nature genetics

Array comparative genomic hybridization and its applications in cancer

Daniel Pinkel & Donna G Albertson

Alteration in DNA copy number is one of the many ways in which gene expression and function may be modified. Some variations are found among normal individuals, others occur in the course of normal processes in some species and still others participate in causing various disease states. For example, many defects in human development are due to gains and losses of chromosomes and chromosomal segments that occur before or shortly after fertilization, and DNA dosage-alteration changes occurring in somatic cells are frequent contributors to cancer. Detecting these aberrations and interpreting them in the context of broader knowledge facilitates the identification of crucial genes and pathways involved in biological processes and disease. Over the past several years, array comparative genomic hybridization has proven its value for analyzing DNA copy-number variations. Here, we discuss the state of the art of array comparative genomic hybridization and its applications in cancer, emphasizing general concepts rather than specific results.

Comparative genomic hybridization

Comparative genomic hybridization (CGH)[1,2] was the first efficient approach to scanning the entire genome for variations in DNA copy number (**Fig. 1a**). In a typical CGH measurement, total genomic DNA is isolated from test and reference cell populations, differentially labeled and hybridized to metaphase chromosomes or, more recently, DNA microarrays. The relative hybridization intensity of the test and reference signals at a given location is then (ideally) proportional to the relative copy number of those sequences in the test and reference genomes. If the reference genome is normal, then increases and decreases in the intensity ratio directly indicate DNA copy-number variation in the genome of the test cells. More than two genomes can be compared simultaneously if distinguishable labels are available (ref. 3 and A. Estep, unpublished data; **Fig. 1b**). Data are typically normalized so that the modal ratio for the genome is set to some standard value, typically 1.0 on a linear scale or 0.0 on a logarithmic scale. Additional measurements

Daniel Pinkel and Donna G. Albertson are in the Department of Laboratory Medicine and Comprehensive Cancer Center, University of California San Francisco, Box 0808, San Francisco, California 94143, USA. Donna G. Albertson is also in the Cancer Research Institute, University of California San Francisco, San Francisco, California, USA. e-mail: pinkel@cc.ucsf.edu or albertson@cc.ucsf.edu

such as FISH or flow cytometry[4] can be used to determine the copy number associated with a given ratio level.

Array CGH has been implemented using a wide variety of techniques. The initial approaches used arrays with elements produced by spotting DNA obtained directly from large-insert genomic clones such as BACs[5,6]. Because producing sufficient BAC DNA of adequate purity to make arrays is arduous, a number of techniques for amplifying small amounts of starting material have been used. These techniques include ligation-mediated PCR[7], degenerate primer PCR using one[8] or several[9] sets of primers and rolling circle amplification[10]. BAC arrays providing complete genome-tiling paths of mammalian genomes are now being produced[11,12]. Arrays made from less complex nucleic acids such as cDNAs[13], selected PCR products[14,15] and oligonucleotides[16,17] are also being used for array CGH. Although most CGH procedures use hybridization with total genomic DNA, some use reduced-complexity representations of the genome produced by PCR techniques. Computational analysis of the genome sequence is used to design array elements complementary to the sequences contained in the representation[18]. Currently, various SNP genotyping platforms are being evaluated for their ability to determine both DNA copy number and allelic content across the genome, some of which use reduced-complexity genomic representations[19,20].

The different basic approaches to array CGH provide different levels of performance, such that some are more suitable for particular applications than others. Factors that determine performance requirements include the magnitudes of the copy-number changes, their genomic extents, the state and composition of the specimen, how much material is available for analysis and how the results of the analysis will be used (**Fig. 2**). Many applications require reliable detection of copy-number changes of much less than 50%, a much more stringent requirement than for other microarray technologies. Technical details are extremely important, such that different implementations of the same array CGH approach may yield different levels of performance. Here, we discuss some of the issues related to array CGH performance, including biophysical, genomic and sample preparation considerations, and conclude with a general discussion of applications to cancer. Array CGH is also used extensively for analysis of constitutional abnormalities, but that topic is beyond our scope. Information on these applications is contained in a related review[21].

Technical considerations in array CGH

Hybridization signals. The major technical challenge of array CGH is the generation of hybridization signals that are sufficiently intense, specific and quantitative that copy-number changes can be detected. The
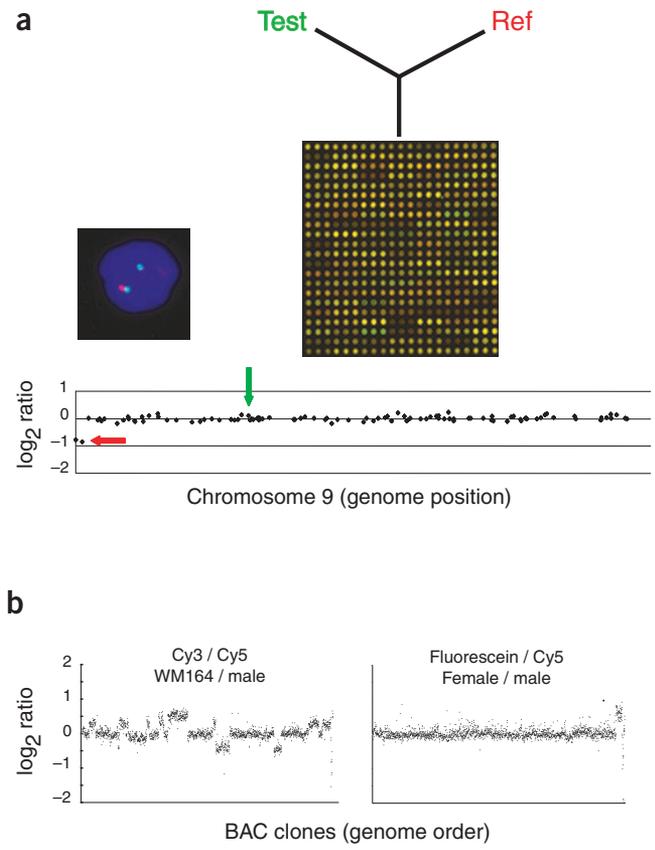
a

Test   Ref

Figure 1 Overview of array CGH. (**a**) Genomic DNA from two cell populations is differentially labeled and hybridized to a microarray. The fluorescent ratios on each array spot are calculated and normalized so that the median $\log_2$ ratio is 0. Plotting of the data for chromosome 9 from pter to qter shows that most elements have a ratio near 0. The two elements nearest pter have a ratio near −1, indicating a reduction by a factor of two in copy number. FISH with a red-labeled probe for the deleted region and a green-labeled control probe (genome locations indicated by the red and green arrows on the ratio profile) shows that the cells contain two copies bound by the green probe and only one bound by the red, consistent with the array CGH analysis[7]. (**b**) Simultaneous comparison of three genomes. Cy3-labeled genomic DNA from melanoma cell line WM-164, Cy5-labeled normal male genomic DNA and fluorescein-labeled normal female genomic DNA were simultaneously hybridized to a BAC array and imaged using a custom-build CCD imaging system (D.P., D.G.A. *et al.*, unpublished data). The left panel shows the Cy3/Cy5 ratio (cell line/normal male) for each clone plotted in genome order. A large number of aberrations are evident in this nondiploid cell line. The right panel shows the fluorescein/Cy5 ratio (normal female/normal male) with the X chromosome copy-number difference evident. Multigenome hybridizations of this type permit more efficient use of arrays and the inclusion of a control with each unknown specimen. Unpublished data courtesy of A. Estep (University of California San Francisco, California, USA).

b

signal intensity on an array element is affected by a number of factors, including base composition, proportion of repetitive sequence content and amount of 'hybridizable' DNA in the array element. Intensities may vary by a factor of 30 or more among array elements even if there are no copy-number changes. If the entire hybridization and measurement process is well behaved (*i.e.*, the signals are linearly proportional to sequence abundance), then the comparative hybridization strategy shown in **Figure 1a** provides ratios that are quantitatively proportional to relative copy number. Most importantly, production variability among different arrays, such as the amount of DNA in array elements or element morphology, is accurately compensated. Ratio accuracy is maintained even if the intensities become nonlinearly related to

genomic abundance owing to processes that affect the test and reference genomes equally, such as saturation of array elements or reassociation of double-stranded nucleic acids during hybridization. The alternative strategy, hybridization of a single genome to an array and comparison of the result to a set of historical controls, places more stringent requirements on reproducibility of array manufacture and hybridization conditions to avoid reduction in data quality.

The complexities of both the genomic DNA and the DNA in the array element affect the signal intensities and therefore have a dominant role in determining the data quality obtained from different array CGH technologies. For example, it is much easier to obtain copy-number information from genomes such as those of bacteria and yeast[22,23] than from mammalian genomes (which are 100–1,000 times more complex) because the concentration of each portion of the genome in the hybridization is correspondingly higher. Similarly, owing to a number of complex kinetic factors, array elements made from genomic BAC clones (complexity ~100–200 kb) typically provide more intense signals than do elements with shorter sequences such as cDNAs, PCR products and oligonucleotides. The more intense signals from higher-complexity array elements result in better measurement precision, allowing detection of single-copy transition

Figure 2 Factors influencing the success of array CGH. The difficulty of array CGH analysis varies among different applications. It is much easier to detect large increases in copy number associated with amplification of a genomic region than single-copy gains or losses. Aberrations affecting an extended genomic region spanning multiple array elements are easier to detect than focal events. Measurements on cell lines are the least difficult because isolation of high-quality DNA is straightforward and the genomes are relatively homogeneous. Fresh or frozen tumor tissues present additional challenges owing to possible tissue-specific factors and the potential for genomic heterogeneity in a tumor or inclusion of normal cells. Measurements on formalin-fixed, paraffin-embedded tissue present the greatest challenges. Research studies aimed at profiling a group of tumor specimens that have a large number of highly recurrent aberrations can be informative even if a considerable number of errors are made in the analysis of each tumor. In contrast, the detection of small or rare aberrations and clinical applications present challenging specificity and sensitivity requirements.

Increasing difficulty

|			
Aberration:	Amplification		Single-copy change
Aberration size:	Multiarray element		Single array element
Specimen:	Cell line	Fresh/frozen tissue	Fixed archival tissue
Composition:	Homogeneous		Heterogeneous (normal cells)
Material:	Lots		Small primary tumor
Data utility:	Population overview		Accuracy for each specimen

Figure 3 Array CGH analysis of a deletion boundary using arrays with elements of different complexity. Data from a BAC array (element complexity ~150–200 kb; black bars) indicate the decrease in ratio due to the deletion. Data from a fosmid array (element complexity 30–40 kb; green bars) provide a more precise indication of the deletion boundary, but the variation in ratios among the different array elements is increased. Finally, data from an array of genomic PCR products (element complexity 1.5–4 kb; red bars) provide even higher-resolution information on the boundary, but with a larger ratio variation. Noise from low-complexity array elements decreases in importance as the magnitude of the copy-number change increases, so that boundaries of amplified regions, if they are abrupt, can be determined even if the measurements are very noisy. The data indicate that one of the BAC clones is partially contained in the deletion, which might underlie the slightly reduced ratio seen for this clone. Thus, a tiling path of BAC clones can map the position of the copy-number transition to a fraction of the length of a clone[24]. Unpublished data courtesy of R. Redon (Wellcome Trust Sanger Institute, UK).

boundaries even in specimens with a high proportion of normal cells and localization of copy-number transitions to a fraction of the length of the array element in some circumstances[24].

Smaller array elements can potentially provide higher genomic resolution if measurement precision can be maintained or if the application focuses on the detection of multicopy aberrations such as gene amplifications. The advantages of using shorter sequences, including the opportunity to design arrays directly from genome sequence, the ability to use the same arrays for expression and genomic analysis and the possibility of higher genomic resolution, drive efforts to improve the performance of arrays with low-complexity elements. Detection of single-copy changes on individual array elements has been demonstrated for sequences as short as several kilobases[14], and even several hundred bases[15], but combining data from multiple elements is currently required for oligonucleotide arrays[17,25,26]. **Figure 3** illustrates the relationship between measurement precision and genomic resolution for analysis of a single-copy deletion boundary using arrays made from BACs, fosmids and PCR products of several kilobases in length. As indicated above, some measurement approaches reduce the complexity of the genomic DNA to increase signal intensities and allow the use of low-complexity array elements[19,20,25]. Published data from these procedures indicate that the noise levels are too high to allow detection of single-copy changes affecting individual array elements.

Genome characteristics and copy-number measurement. The change in ratio produced by a copy-number change is affected by several intrinsic characteristics of the specimen DNA. Most important are the high-copy repetitive sequences dispersed throughout mammalian genomes. These can hybridize to array elements that contain copies of the repeats, such as those made from genomic and cDNA clones, overwhelming the signal from the unique sequences. Therefore, hybridization from these sequences must be blocked, typically by adding unlabeled Cot-1 DNA to the hybridization, or the repetitive sequences must be removed from the genomic DNA or be absent from the array elements. Blocking is not perfectly effective, and signals are biased by the residual repetitive sequence hybridization. In addition, there may be general nonspecific binding of test and reference signals to the array elements.

A simple model for the effect of biases that equally affect the test and reference signals, such as that from the repetitive sequences, shows that the change in ratio produced by a copy-number change remains linear, but the slope is reduced as the bias increases because the signal does not become zero even if all copies of a locus are absent (**Fig. 4a**)[4]. Linear-ratio increases of more than several orders of magnitude have been demonstrated in some array CGH systems[6]. **Figure 4b** shows data from a cell line
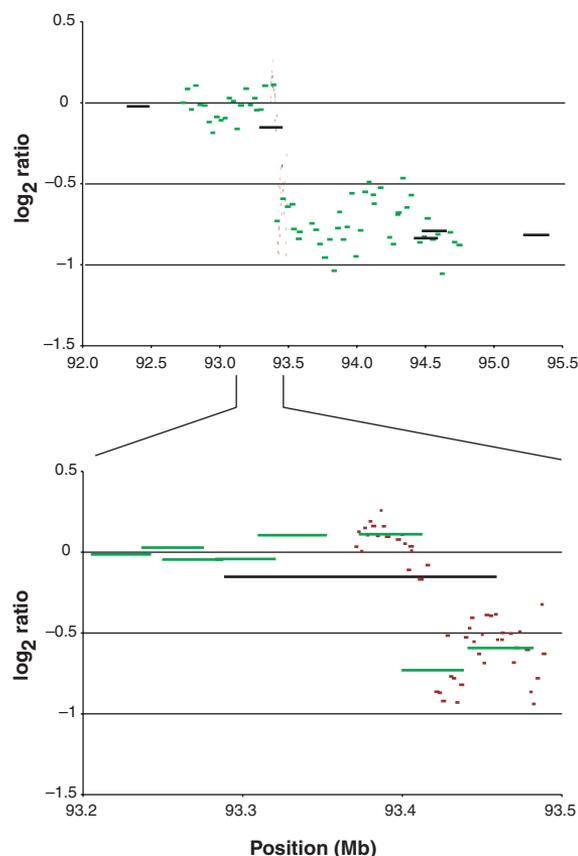
that is reasonably homogeneous in its genomic constitution and has true copy-number levels ranging from 0 to 3 as determined by FISH. The measured ratios for this sample, and published data[27], show that this simple model provides an accurate description of the behavior of some array CGH systems. If the magnitude of the biases differs significantly among array elements (*e.g.*, owing to different repetitive sequence content), then the elements will reproducibly follow different curves in **Figure 4a**. Such behavior may lead to false indications of recurrent copy-number structure in a region where the aberrant copy number is constant, producing false indications of the potential locations of crucial genes.

The performance of an array system for measurement of heterogeneous specimens (*e.g.*, normal cells in tumor specimens) can be estimated by first establishing its behavior with a well-characterized homogeneous specimen. The expected ratio changes in the heterogeneous specimen can then be obtained using the measured response curve in conjunction with values of the normalized copy number appropriate for the expected aberrations in the specimens. For example, a single-copy deletion in diploid tumor cells that were mixed with an equal number of normal cells would result in a normalized copy number of 0.75 instead of 0.5, the value expected if the specimen were pure tumor (**Fig. 4a**). Comparing the expected ratio changes with the noise level characteristic of the measurement platform then allows determination of the ranges of copy-number change and specimen heterogeneity for which acceptable performance might be expected.

Finally, this simple model does not describe the behavior of the measurements if the effective biases on array elements have contributions from autofluorescence, differential nonspecific behavior of the genomic DNAs owing to differences in the labels or high levels of nonspecific binding to the array substrate, or if the measurement process has artifacts introduced by nonlinearities in the imaging systems or characteristics of

Figure 4 Relationship of measured ratios to copy-number change. (**a**) Calculated ratios (linear representation, left; logarithmic, right) as a function of copy number using a simple model that includes the signal from unsuppressed repetitive sequences and nonspecific hybridization. The ratios are plotted relative to the normalized copy number, which is set to 1.0 for the median copy number in the genome. The heavy line shows the dependence when the signal is entirely due to sequences uniquely associated with the locus corresponding to the array element. The five lighter lines show the dependence when test and reference signals on the array element include a bias equal to 10%, 20%, 30%, 40% or 50% of the signal that would be present when the normalized copy number of the locus is 1. The circles indicate the ratios corresponding to true copy number of 0, 1, 2 and 3 found in the profile in **b**. (The model assumes that the bias, β, is proportional to the total amount of genomic DNA used in the hybridization but independent of the copy number of a particular locus because it is generated by sequences distributed throughout the genome. Because the unique sequence signal on an array element is also proportional to the amount of genomic DNA, after normalization, one can write that the test signal is $C+\beta$, where C is the copy number of the locus normalized to the median, or any other similar value, for the genome, whereas the reference signal is $1+\beta$. Thus, ratio = $(C+\beta)/(1+\beta)$.

Lines show behavior for β=0, 0.11, 0.25, 0.43, 0.67 and 1.0. (**b**) Ratio profile of a variant of cell line HCT-116 undergoing selection for resistance to methotrexate[54]. Array CGH was done using the BAC arrays, and copy-number levels for the parental HCT-116 cells were previously determined using FISH[7]. The ratios were directly calculated from the total test and total reference signal intensities for each element, corrected for local background. An overall normalization factor was applied to set the median $\log_2$ ratio = 0. No other computational adjustments were used. The cell line contains a well-established homozygous deletion on chromosome 16p ($\log_2$ ratio $\approx$ –3.2 in this analysis) as well as single-copy deletions ($\log_2$ ratio $\approx$ –0.8) and single-copy gains ($\log_2$ ratio $\approx$ 0.5). Plotting these points in **a** demonstrates that, in this data set, the typical bias on the array elements was equal to ~10% of the diploid signal level, and the response slopes for all array elements were very similar. Individual clones with ratios much different from 0 indicate copy-number polymorphisms, focal aberrations or noise. Close examination of the ratios indicates that some genomic regions are heterogeneous in copy number in this population, presumably owing to the ongoing selection. In particular, the ratio on chromosome 5q, the site of *DHFR* (the target of methotrexate), is slightly higher than on other regions of the HCT116 genome that are characteristically present at three copies. Unpublished data courtesy of A. Snijders (University of California San Francisco, California, USA).
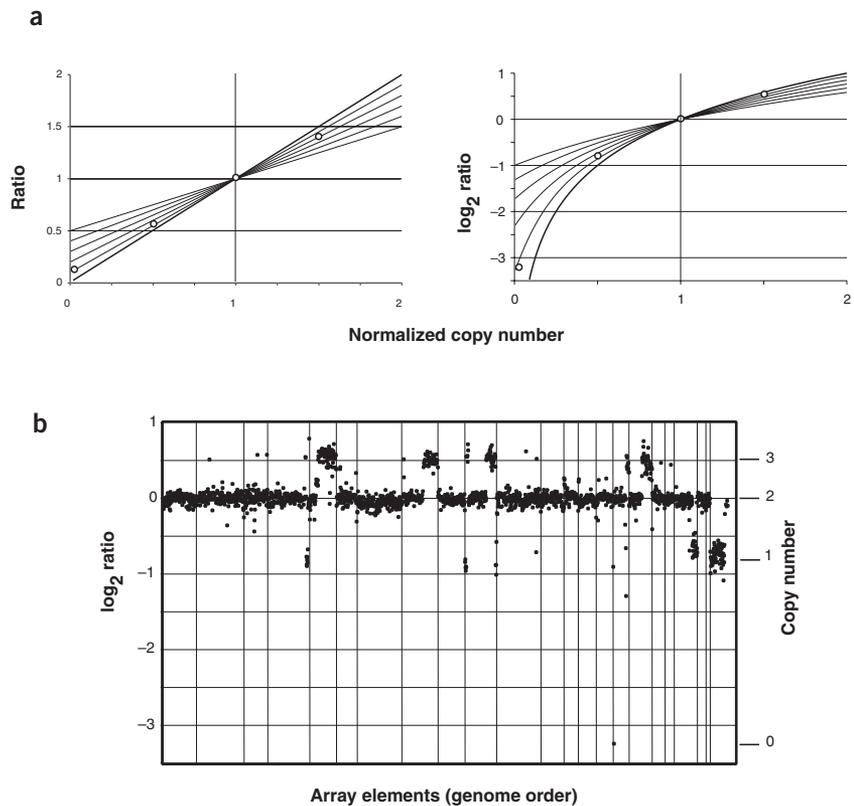
the image analysis software. These effects may lead to very complex and idiosyncratic behavior of different array elements.

CGH measurements are also affected by low-copy reiterated sequences that are common to all individuals and by copy-number polymorphisms among individuals. Low-copy reiterated sequences include members of gene families and blocks of duplicated sequences[28–30]. If a locus that contains such a sequence is changed in copy number, the corresponding ratio change may underestimate the magnitude of the aberration because the other loci with copies of that sequence remain at normal copy number[31]. Conversely, all loci that contain a copy of the sequence may show a ratio change when one locus is altered[28,32]. Copy-number polymorphisms can involve variable numbers of members of gene families, different numbers of copies of a sequence motif within a gene or losses and gains of sequence segments. Many polymorphisms affecting genomic regions of tens of kilobases to several megabases have been detected in human and mouse genomes using array CGH[12,25,33–35]. In addition, the aggregate effect of much smaller-scale, highly distributed, sequence variation among mouse strains and species has also been detected[35].

Our current understanding of copy-number polymorphisms is far from complete. In particular, different publications have used technologies with quite different performance characteristics, and most have

not included sufficient primary data or analytical details to allow comprehensive interstudy comparisons. The further elucidation of dosage polymorphisms will be an experimental rather than a computational endeavor until high-quality sequence is available from a large number of individuals. Understanding the copy-number polymorphisms that are detectable by a particular technique is important so that normal variations are not falsely associated with disease and, conversely, to determine whether some 'normal' variation underlies phenotypic characteristics such as disease susceptibility[36].

Specimen preparation. The quality of genomic DNA preparations has a substantial effect on the resulting data. Although isolation of genomic DNA from fresh and frozen specimens is routinely accomplished through use of numerous published protocols and commercial kits, there is an unknown class of contaminants that occasionally copurify with the DNA and produce abnormally high noise in the ratios. This noise is typically not random because relabeling a different aliquot of the same DNA reproduces the same pattern. In these cases, repurifying or reisolating the DNA may help.

DNA quality issues are especially acute when analyzing formalin-fixed archival tissue. Data obtained from such specimens can range from

excellent (*i.e.*, indistinguishable from fresh tissue) to unusable. Simple diagnostics such as fragment size distribution of the isolated DNA have not been reliable predictors of performance. One of the difficulties may be determining accurately the amount of DNA present in a specimen because contaminants from the tissue section or isolation procedure may interfere with standard fluorimetry and absorbance measurements. Some investigators use PCR to assess the quantity of effective DNA in a specimen[37]. Empirically, increasing the amount of DNA by a factor or two in the labeling reactions often improves results[38].

The amount of specimen DNA is frequently a constraint on CGH measurements. Typical array CGH procedures use between 300 ng and 3 μg of specimen DNA in the labeling reaction, equivalent to ~50,000–500,000 cells. Most protocols use random primer labeling, which also amplifies the DNA, so that several micrograms are used in the hybridization. The need to obtain analyses from small specimens, or small regions of heterogeneous specimens, has motivated efforts to develop whole-genome amplification procedures. The strand-displacing polymerase ϕ29 has been used when the genomic DNA is present in long fragments, permitting analysis of nanogram quantities[39,40]. Several companies offer kits for such amplifications. DNA from formalin-fixed specimens is typically too short for use of this approach. A number of other procedures, including degenerate primer PCR[37,41], two-stage random primer labeling reactions[38], balanced PCR[42], ligation-mediated PCR[43–45] and ligation-circularization of degraded DNA[46], have also been used for DNA from both fresh or frozen and fixed specimens. The use of PCR to generate genomic representations for some methods also amplifies the DNA, allowing analysis of tens to hundreds of nanograms of input DNA[19,20,25]. The judgment of how well any of these techniques works depends on the requirements of the desired application (**Fig. 2**).

Data analysis. A number of primary processing approaches have been applied to obtain ratio profiles. In some cases, normalization involves only a simple overall factor to set the median ratio to some standard value, whereas in others, additional procedures based on spatial and intensity dependence and historical data specific to each array element may also be applied. Occasionally, genomes have so much copy-number variation that the biological importance of the normalization is uncertain because only a very small proportion of the genome is at the 'normal' ratio. Some platforms use data from a single hybridization, whereas others combine data from two measurements with dye reversal. Use of any of the myriad data-adjustment procedures that have been proposed without understanding the underlying processes responsible for the distortions, or without a robust phenomenological validation that the procedures are stable and give reasonable results, runs the risk of introducing systematic errors.

Although the major aberrations in a genome are frequently evident by inspection, a large number of approaches have been developed to improve interpretation in the face of measurement noise. The simplest is to apply thresholds. If the ratio profile has only a few well-spaced ratio levels, then thresholds can be chosen by examination of the distribution of all measured ratios[8]. But many tumors, owing to their non-diploid genomes or heterogeneity, have closely spaced ratio levels that partially overlap because of measurement noise. Thresholds for these tumors cannot be determined by this simple method. Use of smoothing by averaging the ratios on neighboring array elements improves the behavior of thresholding but blurs the locations of boundaries and reduces the amplitude of aberrations involving fewer elements than the smoothing window.

More sophisticated analytical approaches rely on the fact that copy-number changes involve chromosome segments, and so ratios at contiguous sets of loci should be identical, except for an occasional

abrupt step to a new plateau. These methods statistically assess the status of each array element in the context of its neighbors. Among the approaches that have been used are hidden Markov models[47], change-point analysis[48], adaptive weights smoothing[49], Bayesian maximum *a posteriori* probabilities[50] and clustering[51], and many more are under development (several of the software packages are available at http://www.bioconductor.org/). Statistical approaches limited to examination of ratio profiles cannot evaluate the reliability of an aberrant ratio that affects only a single array element. The underlying image data need to be examined to determine their quality, and the interpretation needs be accomplished in light of experience. Single-copy aberrations that affect only one array element can be detected with high sensitivity and specificity with some BAC technologies and may be highly informative[38].

Applications of array CGH in oncology

Tumors develop through the combined processes of genetic instability and selection, resulting in clonal expansion of cells that have accumulated the most advantageous set of genetic aberrations. Many types of instability may occur, resulting in point mutations, chromosomal rearrangements, DNA dosage abnormalities, alteration of microsatellite sequences and epigenetic changes such as methylation. These abnormalities act alone or in combination to alter the functions or expression levels of cellular components. Tumors contain the genetic history of their development, but this history may be difficult to decipher. Some aberrations that are important early in tumor development may be lost or obscured by subsequent events, or no longer be functionally relevant. Others may be neutral or even somewhat detrimental to the tumor but are found because they are present in a cell that developed a sufficiently protumorigenic aberration, or because they are obligate products of the event that produced a crucial aberration. Array CGH, with its ability to analyze DNA from a wide variety of specimens, including those not amenable to other forms of global analysis, provides a powerful entry point for studies of cancer. The leads that are generated frequently motivate follow-up studies that use the complete range of biological approaches, including expression analysis, immunohistochemistry, FISH, DNA sequencing, tissue microarrays and functional studies in tissue culture and animal models.

Tumor genomes have a wide variety of copy-number phenotypes, indicating different types of genetic instability. For example, colon tumors have different levels and types of genomic aberrations, which can be attributed to differences in mismatch-repair competence[52,53]. Analysis of mismatch-repair–proficient and –deficient cell lines has found that the exact nature of the repair deficiency also affects the characteristics of the copy-number changes[54]. Tumors in mouse model systems do not typically contain a large number of informative copy-number changes unless they have been engineered to carry specific genetic defects such as impaired telomeres[55]. The wide range of genomic phenotypes in cancer means that, for some sets of specimens, array CGH will provide information on the locations of important cancer genes, whereas in others, it will be uninformative. Copy-number profiles of cell populations uncover the past genomic instability that led to the clonal expansion of a cell population, the genome of which may then remain relatively stable, at least within its selective environment. For example, some tumors seem very stable *in vivo*, with primaries and recurrences having nearly identical copy-number profiles even though there are many years between them[56,57]. Ongoing genomic instability results in heterogeneity that is not detectable by CGH and is best assessed by techniques that examine individual cells[58].

Knowledge of copy-number aberrations can have immediate clinical use in diagnosis and can, in some cases, provide useful prognostic information. Microarrays designed to analyze targeted genomic regions

relevant to chronic lymphocytic leukemia have been produced for use with clinical trials to facilitate determinations of the relationship between therapeutic options and genomic aberrations[59]. Association of DNA copy-number aberrations with prognosis has been found for a variety of tumor types, including prostate cancer[60], breast cancer[61], gastric cancer[62] and lymphoma[63,64]. Many more studies are in press or nearing completion. As with other types of statistical study, these results require validation on independent sets to control for the possibility of unanticipated systematic factors in the initial groups[65,66].

Identification of important genes in regions of copy-number change is a complex task. If narrow regions of highly elevated copy number or total deletion that contain previously known cancer genes or genes with suggestive function are found, high-probability candidates may be immediately evident. Even if such aberrations are very rare, they may suggest additional measurements that develop support for the common involvement of a particular candidate gene or pathway[38]. In many cases, however, even minimally defined aberrant regions resulting from combining data from many specimens may contain several attractive candidates (or none), or the copy-number aberrations may be complex, making it difficult to determine how many different loci may be under selection. If a gain is greater than a single copy, it is possible that more than one evolutionary step was involved in its formation. This sometimes results in a profile that resembles a peak with sloped sides, suggesting, but not proving, that the critical genes are located near the center of the peak[24,54]. Therefore, it is sometimes useful to interpret the amplitude of copy-number changes in addition to noting their locations.

Measurement of gene expression at the RNA or protein level is crucial for candidate evaluation. If a gene is a target of selection in a region of copy-number increase, it should be overexpressed in tumors in which it is amplified. Unfortunately, overexpression does not distinguish it from other genes in the aberrant region, because as many as 40–60% of all genes in the region may show elevated expression[67–69]. Moreover, genes can be overexpressed for reasons other than dosage increase and therefore may be involved in tumor development even if present at normal copy number. Only rarely (the classic example being human *ERBB2*) are expression changes at the RNA and protein levels essentially perfectly coupled to dosage[70]. Therefore, finding that a gene is always overexpressed when showing an increase in copy number, and sometimes overexpressed when it does not, supports its functional role in cancer. Genes that drive copy-number gains may also be altered by mutation[71], so that sequencing of candidates in tumors with and without increases may provide important information. Similarly, particular alleles of a gene may contribute to tumorigenesis, so that finding a preferential gain of one variant may indicate its functional involvement[72].

Evaluation of genes in regions of copy-number losses is also complex. In some cases, the decrease in expression caused by deletion of a single copy of a gene contributes to tumor development. But in the classic case of tumor-suppressor genes, function is totally abrogated by deletion of all copies of a gene, deletion of one copy and mutation or epigenetic alteration of the other[73], or alteration of one copy and replacement of the other by a duplicate of the altered copy. Deletions of all copies of a genomic region are easily detectable in cell lines by array CGH and other techniques[74], but their reliable detection in tumor specimens is complicated by the likely inclusion of normal cells. The finding of focal homozygous deletions in regions of frequent heterozygous deletion or loss of heterozygosity can provide crucial information to focus searches for important genes. Aberrations that result in loss of heterozygosity but no copy-number change are not detectable by array CGH. The developing SNP-profiling technologies may be able to provide additional information concerning these events, perhaps eventually providing information on heterozygosity and dosage for some types of specimens[19,20]. Candidate genes in recurrent regions of loss can be assessed for expression changes and examined to determine whether the remaining copies are mutated or methylated[73]. One general approach that has proven useful to screen for mutated genes in cultured cells uses nonsense-mediated decay. If a mutation produces a premature stop codon, transcripts are rapidly degraded, and global comparison of expression levels before and after inactivating nonsense-mediated decay identifies genes whose transcript levels have increased. Those that are contained in deletions are candidate tumor suppressors[75].

Conclusion

Array CGH is one of a growing number of 'top-down' approaches that are able to provide comprehensive information about aspects of biological status or function. In the near term, these techniques can provide correlative information that is useful for important clinical applications. In the longer term, there is the hopeful vision that the combination of global measurements will lead to substantial advances in our fundamental understanding of biological processes. But the more traditional 'bottom-up' studies focused on individual functional units always uncover details that are not glimpsed by the global approaches. Conversely, focused studies may be misinterpreted owing to the lack of global information. Thus, improving the ability to integrate bottom-up and top-down information is essential. Looking backward at past accomplishments and forward to the increasingly powerful technical and computational tools that continue to become available may lead to overoptimism about the ease of taking the next steps. Profiling technologies can fill databases at prodigious rates. But they provide little value unless the data are of sufficient quality and are interpreted in the richest possible context. We hope we have succeeded in addressing some of these issues as they pertain to array CGH.

1. Kallioniemi, A. *et al*. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821 (1992).
2. du Manoir, S. *et al*. Detection of complete and partial chromosome gains and losses by comparative genomic in situ hybridization. *Hum. Genet.* **90**, 590–610 (1993).
3. Karhu, R., Rummukainen, J., Lorch, T. & Isola, J. Four-color CGH: a new method for quality control of comparative genomic hybridization. *Genes Chromosomes Cancer* **24**, 112–118 (1999).
4. Mohapatra, G. *et al*. Analyses of brain tumor cell lines confirm a simple model of relationships among fluorescence in situ hybridization, DNA index, and comparative genomic hybridization. *Genes Chromosomes Cancer* **20**, 311–319 (1997).
5. Solinas-Toldo, S. *et al*. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer* **20**, 399–407 (1997).
6. Pinkel, D. *et al*. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**, 207–211 (1998).
7. Snijders, A.M. *et al*. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat. Genet.* **29**, 263–264 (2001).
8. Hodgson, G. *et al*. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nat. Genet.* **29**, 459–464 (2001).
9. Fiegler, H. *et al*. DNA microarrays for comparative genomic hybridization based on DOP-PCR amplification of BAC and PAC clones. *Genes Chromosomes Cancer* **36**, 361–374 (2003).
10. Smirnov, D.A., Burdick, J.T., Morley, M. & Cheung, V.G. Method for manufacturing whole-genome microarrays by rolling circle amplification. *Genes Chromosomes Cancer* **40**, 72–77 (2004).
11. Ishkanian, A.S. *et al*. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.* **36**, 299–303 (2004).
12. Li, J. *et al*. Genomic segmental polymorphisms in inbred mouse strains. *Nat. Genet.* **36**, 952–954 (2004).
13. Pollack, J.R. *et al*. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23**, 41–46 (1999).
14. Mantripragada, K.K. *et al*. DNA copy-number analysis of the 22q11 deletion-syndrome region using array-CGH with genomic and PCR-based targets. *Int. J. Mol. Med.* **13**, 273–279 (2004).
15. Dhami, P. *et al*. Exon array-CGH: detection of copy number changes at the resolution of individual exons in the human genome. *Am. J. Hum. Genet.* **76**, 750–762 (2005).

16. Brennan, C. *et al*. High-resolution global profiling of genomic alterations with long oligonucleotide microarray. *Cancer Res.* **64**, 4744–4748 (2004).

17. Carvalho, B., Ouwerkerk, E., Meijer, G.A. & Ylstra, B. High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J. Clin. Pathol.* **57**, 644–646 (2004).

18. Lucito, R. *et al*. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Res.* **10**, 1726–1736 (2000).

19. Zhao, X. *et al*. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64**, 3060–3071 (2004).

20. Zhou, X., Mok, S.C., Chen, Z., Li, Y. & Wong, D.T. Concurrent analysis of loss of heterozygosity (LOH) and copy number abnormality (CNA) for oral premalignancy progression using the Affymetrix 10K SNP mapping array. *Hum. Genet.* **115**, 327–330 (2004).

21. Pinkel, D. & Albertson, D. Array comparative genomic hybridization. *Ann. Rev. Genet. Genomics* (in the press).

22. Fukiya, S., Mizoguchi, H., Tobe, T. & Mori, H. Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.* **186**, 3911–3921 (2004).

23. Watanabe, T., Murata, Y., Oka, S. & Iwahashi, H. A new approach to species determination for yeast strains: DNA microarray-based comparative genomic hybridization using a yeast DNA microarray with 6000 genes. *Yeast* **21**, 351–365 (2004).

24. Albertson, D.G. *et al*. Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nat. Genet.* **25**, 144–146 (2000).

25. Sebat, J. *et al*. Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).

26. Barrett, M.T. *et al*. Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA. *Proc. Natl. Acad. Sci. USA* **101**, 17765–17770 (2004).

27. Wang, N.J., Liu, D., Parokonny, A.S. & Schanen, N.C. High-resolution molecular characterization of 15q11-q13 rearrangements by array comparative genomic hybridization (array CGH) with detection of gene dosage. *Am. J. Hum. Genet.* **75**, 267–281 (2004).

28. Lupski, J.R., Roth, J.R. & Weinstock, G.M. Chromosomal duplications in bacteria, fruit flies, and humans. *Am. J. Hum. Genet.* **58**, 21–27 (1996).

29. Eichler, E.E. Segmental duplications: what's missing, misassigned, and misassembled–and should we care? *Genome Res.* **11**, 653–656 (2001).

30. Emanuel, B.S. & Shaikh, T.H. Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat. Rev. Genet.* **2**, 791–800 (2001).

31. Zhang, X. *et al*. High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization. *Am. J. Hum. Genet.* **76**, 312–326 (2005).

32. Locke, D.P. *et al*. BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications. *J. Med. Genet.* **41**, 175–182 (2004).

33. Albertson, D.G. & Pinkel, D. Genomic microarrays in human genetic disease and cancer. *Hum. Mol. Genet.* **12** special issue 2, R145–R152 (2003).

34. Iafrate, A.J. *et al*. Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).

35. Snijders, A.M. *et al*. Mapping segmental and sequence variations among laboratory mice using BAC array CGH. *Genome Res.* **15**, 302–311 (2005).

36. Utermann, G. *et al*. Lp(a) glycoprotein phenotypes. Inheritance and relation to Lp(a)-lipoprotein concentrations in plasma. *J. Clin. Invest.* **80**, 458–465 (1987).

37. DeVries, S. *et al*. Array-based comparative genomic hybridization from formalin-fixed, paraffin-embedded breast tumors. *J. Mol. Diagn.* **7**, 1–7 (2005).

38. Snijders, A.M. *et al*. Rare amplicons implicate misspecification of cell fate in oral squamous cell carcinoma. *Oncogene* published online, 11 April 2005 (doi:10.1038/sj.onc.1208601).

39. Lage, J.M. *et al*. Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome Res.* **13**, 294–307 (2003).

40. Hosono, S. *et al*. Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* **13**, 954–964 (2003).

41. Daigo, Y. *et al*. Degenerate oligonucleotide primed-polymerase chain reaction-based array comparative genomic hybridization for extensive amplicon profiling of breast cancers: a new approach for the molecular analysis of paraffin-embedded cancer tissue. *Am. J. Pathol.* **158**, 1623–1631 (2001).

42. Wang, G. *et al*. Balanced-PCR amplification allows unbiased identification of genomic copy changes in minute cell and tissue samples. *Nucleic Acids Res.* **32**, e76 (2004).

43. Guillaud-Bataille, M. *et al*. Detecting single DNA copy number variations in complex genomes using one nanogram of starting DNA and BAC-array CGH. *Nucleic Acids Res.* **32**, e112 (2004).

44. Tanabe, C. *et al*. Evaluation of a whole-genome amplification method based on adaptor-ligation PCR of randomly sheared genomic DNA. *Genes Chromosomes Cancer* **38**, 168–176 (2003).

45. Liu, D. *et al*. LM-PCR permits highly representative whole genome amplification of DNA isolated from small number of cells and paraffin-embedded tumor tissue sections. *Diagn. Mol. Pathol.* **13**, 105–115 (2004).

46. Wang, G. *et al*. DNA amplification method tolerant to sample degradation. *Genome Res.* **14**, 2357–2366 (2004).

47. Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G. & Jain, A.N. Hidden Markov models approach to the analysis of array CGH data. *J. Multivariate Anal.* **90**, 132–153 (2004).

48. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).

49. Hupe, P., Stransky, N., Thiery, J.P., Radvanyi, F. & Barillot, E. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**, 3413–3422 (2004).

50. Daruwala, R.S. *et al*. A versatile statistical analysis algorithm to detect genome copy number variation. *Proc. Natl. Acad. Sci. USA* **101**, 16292–16297 (2004).

51. Wang, P., Kim, Y., Pollack, J., Narasimhan, B. & Tibshirani, R. A method for calling gains and losses in array CGH data. *Biostatistics* **6**, 45–58 (2005).

52. Bocker, T., Ruschoff, J. & Fishel, R. Molecular diagnostics of cancer predisposition: hereditary non-polyposis colorectal carcinoma and mismatch repair defects. *Biochim. Biophys. Acta* **1423**, O1–O10 (1999).

53. Esteller, M. Epigenetic lesions causing genetic lesions in human cancer: promoter hypermethylation of DNA repair genes. *Eur. J. Cancer* **36**, 2294–2300 (2000).

54. Snijders, A.M. *et al*. Shaping of tumor and drug-resistant genomes by instability and selection. *Oncogene* **22**, 4370–4379 (2003).

55. O'Hagan, R.C. *et al*. Telomere dysfunction provokes regional amplification and deletion in cancer genomes. *Cancer Cell* **2**, 149–155 (2002).

56. Waldman, F.M. *et al*. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. *J. Natl. Cancer Inst.* **92**, 313–320 (2000).

57. Albertson, D.G. Profiling breast cancer by array CGH. *Breast Cancer Res. Treat.* **78**, 289–298 (2003).

58. Chin, K. *et al*. In situ analyses of genome instability in breast cancer. *Nat. Genet.* **36**, 984–988 (2004).

59. Schwaenen, C. *et al*. Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations. *Proc. Natl. Acad. Sci. USA* **101**, 1039–1044 (2004).

60. Paris, P.L. *et al*. Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. *Hum. Mol. Genet.* **13**, 1303–1313 (2004).

61. Callagy, G. *et al*. Identification and validation of prognostic markers in breast cancer with the complementary use of array-CGH and tissue microarrays. *J. Pathol.* **205**, 388–396 (2005).

62. Weiss, M.M. *et al*. Genomic alterations in primary gastric adenocarcinomas correlate with clinicopathological characteristics and survival. *Cell. Oncol.* **26**, 307–317 (2004).

63. Martinez-Climent, J.A. *et al*. Transformation of follicular lymphoma to diffuse large cell lymphoma is associated with a heterogeneous set of DNA copy number and gene expression alterations. *Blood* **101**, 3109–3117 (2003).

64. Rubio-Moscardo, F. *et al*. Mantel cell lymphoma genotypes identified with CGH to BAC microarrays define a leukemic subgroup of disease and predict patient outcome. *Blood* published online 17 February 2005 (doi: 10.1182/blood-2004-10-3907).

65. Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A. & Contopoulos-Ioannidis, D.G. Replication validity of genetic association studies. *Nat. Genet.* **29**, 306–309 (2001).

66. Freedman, M.L. *et al*. Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).

67. Pollack, J.R. *et al*. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci. USA* **99**, 12963–12968 (2002).

68. Heidenblad, M. *et al*. Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene* **24**, 1794–1801 (2005).

69. Hyman, E. *et al*. Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.* **62**, 6240–6245 (2002).

70. Press, M.F. *et al*. Evaluation of *HER-2/neu* gene amplification and overexpression: comparison of frequently used assay methods in a molecularly characterized cohort of breast cancer specimens. *J. Clin. Oncol.* **20**, 3095–3105 (2002).

71. Maldonado, J.L. *et al*. Determinants of BRAF mutations in primary melanomas. *J. Natl. Cancer Inst.* **95**, 1878–1890 (2003).

72. Ewart-Toland, A. *et al*. Identification of *Stk6/STK15* as a candidate low-penetrance tumor-susceptibility gene in mouse and human. *Nat. Genet.* **34**, 403–412 (2003).

73. Zardo, G. *et al*. Integrated genomic and epigenomic analyses pinpoint biallelic gene inactivation in tumors. *Nat. Genet.* **32**, 453–458 (2002).

74. Cox, C. *et al*. A survey of homozygous deletions in human cancer genomes. *Proc. Natl. Acad. Sci. USA* **102**, 4542–4547 (2005).

75. Huusko, P. *et al*. Nonsense-mediated decay microarray analysis identifies mutations of EPHB2 in human prostate cancer. *Nat. Genet.* **36**, 979–983 (2004).