitated discovery of oncogenes. DNA copy-number profiles for chromosomes 17 and 20 are shown (Fig. 4), generated for three breast tumour cell lines (BT474, MCF7 and UACC-812) and one primary breast tumour (BC-2, a poorly differentiated infiltrating ductal carcinoma). We

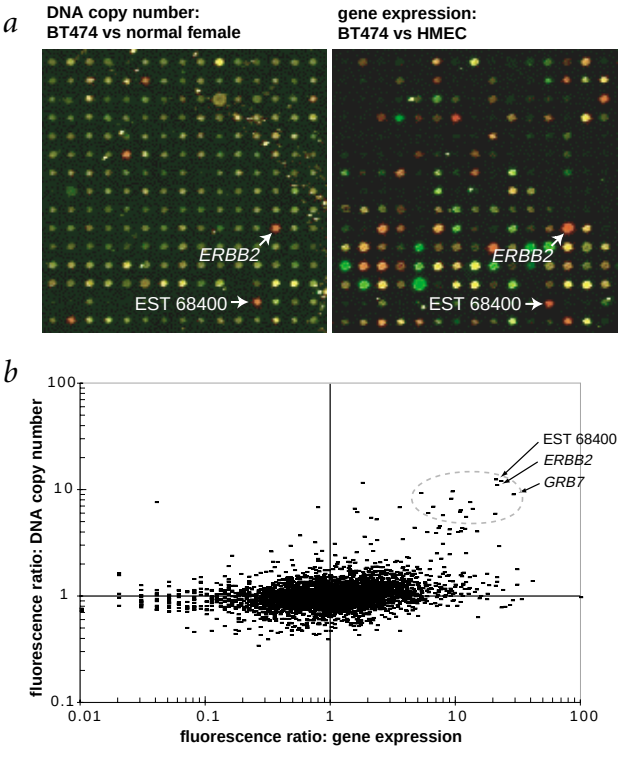

Fig. 5 Parallel analysis of DNA copy number and gene expression. ***a***, Identical portion of an ~5,000-gene microarray. Left, analysis of DNA copy-number variation in breast carcinoma cell line BT474 (red) compared with normal female genomic DNA (green). Right, the identical portion of the microarray with gene expression analysis of the same tumour sample, with BT474 poly(A)$^+$ mRNA (red) compared with normal human mammary epithelial cell (HMEC) poly(A)$^+$ mRNA (green). Selected genes are identified. ***b***, DNA copy number (ordinate, $\log_{10}$ scale) is plotted against gene expression (abscissa, $\log_{10}$ scale) for the ~4,000 mapped genes from the arrays depicted in Fig. 5*a*. The enclosed region identifies genes that are both highly amplified (>fivefold) and highly expressed (>fivefold compared with reference); selected genes are identified.

identified several recurrent regions of DNA amplification (Fig. 4). In addition to providing markers for localizing recurrent amplicons with high resolution, amplified genes represented on the array are themselves candidate oncogenes. For example, the recurrent amplicon at 17q12–q21 (present in BT474, UACC-812 and BC-2) corresponds to the well-studied *ERBB2* amplicon. In addition to *ERBB2*, we have detected high fluorescence ratios for named genes in this amplicon, including *GRB7* (ref. 17) and *MLN64*/*CAB1* (ref. 18), which may contribute to tumour progression[17,18], as well as previously uncharacterized ESTs (for example EST IMAGE 68400; Fig. 1*a*, inset). The region of amplification shared by BT474 and MCF at approximately 20q12 (Fig. 4) contains the candidate oncogene *AIB1* (refs 19,20), and the region of amplification shared by BT474 and UACC-812 (~20q13) contains the candidate oncogenes *TFAP2C* (ref. 21) and *STK15* (ref. 22). It is notable, however, that in both of these recurrent chromosome 20 amplicons the genes that have the highest levels of amplification (and therefore presumably the targets of greatest phenotypic selection) are not the previously recognized candidate oncogenes, but anonymous ESTs, suggesting that the relevant oncogenes in these regions are yet uncharacterized.

The cDNA microarrays used in this study provided an opportunity to analyse in parallel the changes in DNA copy number and expression levels of thousands of genes in the same tumour sample. We compared poly(A)$^+$ mRNA from each of the breast cancer cell lines and primary tumours (red) with poly(A)$^+$ mRNA prepared from normal human mammary epithelial cells (green), which, for the purpose of this experiment, served as an imperfect approximation of the normal counterpart of the tumour cells. Pseudocolour images from an identical portion of an array comparing DNA copy-number variation and gene expression for the BT474 breast cancer cell line are shown (Fig. 5*a*). The DNA copy number and gene expression data for the approximately 4,000 mapped cDNAs on these arrays are also shown (Fig. 5*b*). As expected, most highly expressed genes were not amplified, nor were all amplified genes highly expressed. A small number of genes were both amplified and highly expressed (Fig. 5*b*), and therefore are more likely to include important mediators of tumour formation or progression. Many of the genes in the *ERBB2* amplicon that were highly amplified in BT474 cells were also highly expressed (including *GRB7*, anonymous EST IMAGE68400 (Fig. 5*b*) and *MLN64*/*CAB1*), consistent with the possibility that genes in this region other than *ERBB2* may contribute to the tumour phenotype[17,18].

There is insufficient data in the literature to comprehensively compare the performance of our array CGH method with that of others[2–4]. When a moving-average ratio is used in our analysis of X-chromosomal copy-number changes, our ability to detect single-copy deletions/gains appears comparable to that reported[3] using larger BAC array targets. The DNA copy-number profiles derived from hybridization to cDNA microarrays appear comparable to those reported using BAC arrays. For example, the DNA copy-number profile for BT474 across chromosome 20 (Fig. 4), based on the 87 unselected chromosome 20 cDNAs present on our arrays, closely approximates that obtained[3] using 22 BAC targets selected to include loci known to be amplified. The concordance between the DNA copy-number profiles includes a region of DNA deletion (Fig. 4, arrow).

The use of cDNA microarrays for analysis of DNA copy-number variation offers some significant advantages over other array-based CGH methods[2–4], which have relied on array targets comprised of large genomic DNA clones (for example BACs, or BAC-derived inter-*Alu* PCR products). High-density cDNA microarrays containing 10,000 genes or more are routinely employed for gene expression analyses (ref. 9, and unpublished data), but no resource currently exists for full-genome coverage with large genomic clones. Ultimately, we would like not only to map DNA copy-number variation at high resolution, but also to measure changes in DNA copy number gene by gene, for every human gene. Our results suggest that, using presently available methods and cDNA resources, the cDNA microarrays can help us attain this goal. Another important advantage of cDNA microarray-based CGH is that DNA copy number and gene expression patterns can be characterized in parallel in the same sample. The ability to monitor gene amplification and expression in parallel and at high resolution may facilitate the identification of pathogenetically important genes in amplicons, and aid in the interpretation of the gene expression data being collected in studies of human tumours.

The detection of twofold and smaller differences in DNA copy number (twofold in XO versus XX, 1.5-fold in XXX versus XX) is notable. The cDNA elements on our microarrays averaged approximately 1 kb, or about 1.5×10^{-7} of the mass of the diploid human genome. Thus, for example, the difference in copy number of X-chromosomal genes in a comparison of genomic DNA from a Turner syndrome cell line with that from normal female cells represents a twofold difference in the partial concentration of a DNA sequence present at one part in six million in the complex DNA sample. Our ability to detect single-copy DNA deletions at the genome level should aid in the localization and identification of novel tumour-suppressor genes. Single-copy deletions (and gains) are also a characteristic feature of many constitutional genetic syndromes, and the cDNA microarrays may allow mapping and identification of genes whose copy number is altered in these syndromes. The ability to scan complex genomes for DNA copy-number variations using cDNA microarrays should have broad applications in cancer biology, human genetics, comparative genomics and other whole-genome studies in a variety of organisms.

Methods

Genomic DNA, mRNA and cell lines. We obtained BT474, COLO320HSR and NCI-H358 cell lines, and UACC-812 genomic DNA (American Type Culture Collection). We prepared genomic DNA from cell lines and peripheral blood using a Blood and Cell Culture DNA Maxi kit (Qiagen). We obtained genomic DNA samples from 45,XO (repository number NA01723A), 47,XXX (NA04626), 48,XXXX (NA040695) and 49,XXXXX (NA06061C) cell lines from the NIGMS Human Genetic Mutant cell repository. We obtained normal human mammary epithelial cells (Clonetics) and grew them according to the manufacturer's instructions. We snap froze primary breast tumours in liquid nitrogen within 20 min of devascularization and stored them at –80 °C. For all cell lines, we isolated poly(A)$^+$ mRNA using a FastTrack 2.0 kit (Invitrogen), whereas for primary breast tumours, we first isolated total RNA using Trizol reagent (Gibco BRL) followed by poly(A)$^+$ mRNA isolation as above.

cDNA microarrays. We fabricated cDNA microarrays essentially as described[23,24]. In brief, we PCR-amplified IMAGE (ref. 25) human cDNAs (ESTs) in 96-well format from DNA minipreps (Qiagen) using modified M13 universal primers. Most PCR products were 0.5–2 kb. We suspended purified PCR products in 3×SSC and robotically arrayed them (spotting ~1–5 ng each PCR product, 200 μm spacing between spots) onto polylysine-coated glass microscope slides[24]. We then processed the microarrays to immobilize the DNA (ref. 24). The cDNA microarrays described here contained 5,240 sequence-validated human cDNAs, representing 4,915 different human genes (UniGene clusters[5,26,27]), including 5,184 cDNAs of the Research Genetics GeneFilters Release I. Approximately one-half the cDNAs on the microarray were either known genes or similar to known genes in other organisms, whereas the remainder were anonymous ESTs.

Labelling and hybridizations. For each labelling, we *Dpn*II-digested (New England Biolabs) genomic DNA (2 μg), which was then purified (Qiaquick PCR kit) and random-primer labelled using a Bioprime Labeling kit (Gibco BRL), modified to include in a 100 μl reaction, dATP, dGTP and dTTP (120 μM each), dCTP (60 μM) and Cy5-dCTP or Cy3-dCTP (60 μM). (We subsequently found that labelling 2 μg of genomic DNA in a 50 μl reaction volume performed equivalently, and with less reagent cost.) We purified labelled products using a microcon 30 filter (Amicon). We then combined Cy5- and Cy3-labelled probes from the entire labelling reactions along with human Cot-1 DNA (30–50 μg; Gibco BRL), yeast tRNA (100 μg; Gibco BRL) and poly(dA-dT) (20 μg; Sigma). We concentrated the hybridization mixture using a microcon 30 filter (Amicon) and adjusted it to contain 3.4×SSC and 0.3% SDS in a 15 μl final volume. Following denaturation (100 °C, 1.5 min) and a 30 min Cot-1 preannealing step (37 °C), we hybridized the probe to the array under a glass coverslip at 65 °C for 16–20 h. Following hybridization, we washed the arrays in 2×SSC, 0.03% SDS (65 °C, 5 min), followed by 5 min each at RT in 1×SSC and 0.2×SSC. We labelled poly(A)$^+$ mRNA for gene expression analysis as described[24].

Optimization. During optimization of the cDNA microarray CGH procedure, we found that the labelling efficiency was increased by reducing the average fragment size of the genomic DNA before random-primed labelling. This may reflect the increased accessibility of the DNA template following digestion. We also found it important to add to the hybridization mixture not only human Cot-1 DNA (to block hybridization to repetitive elements contained on ~3% of cDNAs) but also poly(dA-dT) to block non-specific hybridization to extended poly(A) tails present on some of the cDNA clones. A small number (~0.2%) of cDNAs on the array appeared amplified in most if not all tumour samples tested. Further characterization has shown these putative cDNAs to be derived from the mitochondrial genome, consistent with previous observations that mitochondria are typically more abundant in tumour cells than in their normal counterparts. The mitochondrial genome-derived clones were identified by hybridization with purified mitochondrial DNA, allowing them to be annotated appropriately (data not shown).

Imaging and data analysis. We scanned hybridized arrays by fluorescence confocal microscopy as described[23,24], collecting measurements for each fluor separately. We calculated fluorescence ratios after background subtraction (we calculated background as the median fluorescence signal of non-target pixels) using the ScanAlyze software package (M.B.E., D.B. and P.O.B., unpublished data, available at http://rana.stanford.edu/software). To correct for differences in DNA labelling efficiency between samples, we then normalized fluorescence ratios across the array to achieve an average log ratio of 0 (average ratio of 1, that is no DNA copy-number change) for all cDNA elements on the array. We have corrected ratio values and pseudocolour array images presented to reflect these normalized ratio values. We excluded array spots with insufficient fluorescence signal in the normal genomic DNA sample (fluorescence signal <20% above background, reflecting PCR or printing failures, <1% of spots), and array spots with overlying fluorescent debris (manually flagged, <0.5% of spots) from data analysis. We calculated means and standard deviations of fluorescence ratios in log space to weight DNA amplifications (fluorescence ratios >1) and deletions (fluorescence ratios <1) equally. We interpreted DNA copy-number profiles that deviated significantly (>1s.d.) from background ratio variations seen in normal genomic DNA samples as evidence of real copy-number differences[4,13]. When indicated, we used a 'moving average' of fluorescence ratios, calculated for sets of three adjacent genes along the chromosome, as determined by RH map position. A moving average ratio served to average across multiple elements any imprecision in measurement along with inaccuracies due to uncommon RH-mapping/UniGene misassignments. We obtained RH map positions of cDNAs from the GeneMap'98 RH mapping database[12] (database accessible at http://www.ncbi.nlm.nih.gov/genemap98/), using the UniGene clustering scheme[5,26,27] to assign map positions to cDNA elements on the array.

Acknowledgements

We thank K. Ranade for assistance with quantitative PCR analysis; R. Sutton, C. Rees and members of the Brown and Botstein Labs for helpful discussions; and J. Doda for human mitochondrial DNA. This work was supported by grants from the National Cancer Institute, the National Human Genome Research Institute and the Howard Hughes Medical Institute. J.R.P. is a Physician Postdoctoral Fellow and P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute. C.M.P. is a SmithKline Beecham Pharmaceuticals Fellow of the Life Sciences Research Foundation. M.B.E. is supported by a postdoctoral fellowship from the Alfred E. Sloan Foundation.

Received 24 February; accepted 26 July 1999.

1. Kallioniemi, A. *et al*. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821 (1992).
2. Solinas-Toldo, S. *et al*. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**, 399–407 (1997).
3. Pinkel, D. *et al*. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet*. **20**, 207–211 (1998).
4. Geschwind, D.H. *et al*. Klinefelter's syndrome as a model of anomalous cerebral laterality: testing gene dosage in the X chromosome pseudoautosomal region using a DNA microarray. *Dev. Genet*. **23**, 215–229 (1998).
5. Schuler, G.D. *et al*. A gene map of the human genome. *Science* **274**, 540–546 (1996).
6. Deloukas, P. *et al*. A physical map of 30,000 human genes. *Science* **282**, 744–746 (1998).
7. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
8. DeRisi, J. *et al*. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet*. **14**, 457–460 (1996).
9. Iyer, V. *et al*. The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
10. Kallioniemi, O.P. *et al*. ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. *Proc. Natl Acad. Sci. USA* **89**, 5321–5325 (1992).
11. Lucito, R. *et al*. Genetic analysis using genomic representations. *Proc. Natl Acad. Sci. USA* **95**, 4487–4492 (1998).
12. Gyapay, G. *et al*. A radiation hybrid map of the human genome. *Hum. Mol. Genet*. **5**, 339–346 (1996).
13. Kallioniemi A. *et al*. Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. *Proc. Natl Acad. Sci. USA* **91**, 2156–2160 (1994).
14. Gelmini, S. *et al*. Quantitative polymerase chain reaction-based homogeneous assay with fluorogenic probes to measure c-erbB-2 oncogene amplification. *Clin. Chem*. **43**, 752–758 (1997).
15. Barlund, M. *et al*. Increased copy number at 17q22–q24 by CGH in breast cancer is due to high-level amplification of two separate regions. *Genes Chromosomes Cancer* **20**, 372–376 (1997).
16. Kallioniemi, O.P. *et al*. Optimizing comparative genomic hybridization for analysis of DNA sequence copy number changes in solid tumors. *Genes Chromosomes Cancer* **10**, 231–243 (1994).
17. Stein, D. *et al*. The SH2 domain protein GRB-7 is co-amplified, overexpressed and in a tight complex with HER2 in breast cancer. *EMBO J*. **13**, 1331–1340 (1994).
18. Tomasetto C. *et al*. Identification of four novel human genes amplified and overexpressed in breast carcinoma and localized to the q11–q21.3 region of chromosome 17. *Genomics* **28**, 367–376 (1995).
19. Guan, X.Y. *et al*. Hybrid selection of transcribed sequences from microdissected DNA: isolation of genes within amplified region at 20q11–q13.2 in breast cancer. *Cancer Res*. **56**, 3446–3450 (1996).
20. Anzick, S.L. *et al*. AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. *Science* **277**, 965–968 (1997).
21. Williamson, J.A. *et al*. Chromosomal mapping of the human and mouse homologues of two new members of the AP-2 family of transcription factors. *Genomics* **35**, 262–264 (1996).
22. Sen, S., Zhou, H. & White, R.A. A putative serine/threonine kinase encoding gene BTAK on chromosome 20q13 is amplified and overexpressed in human breast cancer cell lines. *Oncogene* **14**, 2195–2200 (1997).
23. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*. **6**, 639–645 (1996).
24. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
25. Lennon, G., Auffray, C., Polymeropoulos, M. & Soares, M.B. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33**, 151–152 (1996).
26. Boguski, M.S. & Schuler, G.D. ESTablishing a human transcript map. *Nature Genet*. **10**, 369–371 (1995).
27. Schuler, G.D. Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med*. **75**, 694–698 (1997).
28. Alitalo, K., Schwab, M., Lin, C.C., Varmus, H.E. & Bishop, J.M. Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. *Proc. Natl Acad. Sci. USA* **80**, 1707–1711 (1983).
29. Takahashi, T. *et al*. p53: a frequent target for genetic abnormalities in lung cancer. *Science* **246**, 491–494 (1989).
30. Saxena, R. *et al*. The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nature Genet*. **14**, 292–299 (1996).