Laboratory Investigation (2007) 87, 737–754
© 2007 USCAP, Inc All rights reserved 0023-6837/07 $30.00

MINI REVIEW

Getting it right: designing microarray (and not 'microawry') comparative genomic hybridization studies for cancer research

David SP Tan, Maryou BK Lambros, Rachael Natrajan and Jorge S Reis-Filho

The development of high-resolution microarray-based comparative genomic hybridization (aCGH), using cDNA, bacterial artificial chromosome (BAC) and oligonucleotide probes, is providing tremendous opportunities for translational research by facilitating detailed analysis of entire cancer genomes in a single experiment. However, this technology will only fulfil its promise if studies incorporating aCGH are designed with a full understanding of its current limitations and the strategies available to circumvent them. While there have been several excellent reviews on the current status of this technology, there is currently very little guidance available regarding the appropriate design of experiments incorporating aCGH (including the strengths and weaknesses of each platform), and how best to combine the results obtained from aCGH with other 'omic' technologies, including gene expression. In this review, we present the key design issues that need to be considered in order to optimize aCGH studies, including sample selection, the definition of appropriate experimental objectives, arguments for and against the various microarray platforms that are currently available, and methods for data validation and integration. It is envisaged that future well-designed aCGH studies will enhance our understanding of the genetic basis of cancer, and lead to the identification of novel predictive and prognostic cancer biomarkers, as well as molecular therapeutic targets in cancer.

Laboratory Investigation (2007) **87,** 737–754; doi:10.1038/labinvest.3700593; published online 11 June 2007

KEYWORDS: bacterial artificial chromosomes; oligonucleotides; microarrays; comparative genomic hybridization; design

Microarray-based comparative genomic hybridization (aCGH) was developed in the late 1990s and brought with it the advantages of rapid, high-resolution screening of entire genomes with minimal cytogenetic expertise required for analysis.[1] Since then, progress in microarray technologies have led to the development of various genomic analysis array platforms with even higher resolution, including tiling path bacterial artificial chromosome (BAC) arrays of up to ~50–100 kb resolution[2] and oligonucleotide arrays with a theoretical resolution of up to ~2 kb.[3] By facilitating the detailed analysis of global genomic copy number changes in tumours, new vistas of enquiry can now be explored. These include the possible derivation of new molecular classifications of tumours based on common patterns of genetic aberration,[4,5] remodelling of carcinogenesis and tumour progression by comparing genetic profiles of normal, preinvasive, invasive primary and metastatic lesions,[6–8] and, consequently, the identification of novel molecular therapeutic targets.[9]

The principles of aCGH are similar to those of chromosomal CGH, a technique that has been extensively used for the characterization of the genomes of solid tumours.[1,7,10] Briefly, the procedure was first developed in the form of an experiment where labelled test and reference DNA is hybridized to probes on a microarray, which are then scanned to produce an image of differential signal intensities (ie dual channel/colour microarrays). Based on the normalized Log2 ratios for each specific clone, a genome-wide (semi)-quantitative analysis of copy number changes in a given locus is defined.[10] More recently, single colour (single channel) microarray CGH analysis tools have been developed (see below). Hence, aCGH provides a genome-wide assessment of numerical genomic aberrations in tumours and, depending on the platform, of loss of heterozygosity (LOH) as well (see

Molecular Pathology Team, The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London, UK
Correspondence: Dr JS Reis-Filho, MD, PhD, The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK. E-mail: Jorge.Reis-Filho@icr.ac.uk

Received 24 April 2007; revised 2 May 2007; accepted 3 May 2007

below). This technique is an excellent screening tool for the identification of deletions, gains and amplification, but, via conventional protocols, is unable to detect polyploidy and balanced chromosomal translocations.[7,10]

Microarray techniques are subject to considerable data variability, due in part to variations in methods of DNA extraction, probe labelling and hybridization, the type of microarray platform used, the number and histological type of samples analysed, the methods used for microarray and statistical analysis and results validation.[11] Hence, numerous methodological issues need to be addressed before its impact on translational research can be fully realized. Although there have been several excellent reviews[10,12,13] on the current status of this technology, there is currently very little guidance available regarding the appropriate design of experiments incorporating aCGH, and how best to integrate the results obtained from aCGH with other 'omic' technologies including gene-expression arrays. The 'omic' field is plagued with acronyms and jargons, some of which are summarized in Table 1. In this review, we present the key design issues that need to be considered in order to optimize aCGH studies. These are summarized in Tables 2 and 3.

CHOOSING AND OPTIMIZING SAMPLES: WHAT SAMPLES ARE AVAILABLE?

In designing any microarray studies, the key determinant for all subsequent decisions is the type of samples available for the study as this has a direct impact on the quantity, quality and purity (ie the proportion of DNA belonging to the cells of interest) of extractable DNA for analysis. While a high quantity of 100% pure, good quality DNA can easily be extracted from cell lines, the impact of aCGH studies on translational research in oncology can only be maximized if human tissue is used.

The vast majority of translational research studies successfully incorporating aCGH analyses have used fresh frozen tissues to provide the highest quality nucleic acid for analysis. However, the most widely available resource for DNA remains locked in archival formalin-fixed, paraffin embedded (FFPE) material. This is particularly the case with rare and unusual tumour subtypes where there is a paucity of fresh frozen tissues. In addition, FFPE material is accompanied by a wealth of long-term clinical follow up data, which lend further strength to these studies. Hence, there are tremendous advantages to be gained if FFPE tissue can be utilized in aCGH studies. However, extracted DNA from FFPE is often heavily crosslinked, degraded and fragmented, heterogeneous (ie mix of cells of different genomic composition), and rarely composed of 100% tumour cells, and therefore suboptimal for microarray analysis.[14] Consequently, aCGH profiles of FFPE material generally have larger variances, lower intensities and lower dynamic range compared with hybridizations of fresh frozen tissue and cell-line-derived DNA.

Table 1 Abbreviations and jargons used in microarray-based comparative genomic hybridization analysis

Abbreviation or jargon	
aCGH	Microarray-based comparative genomic hybridization
ASPE	Allele-specific primer extension
BAC	Bacterial artificial chromosome
Chip	Microarray platform
CISH	Chromogenic *in situ* hybridization
CNV	Copy number variations, aka copy number polymorphisms
DOP-PCR	Degenerate oligonucleotide primer polymerase chain reaction
ESP	End sequence profiling
FFPE	Formalin-fixed paraffin embedded
FISH	Fluorescent *in situ* hybridization
HMW	High molecular weight
Log *R* ratios	Log2 of the normalized hybridization intensity of both alleles obtained with Illumina single-nucleotide polymorphism arrays
LOH	Loss of heterozygosity
MIP arrays	Molecular inversion probe arrays
MiRNA	Micro RNA
OaCGH	Oligonucleotide array comparative genomic hybridization
PGL	Predictive gene lists, aka 'gene signature' (ie lists of genes that are predictive of a given outcome)
QPCR	Real-time polymerase chain reaction
SBE	Single base extension
SCOMP	Single-cell comparative genomic hybridization
SISH	Silver *in situ* hybridization
SNP	Single-nucleotide polymorphism
UPD	Uniparental disomy
WGA	Whole genome amplification
WGG	Whole genome genotyping

Currently, the majority of aCGH studies that have reported success in using DNA extracted from archival FFPE cancers to identify copy number changes and putative therapeutic targets have been based on BAC array platforms.[9,15–18] Nonetheless, there have recently been limited reports of success in aCGH profiling of FFPE tumours using cDNA arrays,[19,20] Affymetrix[21] and Illumina[22] single-nucleotide polymorphism (SNP) array platforms as well. Furthermore, a multiplex-PCR-based quality control procedure that can predict the viability of the test DNA for the aCGH analysis has recently been described.[14] When using FFPE or fresh

Table 2 Challenges for microarray-based comparative genomic hybridization study design

Sample related
Sample type
Tissue procurement
Tissue heterogeneity
Intra-tumour heterogeneity
DNA yield and quality
DNA amplification methods
Links with clinical databases
Platform related
Reliability
Availability
Cost
Types of aberrations detected
Resolution
Analysis methods
Integration of results with those of other high-throughput methods
Other
Tumour-specific aberrations *vs* copy number polymorphisms

frozen tissue for aCGH analysis, the purity DNA content from tumour tissue can be ensured by careful microdissection of tumour from surrounding stromal components. Where tumours are heavily infiltrated with stromal and inflammatory cells, microdissection of tissue of which at least 70–75% is composed of tumour (or tissue of interest) has been deemed sufficiently pure for aCGH.[14,23] With the advent of laser capture microdissection, it is now possible to study the genetic features of limited number of cells or small lesions of interest. The limiting step for coupling laser capture microdissection or other microdissection techniques with aCGH has been the small amount of DNA retrieved using these methods. In the study of tumours where most diagnoses are currently made on core needle biopsies, eg breast cancer, the lack of material can present a significant obstacle to detailed molecular analysis.

In an effort to increase the yield of DNA for aCGH, whole-genome amplification (WGA) methods have been developed in order to obtain adequate DNA yields with the highest possible fidelity to the original profile.[24–29] PCR-based amplification methods, including degenerate oligonucleotide primer polymerase chain reaction (DOP-PCR) and single-cell comparative genomic hybridization (SCOMP), have been shown to provide a DNA yield sufficient for CGH analysis; however some regions, particularly those with repetitive sequences such as 1p32-pter, 16p, 19p, and 22q, are reported to be affected by amplification bias/genomic distortion.[30] Multiple displacement amplification (MDA) is a non-PCR-based amplification method which uses bacteriophage Phi29 or large fragment-Bst DNA polymerase for WGA, and is supposed to have a much lower propensity for over/under representation (three-fold *vs* 1000-fold) than PCR-based WGA methods.[24,31] In addition, MDA produces longer products from each priming event than PCR-based methods, theoretically generating equal representation of loci, thus providing nearly perfect coverage of the human genome.[24] Phi29 has only been successfully applied to fresh frozen tissue and cell line DNA in aCGH studies;[24,32,33] although there are anecdotal reports of successful Phi29 amplification of DNA amplified from FFPE tissue samples,[34] this has proven to be inconsistent to say the least. Bst amplification has been shown to produce good results with FFPE DNA as well.[29] However, preferential amplification of regions of known copy number variation have also been observed in microarray CGH experiments where Phi29 amplified test DNA has been used.[24,32] Amplification bias can be compensated for by using similarly amplified reference DNA, with a starting template concentration of >10 ng in the assay, which effectively removes these areas of regional misrepresentation.[24] These studies illustrate the need for caution in applying these amplification methods, and the need to be aware of these biases to enable appropriate corrective measures to be taken before accurate interpretation of results from studies using amplified DNA is possible. To reliably translate the increased resolution of microarray-based CGH into the identification of gene-specific copy number changes, particularly where only small quantities of DNA are available, unbiased amplification methods, which remain elusive, are required.

Recent reports estimate that 12% of the entire human genome is prone to copy number variation (CNV) between European, African and Asian populations.[35] The spectrum of CNVs encompass deletions, insertions, duplications and complex multisite variants,[33] ranging in size from kilobases (kb) to megabases (Mb),[36,37] and CNV maps are now publicly available (http://projects.tcag.ca/variation/ and http://www.sanger.ac.uk/PostGenomics/decipher/). Furthermore, the fact that CNVs have been shown to influence gene expression, phenotypic variation and adaptation by disrupting genes and altering gene dosage,[38,39] thus causing disease or increasing one's risk of developing disease, suggests that CNVs are likely to have a real biological impact on tumorigenesis. This poses a potentially significant confounding factor when interpreting results from aCGH studies using pooled reference DNA from several different individuals, since gains or losses may simply represent CNVs rather than a true cancer specific aberration. The problem may be overcome by using individually matched DNA extracted from normal tissue as reference DNA. However, this is not feasible in many circumstances, particularly with archival samples, and future studies using prospectively collected tumour samples should try to include the accrual of matched normal tissue from each patient (eg blood). If this is not

Table 3 Advantages and limitations in the microarray-based comparative genomic hybridization study design[a]

	Advantages	Limitations
Specimen type		
Cell lines	Readily available HMW DNA Can be used in models for biological validation of findings	No direct association with clinical findings Prone to minor but occasionally significant *in vitro* artefacts
Frozen samples	Results can be linked with relevant clinical data HMW DNA	Limited availability Microdissection possible but not trivial
FFPE samples	Readily available Results can be linked with relevant clinical data	Degraded DNA
Tissue harvest		
Nonmicrodissected samples	High DNA yield	Most samples have $<70\%$ of tumour cells
Needle microdissected samples	$>70\%$ of tumour cells High DNA yield	Not all samples are suitable (eg invasive lobular cancers)
Laser microdissected samples	$>70\%$ (usually $>90\%$) of tumour cells Allows for separation of morphologically or immunohistochemically distinct populations	Low DNA yield DNA amplification methods required
Whole genome DNA amplification	Allows for use of microdissected samples	Most methods induce genomic distortion (ie amplification bias)
Array platform[a]		
cDNA arrays	Platforms readily available and cheap	Low sensitivity and resolution Limited types of input DNA No allelic information Suboptimal quantification of copy numbers
BAC arrays	Robust technology Suitable for analysis of FFPE samples BACs can be used in *in situ* assays for validation of aCGH findings Analysis methods already available	Limited availability from commercial suppliers Challenging chip production Limited resolution No allelic information Underestimates high level copy numbers
Oligonucleotide arrays	High resolution Easy production Highly specific probes	Limited types of input DNA No allelic information Questionable accuracy for detection of low-level copy number gains and deletions
SNP arrays	High resolution Easy production Highly specific probes Provides allelic information	Limited types of input DNA Analysis methods under development Questionable accuracy for detection of low-level copy number gains and deletions

Table 3 Continued

	Advantages	Limitations
MIP arrays	Suitable for analysis of FFPE samples Very accurate copy number analysis	Challenging protocol optimisation Analysis methods under development Limited availability
Data integration	Detailed analysis of the genome, transcriptome, proteome and metabolome	Unprecedented data complexity Analysis tools not fully developed

FFPE: formalin-fixed, paraffin-embedded samples; HMW: high molecular weight.
[a]Please see Tables 4 and 5.

possible, an alternative would be to pool reference DNA from a sufficiently ethnically varied group of control subjects to minimize the prevalence of specific copy number polymorphisms in the reference DNA.

Finally, the issue of tumour heterogeneity also needs to be addressed when designing an aCGH study. A genetic profile derived from aCGH analysis of DNA extracted from a small segment of tumour tissue is representative of the cumulative genetic aberrations of all cells within that tumour segment; in other words, aCGH profiles are usually more representative of, but do not provide an exclusive representation of, the modal population of neoplastic cells. Clonal genetic heterogeneity exists within most tumours,[40] so pooling DNA samples extracted from different regions in a large tumour or extracting DNA from a large tumour area is sensible. In any case, the resulting genomic/gene-expression profile of any tumour obtained from any high-throughput analysis platform will certainly include the profile of the modal tumour population which is likely to be the predominant cell population throughout the tumour, whichever the region sampled. Interestingly, aCGH analysis performed by our lab, comparing three different core biopsies from the same tumour in a cohort of 46 tumours, demonstrated the overwhelming similarity between different samples from the same breast tumour.[41] In a recent unpublished analysis, we observed *KIT* gene amplification by CISH analysis in ~10% of neoplastic cells of a glioblastoma, and subsequently found *KIT* gene amplification by analysing the same sample using aCGH (Reis-Filho JS and Lambros MB, unpublished results). This is not surprising given that neoplastic cells with that amplification harboured >20 copies of the gene. Hence, aCGH preferentially identifies low-level genomic gains and losses seen in the modal population; however, high-level amplification seen in as little as 10% of neoplastic cells can sometimes also be identified using this technique.

WHAT IS THE QUESTION?

Traditionally, molecular genetic research has been hypothesis driven: one hypothesizes the functional role of a candidate gene or molecule and proceeds to validate the hypothesis using various molecular biological approaches with clearly defined objectives.[42] In the process, one might also derive a mechanistic description of its function. Although such approaches are useful when applied to the study of specific molecular pathways in cancer, they become much less applicable in the context of evolving high-throughput technologies and our growing knowledge of the immense heterogeneity of cancer biology. Furthermore, the emerging picture from these high-throughput studies is that of a systems biology disease[43] characterized by multiple defects throughout an overwhelmingly complex interaction of multiple regulatory networks and parallel signalling pathways that would confound any attempt at reducing the molecular pathogenesis of cancer to singular molecular defects, even with a comprehensive mechanistic description of the process. The advent of high-throughput technologies is now placing us in a unique position where we can make use of the increased efficiency afforded by these techniques to devise discovery-based approaches to study different aspects of cancer. This represents a shift from hypothesis-driven *validation* studies to hypothesis *generation* studies. This idea should come as no surprise to the readers of Laboratory Investigation. In 2005, Drs Crawford and Tykocinski[42] emphasized that if we are to capitalize on the unparalleled amount of data generated with high-throughput studies, a paradigm shift in the way data are perceived, hypotheses are tested, and results are shared is absolutely required. Genome-wide profiling of cancer has the potential to identify novel genetic aberrations and therapeutic targets, enhance our understanding of the link between the clinicopathological phenotype and genetics of cancer, and lead to the development of a functional and predictive molecular pathological classification of cancer. In the process, long-held misconceptions regarding the culprit genes and proteins involved in pathogenesis and clinical behaviour of tumours as a result of erstwhile technological constraints may also be rectified. A well-validated, comprehensive molecular genetic characterization of tumours can then serve as a basis for traditional hypothesis-driven approaches to validate the function and tumorigenic role of putative tumour suppressor genes and oncogenes. It is with this in mind that we turn our attention to determining the experimental objectives, and thus asking the right questions in aCGH studies.

A basic principle needs to be acknowledged before deciding on any experimental study: you only get the right answer if you ask the right question and make use of the right tools. When designing microarray studies, there are three commonly adopted approaches, namely class comparison, class discovery or class prediction studies.[11] In class comparison studies, two or more groups or classes of tumours, for example, core biopsies of breast tumours before and after neoadjuvant chemotherapy[41] or invasive ductal *vs* invasive lobular carcinomas,[44,45] are profiled and compared for differences. In class discovery, a group of similar tumours are profiled and subjected to unsupervised cluster analysis in order to derive newly 'discovered' subsets based on differences in the profile. This approach has been employed in deriving molecular subtypes of breast and head and neck cancer[46,47] based on gene-expression profiling. Finally, in class prediction studies, the aim is to derive prognostic or predictive algorithms or patterns based on the profiles derived from tumours. The majority of these class-directed studies have been carried out using microarray-based gene-expression approaches, with the derivation of several predictive gene lists (PGLs) in various tumour types.[25,48] However, the reliability and reproducibility of these early gene-expression-based studies involving limited numbers of patient samples to derive PGLs have been called into question,[49] particularly given the fact that the actual sample sizes required to achieve sufficient power in these class-prediction studies is often considerable, estimated to be more than several thousand, in order to avoid 'over-fitting'.[50] Although aCGH profiles are a direct reflection of structural genomic aberrations, and therefore considerably less capricious than gene-expression profiles, there remains no formal reliable method for optimal power calculations in the design of these studies. However, it would be fair to say that the larger the sample size the higher the accuracy of the class comparison and class discovery analyses. Therefore adequate tissue procurement is paramount in the design and execution of aCGH studies. Although our pathology files are a unique resource of tissue for aCGH translational research studies, the time has come for the development of tissue banks comprising tumour and matched normal samples, linked to an integrated and constantly updated clinical database and the results of other molecular studies performed. We should not forget that pathologists should ultimately be the curators of such tissue banks, given that pathology expertise is of utmost importance for the selection and processing of human tissue samples.[42] Furthermore, if one is to develop accurate PGLs, adequate tissue procurement should be incorporated in the protocols of clinical trials, to enable the development of optimally designed and sufficiently powered high-throughput prognostic and predictive studies. Based more on pragmatism than statistical data, we have adopted a minimum sample size of approximately 50 subjects for class comparison and class discovery studies. Additionally, when deciding on which 'classes' of tumour to use, it needs to be acknowledged that, given the clinical phenotypic diversity exhibited by different histological subtypes of cancer, it is more likely that useful information will be gained from microrarray studies analysing sufficient numbers of individual cancer histological subtypes before attempting cross-subtype comparisons. For instance, when analysing breast cancer samples, it is of paramount importance to remember that histological grade, more than any other clinicpathological feature and known tumour intrinsic characteristic, is associated with the type, pattern and complexity of molecular genetic changes.[7,51–55]

Another way of using aCGH is in the molecular genetic characterization of tumours, leading to the identification of putative genetically important aberrations in carcinogenesis and tumour progression, thus providing a basis for hypothesis testing to determine the clinical relevance of genes within aberrant regions. This approach has led to the identification of novel oncogenes, prognostic markers and putative therapeutic targets in various tumours such as *E2F3* in bladder cancer,[56] *RAB25* in breast and ovarian cancer,[57] *IGF1R* in Wilm's tumours,[58] and *FGFR1* in breast cancer.[9] Here, the emphasis in on discovering individual genetic aberrations that have a significant impact on tumour biology. In addition, this data may be overlaid with data obtained from other high-throughput microarray techniques, for example, expression array data. Amplified and overexpressed genes are likely to represent key 'addictive' oncogene candidates involved in tumour development and progression,[59,60] while homozygously deleted and underexpressed genes may represent important tumour suppressor genes.[61] Once overlaid, putative candidate oncogenes and tumour suppressor genes can be analysed using an integrative biology approach where key molecular pathogenic pathways, and hence potential therapeutic targets, may be identified[62] (see below: Data Integration). It should be noted that homozygous deletions (HOD) are rarely seen in aCGH studies using DNA extracted from tumour samples with a purity of $<70\%$, as the presence of contaminating normal DNA will affect the displacement of the deleted clones, and regions of HOD may only be seen as a region of high level loss. On the other hand, amplifications are much easier to identify, and amplicons as small as 50 kb can be detected using the majority of current platforms.

Clearly the type and number of samples available will affect the feasibility of any of the aforementioned approaches. While thousands of samples will need to be accrued before a class prediction study can be performed, a more limited sample set could be used in class discovery/comparison or molecular genetic characterization studies.

AN ARRAY OF ARRAYS: WHAT IS THE IDEAL MOLECULAR GENETIC PROFILING TOOL?

Overview

In microarray utopia, the ideal genetic profiling platform would have high hybridization intensity, high resolution, low levels of noise or experimental variation, flexibility in terms of input material, a minimal requirement for laboratory work, and a straightforward, user-friendly method for analysis, all being achieved with push-button simplicity—as luck would have it, such a tool does not yet exist. Instead, a host of different platforms, most of which are commercially available, have emerged, each with its inherent pros and cons (Tables 4 and 5).

There are two basic types of genomic array technology: ordered arrays or random arrays. Ordered arrays are manufactured by spotting (using pins) or synthesizing individual probes in an organized pattern on a planar surface. The main problems associated with ordered arrays are concerns with clone management and probe identity due to PCR contamination. Furthermore, spotted (eg cDNA, BAC and some oligonucleotide) arrays are usually prone to batch-to-batch variation of hybridization, and therefore hybridization intensity and spot morphology may vary from batch to batch. Reasons behind this include changes in humidity during the spotting process, ozone content in the microenvironment of the spotting machine and blockages to spotting pins resulting in heterogenous spot morphology. To minimize this, quality control measures need to be adopted during the manufacturing process so that inadequately spotted batches, and the reasons for technical failure, can be identified before any assays are performed. An alternative to spotting is *in situ* (on-chip) light-directed chemical synthesis of the probes on the slide surface, a process known as photolithography,[63] with the advantages of increased ease of manufacture and reduced batch-to-batch variation. However, it is difficult to assess the quality of the oligonucleotides manufactured on the surface. In contrast to ordered arrays, random arrays are constructed by first immobilizing individual probes onto beads which are then pooled and assembled onto a patterned planar surface. This allows an average of ~30 replicates of each probe in the array. The identity of each bead is determined following hybridization of specific labelled complements to the probe sequences on the bead. Currently, the majority of aCGH platforms are ordered arrays, with random bead arrays only commercially available from Illumina (http://www.illumina.com/).

cDNA Arrays

Initial studies on genome-wide approaches to aCGH were performed using cDNA micoarrays which were originally designed for expression profiling.[64] The advantage of this technique is the widespread availability of cDNA clone-sets, thus enabling large-scale production of microarrays, its high spatial resolution, and the ability to directly correlate genomic deletions and amplifications with changes in expression derived from the same method. However, cDNA microarray analysis enables only the detection of aberrations in known genes and expressed sequence tags (ESTs), as cDNA probes are only representative of expressed genes on a chromosome, hence rendering regions without known transcriptionally active genes uncovered. As only exonic regions of the genome are covered by cDNA microarrays, changes in gene regulatory elements such as promoter regions, transcription-factor-binding sites, microRNAs (miRNA) and small nucleolar RNAs (snRNA), and poorly defined ESTs are largely undetectable. The absence of intronic sequences also reduces the stability of the hybridization dynamics, leading to cross-hybridization and reduced sensitivity. Finally, extensive sequence similarities may exist between cDNA clones of paralogous genes, which further complicate the interpretation of array CGH data. Even though this platform has elucidated valuable information, it cannot compete with currently available alternatives in terms of its maximal achievable resolution.

BAC Arrays

BACs, P1-derived artificial chromosomes (PACs) and yeast artificial chromosomes (YACs) are large insert genomic clones which have been widely used in aCGH studies.[3,10,12] BAC probes vary in length from 100 to 200 kb and the resolution (ie the distance between each DNA target represented on the array) of each BAC array is defined by the number of unique probes it contains. The probe content of genome-wide BAC arrays range from 2400 to ~32 000 unique elements (tiling path array). Tiling path arrays (ie arrays where each BAC overlaps with its contiguous BACs) provide a resolution of up to ~50 kb, given that a genomic change can only be detected if it is sufficiently big to significantly change the hybridization intensity in one of the channels (ie change the red:green ratios). These platforms provide sufficiently intense signals for the detection of single-copy number changes, are able to accurately define the boundaries of genomic aberrations, and, importantly, can be applied to archival FFPE tissue as well.[30,65]

One of the main drawbacks with BAC arrays is the high concentrations of high-quality BAC DNA needed to achieve good array performance.[3] Given the low initial yields of DNA from isolated BAC clones, DNA amplification is required to generate sufficient quantities of adequately pure BAC DNA for the assay. Producing a tiling path array is thus costly and highly labour intensive. In addition, as BAC probes are representative of the human genome, they will also contain repetitive sequences, which can lead to nonspecific hybridization. In order to prevent nonspecific hybridization to these repetitive sequences, Cot-1 DNA is often included in the hybridization reaction, adding to the overall cost of the assay. Furthermore, as the human genome is still being updated on a regular basis, mapping inaccuracies of BAC clones often arise. To avoid making incorrect assumptions about the data, all BAC clones should be fluorescent *in situ* hybridization (FISH) mapped and end-sequence verified in the process

Table 4 Comparison of aCGH plaforms

Platform	Number of elements (K)	Type of platform	Size of probes	Channels	Resolution	Optimal sample type	Allelic information
cDNA	Variable (up to ~30)	Spotted cDNAs	Varies from gene to gene	Dual colour	Exonic regions of known expressed genes and ESTs only	Cell lines Frozen tissue FFPE	No
BAC	2–32	Spotted BACs	~100–150 kb	Dual colour	Variable >1 MB–~50 kb	Cell lines Frozen tissue FFPE	No
Agilent	44 244	Spotted Oligonucleotide	60 mers	Dual colour	~35 kb (actual >200 kb) ~6.4 kb (actual ~60 kb)	Cell lines Frozen tissue	No
NimbleGen	385	Photolithography oligonucleotide	45–85 mers	Dual colour	~6 kb (actual ~60 kb)	Cell lines Frozen tissue	No
Affymetrix (SNP array)	250 500 1000	Photolithography oligonucleotide	25 mers	Single colour	~12 kb (actual ~120 kb) ~6 kb (actual ~60 kb) ~3 kb?	Cell lines Frozen tissue FFPE	Yes
Illumina (SNP array)	300 550 650 1000	Oligonucleotide Beadarray	50 mers	Single colour	~5 kb (actual ~50 kb) ~2.8 kb (actual ~28 kb) ~2.0 kb (actual ~20 kb) ~1 kb?	Cell lines Frozen tissue	Yes
Molecular Inversion Probes (MIP)	20	Spotted oligonucleotide	41–61 bp	Four colour	Exon level changes	Cell lines Frozen tissue FFPE	Yes

of array construction, and any data derived from aCGH should be validated using *in situ* hybridization techniques (ISH), for example FISH,[9,58,66] CISH,[15,67,68] or Silver-ISH (SISH) (www.ventanamed.com) or with real-time copy number PCR.

Oligonucleotide Arrays

Oligonucleotide array aCGH (OaCGH) platforms consist of single-stranded 25–85 mer oligonucleotide elements.[3,13] Different types of oligonucleotide arrays have different labelling and hybridization protocols and can provide high-resolution copy number measurements.[3] There are two main types of oligonucleotide arrays: SNP arrays and non-SNP arrays. Non-SNP arrays are comprised of 60–75 mer oligonucleotides with site-specific sequences across the genome. SNP arrays are comprised of oligonucleotides that correspond to SNPs along the human genome and were originally designed for use in linkage analysis and whole genome genotyping (WGG). Hence, in addition to allelic copy number changes, SNP arrays can also provide information regarding LOH and copy neutral genetic anomalies such as uniparental disomy (UPD) and mitotic recombination (Figure 1).

SNP-OaCGH platforms

Affymetrix is a commercial SNP aCGH platform comprised of ~25 mer oligonucleotides photolithographically synthesized on the arrays (http://www.affymetrix.com/). As this is a single channel array, only test DNA needs to be labelled and hybridized. The labelling of the test sample involves the use of a restriction enzyme (*Nsp*I or *Sty*I)-based complexity reduction procedure requiring at least 250 ng of DNA. Digested DNA is ligated to adaptors that recognize the cohesive four base-pair (bp) overhangs and amplified using a generic primer that recognizes the adaptor sequence. Amplified DNA is subsequently fragmented, labelled, and hybridized to the oligonucleotide array. The variation per element on the array is relatively high, which gets compensated by the large amount of elements on the array, currently 250 000 per array

Table 5 Types of array platforms and their impact on the design of microarray comparative genomic hybridization analysis

	BAC arrays	Oligonucleotide arrays		MIP arrays[a]
		Non-SNP arrays	SNP arrays	
Availability				
Academic sources	Yes	Limited	No	No
Industry	Limited	Yes	Yes	Limited
Types of samples[b]				
Cell lines	Yes	Yes	Yes	Yes
Frozen	Yes	Yes	Yes	Yes
FFPE	Yes	Possible, but not trivial	No	Yes
Required amount of DNA input	Variable	High	High	Variable
Cost				
Implementation	High	High	High	High
Chip	Low	High	Variable	High
Reagents	High[c]	Variable	Low	High
Types of changes detected				
Single copy number gains	Yes (optimal)	Yes (challenging)	Yes (challenging)	Yes
Amplifications	Yes	Yes	Yes	Yes
Deletions	Yes	Yes	Yes (challenging)	Yes
Copy number silent LOH	No	No	Yes	Possible
Analysis tools[d]	Readily available	Readily available (principles similar to those of BAC arrays)	Limited[e]	Limited
Publicly available data for data mining[d]	Readily available	Limited	Available	No

[a] Technology still in development.
[b] Using standard protocols.
[c] Cot-1 DNA and fluorophores are particularly expensive.
[d] As of March 2007.
[e] Under development by several groups; MIP: molecular inversion probes.

(Mapping 500K Array Set which comprises 2 × 250K elements per array). A 1 million SNP array is anticipated later this year. Recently, Affymetrix have introduced the SNP Array 5.0 chip which is a single microarray featuring all SNPs from the Mapping 500K Array Set, as well as 420 000 additional nonpolymorphic probes that can measure other genetic differences like CNV.

Illumina is a commercially available random bead-array whole genome genotyping (WGG) platform that also allows combined DNA copy number and LOH analysis. There are currently several types of high-density SNP-array platforms manufactured by Illumina; the exon-centred Sentrix Human-1 SNP beadchip (109K), the HumanHap300 (317 tag SNPs), and a higher density HumanHap550 (550K tag SNPs).[13] The Illumina WGG protocol consists of four automated steps beginning with whole genome amplification, hybridization to an oligonucleotide array, an SNP scoring assay which involves either an allele-specific primer extension (ASPE) one-colour assay (Infinium I) or a single-base extension (SBE) two-colour assay (infinium II).[13] The Illumina platform works best with relatively intact, high-quality DNA. For the Infinium Assay, a total 750 ng of DNA with fragment sizes of at least 2 kb is recommend, which would largely exclude the use of DNA extracted from FFPE tissue. Illumina have also recently released the HumanCNV370-Duo DNA Analysis BeadChip which contains the standard SNP content of the HumanHap300 Genotyping BeadChip with an additional ~55 000 markers designed to specifically target nearly 11 000

CNV regions. A 1 million Illumina SNP array is also scheduled for release this year.

Non-SNP OaCGH platforms

Agilent Technologies (http://www.agilent.com/) originally manufactured array platforms comprised of 60 mer oligonucleotides for expression analysis. They now offer an oligonucleotide array, also comprised of 60 oligonucleotides, designed specifically for aCGH as well. These arrays can be purchased with up to ~236K (Agilent Human Genome CGH Microarray Kit 244A) unique oligonucleotides per array. The assay requires about 1 μg of DNA, which poses a problem for samples where only small quantities of DNA are available. This problem is overcome by incorporating a PCR amplification procedure, but, inevitably, this adds a further source of variation to the assay and increases its overall cost.

NimbleGen (http://www.nimbledgen.com) is another commercially available oligonucleotide platform where 385K isothermal oligonucleotides are photolithographically synthesized on a single glass slide.[3] This is a highly flexible platform in which each array can be designed and produced with different probe sets, thus allowing arrays designed for analysis at whole-genome level or focused on specific chromosomal regions. The probes vary in length between 50–75 mer and the platform affords a theoretical resolution of ~6 kb in a human whole genome CGH array. Using the

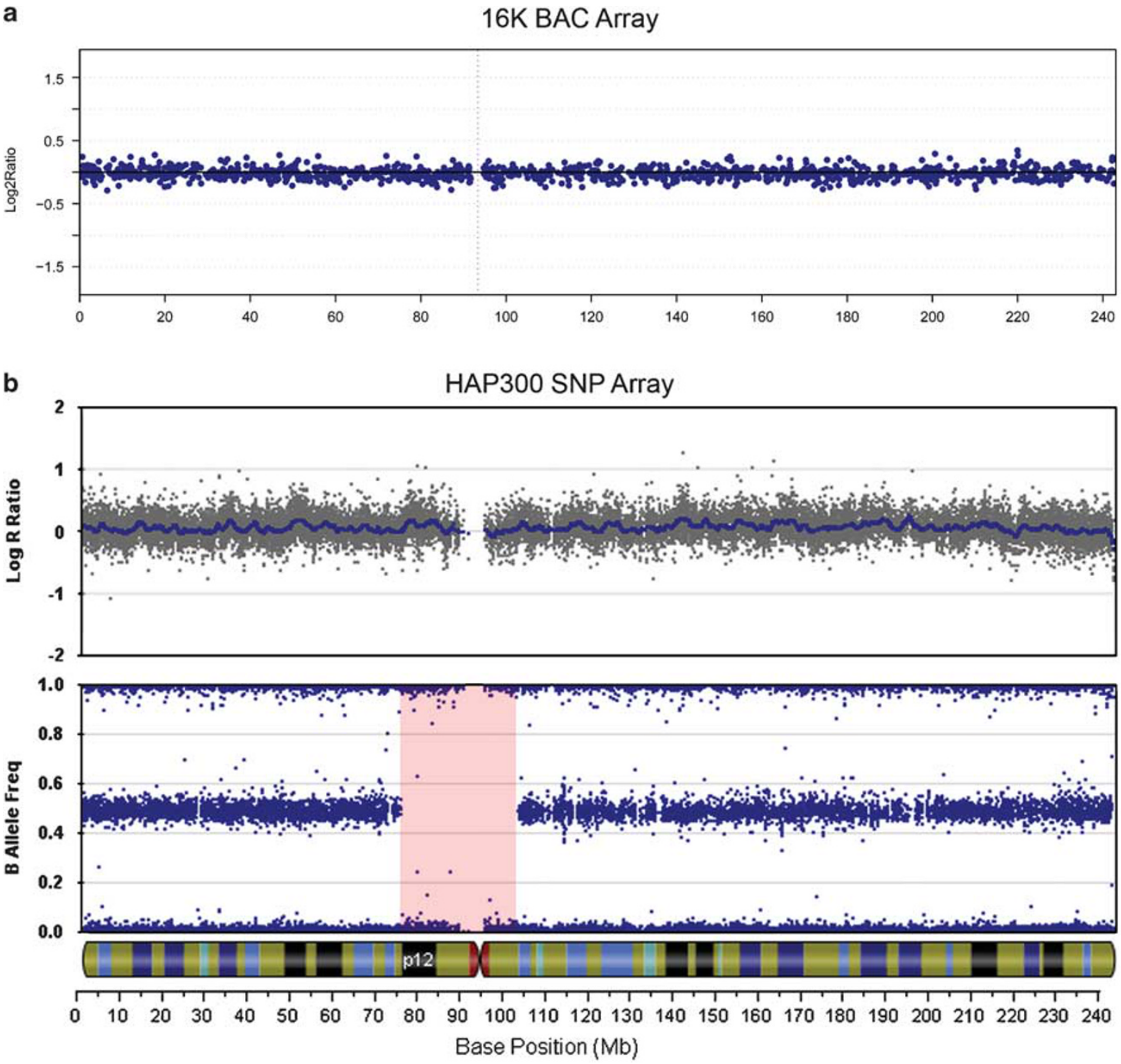

Figure 1 Copy number silent loss of heterozygosity (LOH). Chromosome plot of a high-grade breast cancer displaying no copy number aberrations (ie Log2 ratios centred around 0) on chromosome 2 using our in-house 16K BAC array platform (**a**) and Hap300 Illumina SNP platform (**b**, Log R ratios plot). Note the presence of a loss of the heterozygous allele in the B Allele Frequency Plot (red box). NB: In regions without LOH, B allele frequency data points can be seen at 0 (homozygous A, ie no allele B), 50% (heterozygous) and 100% (homozygous B, ie only allele B). In regions of LOH (red box), the heterozygous features (ie B allele frequency of 0.5) are lost. This figure exemplifies one of the typical profiles found in loci displaying copy number silent LOH, which cannot be identified with BAC arrays or non-SNP oligonucleotide arrays. Given that there is no change in DNA content in copy number silent LOH events, Log2 ratios/Log R ratios do not show any changes in copy number; however, these events can be identified through a thorough analysis of B allele frequency plots by the identification of regions of loss of the ‘heterozygous features’ (ie loss of B allele frequency of 0.5).

NimbleGen platform, Lucito *et al*[69] have described a complexity-reduction method, called ROMA (representational oligonucleotide microarray analysis), involving restriction enzyme digestion of both test and reference DNA followed by PCR amplification, which increases the concentration of DNA complementary to the probes, thus increasing the signal intensity from specific hybridization and consequently reducing the variation in signal intensity from similar copy number changes. Using ROMA, 50 ng of test DNA is sufficient for the assay, provided both test and reference DNA are similarly digested and amplified to exclude biases induced by PCR amplification. This approach has recently been applied to the characterization of breast cancer molecular genetic profiles and provided tantalizing results,[70] which may pave the way for a novel molecular genetic taxonomy for breast cancer.

Oligonucleotide *vs* BAC Arrays

Compared to BAC arrays, oligonucleotide arrays are easier to manufacture, have a greater intrinsic scalability to allow detection of higher feature densities, and can theoretically provide a much higher spatial resolution and locus discrimination. In addition SNP-aCGH oligonucleotide platforms also facilitate the detection copy-neutral allelic imbalances, which may have aetiological significance in carcinogenesis (eg the second event leading to *BRCA1* and *BRCA2* is often copy number neutral LOH).

The main limitation with OaCGH platforms is a higher probe-to-probe variation and sequence dependence of hybridization in the arrays, due in part to greater variation in hybridization dynamics, which may be a function of probe length (oligonucleotide probes are ~100 kb shorter then BAC probes), leading to higher variation in signal intensity for similar copy numbers (Figure 2). On the other hand, the characteristics of the oligonucleotides also translate into more background noise and lower signal intensities for each probe, such that the dynamic range of signal intensity ratios is narrower, thus making it more difficult to discern low-level gains and losses. Consequently, although theoretically affording a resolution as high as ~2 kb, signals from several probes (~5–10) need to be averaged before a call can be made. However, given the scalability of OaCGH arrays, the resolution of these platforms can easily be improved by increasing the feature density (ie number of SNPs) in the array. Hence, an array with 500K SNPs per slide, allowing an averaging of 10 SNPs for each call, will allow a resolution of ~50 kb. With the anticipated introduction of a 1000K SNP array later this year, the effective resolution of OaCGH arrays may increase to as much as ~25 kb. Alternatively, the data need to be analysed using algorithms that reduce the experimental variation for regions with similar copy numbers (ie smoothing algorithms), such as adaptive weighted smoothing (aws),[71] maximum likelihood models, hidden Markov models,[72] or row Loess methods[73] and Gaussian smoothing,[74] as per BAC array data, before confidently defining genomic changes. In simplistic terms, these analytical methods transform aCGH data by organizing a user-defined consecutive sequence of adjacent signals into regions of constant copy number known as segments, which are subsequently classified as a gain, a loss, or no change depending on their signal intensities. As a result, the resolution of OaCGH arrays will decrease depending on the averaging or smoothing window.

Regardless of resolution, however, accurate gene mapping information (NCBI genome build) is of paramount importance to define regions harbouring copy number aberrations, which is the responsibility of the manufacturing companies. Hence, like with BAC arrays, results from OaCGH must be validated as well. However, while the same probes in BAC arrays can be used for validation using CISH or FISH, this is not possible in the case of oligonucleotide probes since, given their small size, the signals will similarly be too small for detection by means of *in situ* techniques (eg FISH or CISH). Finally, the incompatibility of most OaCGH array platforms with FFPE DNA remains to be addressed. However, there have been preliminary encouraging reports suggesting that Agilent Technologies oligonucleotide platforms may be suitable for aCGH analysis of DNA extracted from FFPE sections (http://www.agilent.com).

Molecular Inversion Probe Arrays

Molecular inversion probes (MIPs) represent single oligonucleotides with two inverted recognition sequences at the flanks that recognize and hybridize to targeted genomic DNA sequences ranging between 41 to 61 bp in length (http://www.affymetrix.com/technology/mip_technology.affx). After the probe specifically hybridizes to the target DNA, a single base-pair gap exists in the middle of the two recognition sequences. This gap can either be an SNP or a nonpolymorphic nucleotide. Following a series of specific enzymatic steps, the gap is filled with an appropriate oligonucleotide resulting in the formation of a circularized probe that subsequently undergoes unimolecular rearrangement (ie probe inversion), which enables the probe to be amplified. Crossreacted or unreacted probes are separated from the resulting circularized probe via an exonuclease reaction. Each MIP oligonucleotide has a unique sequence barcode tag which can be assayed via a tag microarray once it anneals to its specific complementary genomic sequence and is circularized.[13]

MIP technology has several theoretical advantages over other array platforms: (1) the non-allele-specific unimolecular design of the assay, coupled with the constraint of dual-recognition sequences, enables multiplexing of >10 000 individual probes without background from crossreactions between probes, thus conferring significant advantages with regard to probe specificity and performance, and thus the robustness of genotype and copy number calling; (2) no PCR amplification is required at the point of mutation detection, thus reducing the risk of amplification bias and overall cost

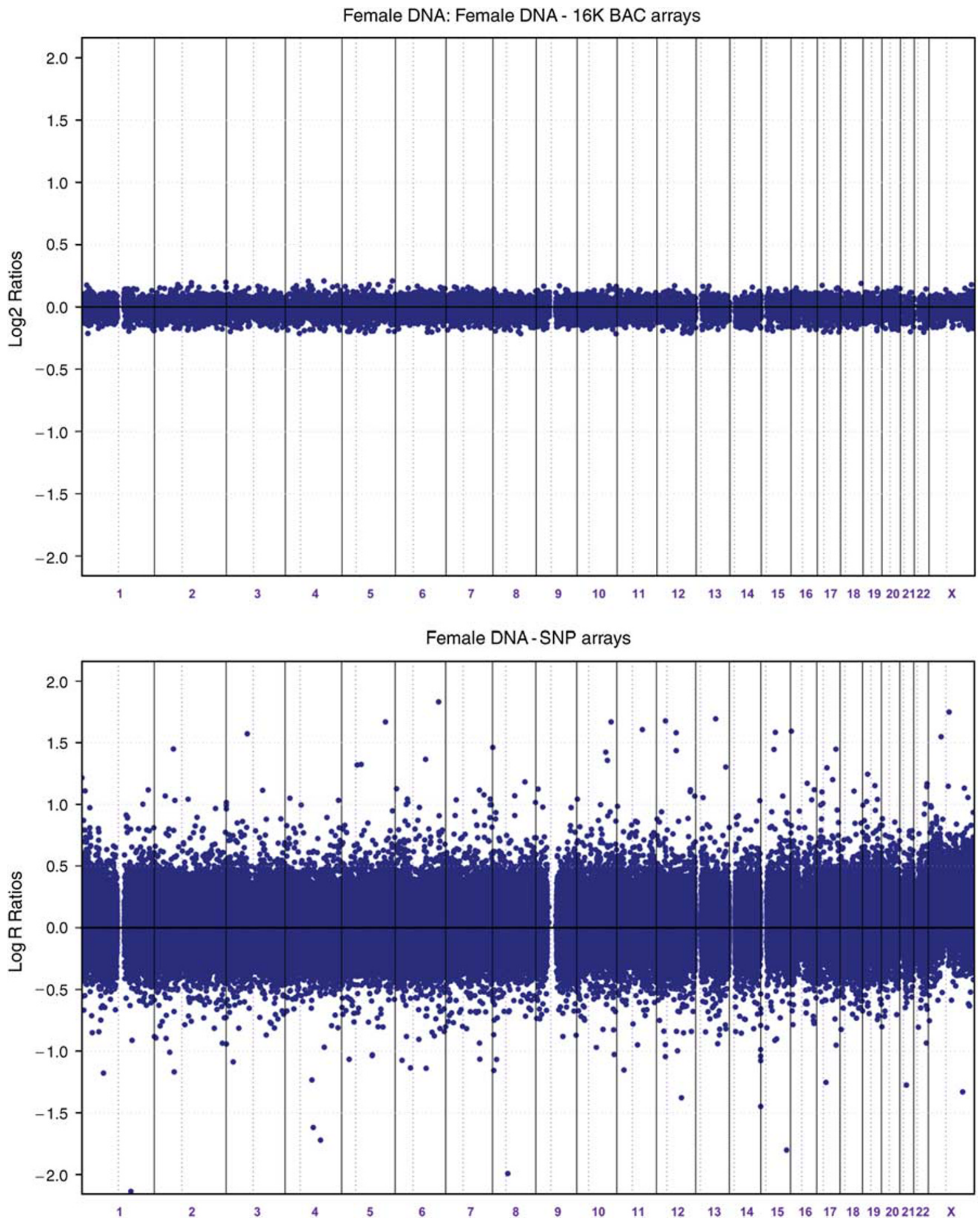

Figure 2 Comparison of the dynamic range of BAC and SNP arrays. Genome plot illustrating the molecular genetic profile obtained with DNA samples extracted from two pools of six healthy female blood donors using our in-house 16K BAC array platform (top) and the molecular genetic profile of a healthy female blood donor using the Illumina Hap300 SNP platform (bottom). Log2 ratios (top) and Log R ratios (ie, Log2 of the sum of the normalized hybridization intensity values for alleles A and B, bottom) are plotted on the *y*-axis against each clone according to genomic location on the *x*-axis. The centromere of each chromosome is represented by a vertical dotted line.

of the assay; (3) unlike aCGH platforms where probes are designed specifically to perform optimally under one set of hybridization conditions, there is greater flexibility in terms of designing MIP probes as signals are assayed using a tag array; (4) hence, one can use nearly any unique sequence and choose specific exons or other interesting sequences when designing MIP probes. Therefore, unlike other aCGH platforms, expanding or refining a gene copy analysis for different genomic regions simply requires the design of new oligonucleotides rather than a new microarray; (5) hence, this technology is particularly well suited to identifying genomic deletion mutations at a very high resolution, for example, at exonic and microsatellite marker level changes,[75] and has been reported to work with DNA from FFPE tissue as well.[75,76] However, this technology has not yet been extensively tested and the current MIP assay requirement for 2 μg test DNA[76] makes this technology prohibitive for samples with limited amounts of DNA available. Currently, a 20K platform is commercially available from Affymetrix but up to 120 000 SNPs can be assayed in a single array.

TILING A PATH TO EUREKA!—DESIGNING AN ACGH STUDY

Once the samples to be tested have been identified and the study objectives have been defined (ie class-directed or molecular characterization), the choice of platform is then dependent on the type of sample available. On current evidence, BAC arrays appear to be more suited to aCGH using DNA extracted from FFPE tissue. Where fresh frozen tumour samples are unavailable, and FFPE DNA is not of sufficient quality for aCGH analysis, then an alternative approach, whereby cell line DNA is interrogated in the first instance followed by validation (eg using *in situ* techniques) of identified genomic aberrations of interest in a larger set of FFPE tissue samples, may be employed (Figure 3). This is a particularly plausible approach for studies where the objective is molecular genetic characterization of tumours for the identification of putative oncogene and tumour suppressor gene candidates. Indeed, recent articles have shown that many of the genomic aberrations observed by using tumour samples for aCGH analysis are also reflected in the genomic profiles obtained from cell lines.[60,77]

Evidently, a platform with higher resolution is likely to provide a more comprehensive picture of the global genomic aberrations in a tumour. For example, while studies had previously reported DNA amplification at 8p12–p11.2[78] in breast cancer, which is associated with a poorer prognosis, it was only after fine-mapping using a high-resolution BAC array that the remarkably complex structure of this amplified region, which is composed of at least four distinct amplification cores, emerged.[79] Similar findings have been described for other amplicons in breast cancer.[80] Clearly, the ease of detecting any particular genomic copy number aberration is inversely proportional to its size (length) and the number of elements involved. Hence, a large 1 Mb region with multiple copy number gains would encompass multiple elements and be detected by most array platforms, while a small 20 kb region with only a single copy number change (eg microdeletion) would be beyond the resolution of most array platforms, with the possible exception of MIP arrays and some SNP arrays. Obviously, cost is also another major consideration, and if high-resolution arrays are unavailable, one might even choose to combine a lower resolution BAC array for global genomic analysis before fine mapping of an individual chromosome or genomic region of interest using custom-designed oligonucleotide arrays (eg NimbleGen).[81]

Regardless of resolution, however, some platforms are better at picking up specific anomalies than others. Hence, if a global picture of large scale gains/amplifications and losses is all that is required, then any array platform will suffice, provided it is of the desired resolution. Alternatively, if the detection of more subtle aberrations involving copy number neutral allelic imbalances such as UPD and mitotic recombination is required, then SNP arrays are the platform of choice. It is also worth noting that a method for identifying balanced chromosomal translocations using both BAC and oligonucleotide arrays, known as array painting, has been developed.[82,83] Similar in concept to reverse chromosome painting, array painting involves fluorescent labelling of flow-sorted chromosomes followed by hybridization to a BAC or oligonucleotide microarray, thus facilitating the detection of cytogenetically balanced chromosome rearrangements.[82,83]

ANALYSIS AND VALIDATION

This review is focussed on study design rather than analysis, and a comprehensive review of analytical methods is thus beyond its scope. Suffice to say that to some extent experimental design and objectives will determine the type of analysis used. Extensive reviews of various analytical tools and methods are available,[84,85] and numerous bioinformatics software packages designed for the analysis of aCGH data are available from commercial sources and the worldwide web. These include regularly updated versions of the *R* data transformation and statistical analysis program (http://www.r-project.org/) and Bio-Conductor (http://www.bioconductor.org/).[86] In addition, all commercially available aCGH platforms come with their own specific data analysis software (eg Illumina Beadstudio and QuantiSNP[87]) and bioinformatics support. Although these methods have the potential to convert aCGH data into meaningful information, there are a few issues in data analysis, directly affected by study design, which should be discussed.

Having established an aCGH profile, the next challenge is to define the boundaries between normal copy number, low-level gains and losses, and amplifications and homozygous deletions. One approach would be to first establish the thresholds for normal copy number, by running normal test *vs* normal reference experiments using matched normal DNA or, if this is not available, a pooled reference DNA sample. Some have suggested that at least six unrelated, healthy individuals should be included in the pooled sample to reduce the false discovery of aberrations in the control sample caused by CNVs.[14] However, as the true prevalence of CNVs remains unknown,[88] we prefer to err on the side of caution, and have incorporated ~24 unrelated, healthy and ethnically varied individuals in our pooled samples for aCGH studies.

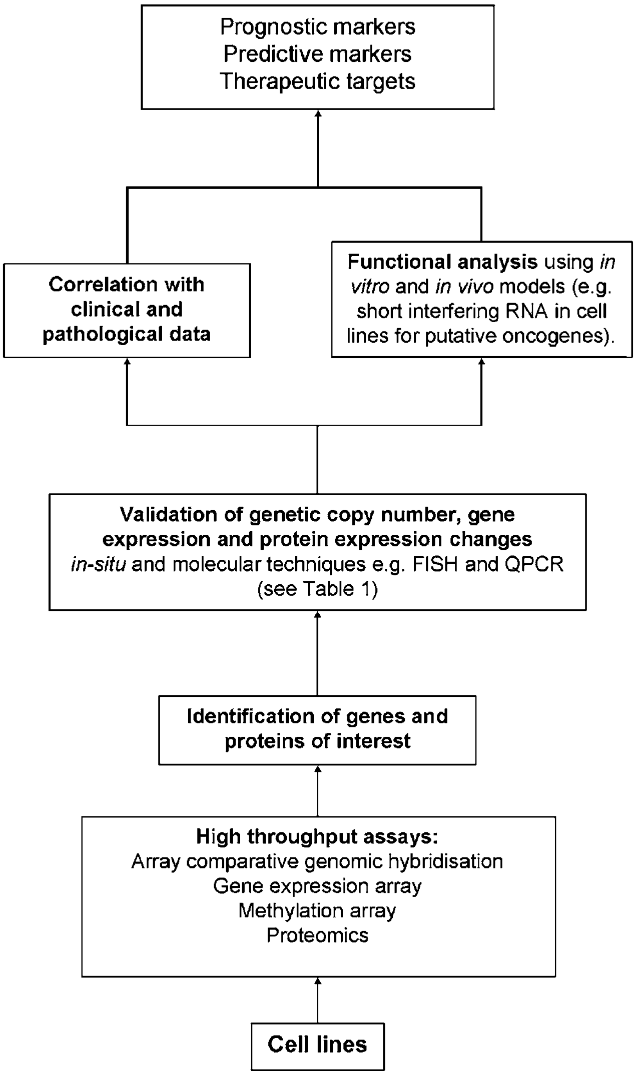
Prognostic markers
Predictive markers
Therapeutic targets
Correlation with clinical and pathological data
Functional analysis using *in vitro* and *in vivo* models (e.g. short interfering RNA in cell lines for putative oncogenes).
Validation of genetic copy number, gene expression and protein expression changes
in-situ and molecular techniques e.g. FISH and QPCR (see Table 1)
Identification of genes and proteins of interest
High throughput assays:
Array comparative genomic hybridisation
Gene expression array
Methylation array
Proteomics
Cell lines

Figure 3 Diagram illustrating an alternative approach for integrating data from different high-throughput methods.

Based on the standard deviation derived from these experiments, thresholds should be set taking into account the balance between the false discovery rate and the ability to detect low-level copy number changes.[8,89,90] Alternatively, as discussed earlier, segmentation algorithms can be used to define the boundaries of copy number changes.[71–73] We have successfully used a combination of the two approaches and identified copy number changes that can be CISH/FISH verified.[8,9,15,91,92]

Whichever methods utilized in the analysis of aCGH data, it is vitally important that the invariably and often excessively large volume of data generated is appropriately curated and validated with *in situ* or molecular methods. These include quantitative real-time PCR (QPCR) techniques[93] and FISH, CISH, or SISH.[9,58,66–68,91] Additionally, in studies where matched normal DNA samples are not available, all regions of copy number change should be cross-referenced with available CNV databases.

DATA INTEGRATION

Having successfully identified and validated an extensive candidate gene list from aCGH analysis, the next questions is how to define a shortlist of the most likely biologically relevant genes. This is where data integration is particularly useful. We have previously discussed the value of overlaying expression and genomic array data to pinpoint ‘addictive oncogenes’ and tumour suppressor genes, but this process can also be extended to using data derived from other high-throughput studies such as proteomics. Recent papers have addressed the issue of integrating protein abundance and mRNA transcript levels from high-throughput analysis,[94] and, indeed, the integration of genomic, gene-expression and proteomic data.[95]

Integrative analysis of high-throughput data confers several advantages:[43,94–96] (1) the impact of methodological unreliability is reduced by cross validation between data from different biological (ie genomic, gene transcription or protein expression) levels; (2) integration of data insensitive to minor spatiotemporal flux (eg genomic copy number changes) with data subject to dynamic changes (eg mRNA and protein expression); (3) improved understanding of the multilayered complexity of various disease and physiological states; (4) facilitate the development of a systems biology approach, for example, mathematical models, in elucidating the intra- and interbiological-level interactions between the signalling networks and pathways that determine disease phenotype. However, such integrative approaches are fraught with complex logistic and analytical challenges, chief among which are: (1) tissue procurement—lack of sufficient or appropriate sample material (eg core biopsies of FFPE tumours) for use in different assays; (2) the immense biological complexity along the progression from genotype to phenotype and sources of variation and functional complementarity within each biological level[95] that exists for any one individual, set against the background of patient and tumour heterogeneity; (3) the need for the development of sufficient bioinformatics expertise, and the necessary software and hardware to process the vast amount of data generated by combining high-throughput technologies. Furthermore, most systems have not incorporated all levels of complexity; for instance, miRNA data have been largely neglected in most models.

While such state-of-art analytical approaches remain aspirational, data integration on a more modest and pragmatic scale is certainly possible. For example, by subjecting a set of tumours to both aCGH and gene-expression array analysis, putative oncogene candidates within a validated region of recurrent amplification can be interrogated at the level of gene expression to identify a shortlist of genes where a good correlation exists between amplification and mRNA overexpression.[9,59,60,79] From this shortlist, immunohistochemical analysis of protein expression in a larger series of FFPE tissues can be performed to confirm the presence of overexpression of certain genes in tumours. Immunohistochemical studies on the prognostic and

Table 6 Six key issues in designing microarray CGH studies

Key issues	Factors to consider
(1) What samples are available?	Sample size (ie number of samples)
	Sample type
	Cell lines
	Formalin fixed paraffin embedded (FFPE)
	Fresh frozen tissue
	Sample DNA quality, quantity and purity
	Reference DNA
	Matched blood samples for DNA extraction (to exclude CNVs)
	Pooled DNA from various ethnic groups
(2) What is the question?	Class prediction
	Class comparison
	Class discovery
	Molecular genetic characterisation
(3) Which array platform?	Cost and availability
	DNA quality, quantity and purity (FFPE/fresh frozen/cell line)
	Type and size of genomic aberration to be detected
	Large copy number changes *vs* single copy number changes—resolution of aCGH platform
	Global genomic profile *vs* copy number neutral allelic imbalances—non-SNP *vs* SNP arrays
	Balanced translocations (array painting)
(4) Analysis	Normalisation
	Data analysis software
	Commercially available platform-specific data analysis software
	Publicly available data analysis software (eg bioconductor in R and BRB array tools)
	Bioinformatics support
	Awareness of CNV in regions with identified copy number changes
(5) Validation	Molecular biology techniques, for example, real-time quantitative PCR
	In situ and cytogenetic techniques, for example, FISH, CISH or SISH on samples tested (test set), followed by a larger sample set (validation set)
(6) Data integration	Immunohistochemistry for protein expression
	RNA/microRNA expression and methylation arrays
	Exon expression arrays
	ESP and other genome-wide sequencing techniques
	Proteomics
	Integrated Systems approach: overlaying aCGH, RNA and/or protein expression data with other high-throughput technologies followed by biological systems modelling → molecular pathways/networks
	Gene functional studies using *in vitro* (eg siRNA for putative oncogenes) or *in vivo* model systems, for example, transgenic mice
	Gene sequencing for mutations

CISH, chromogenic *in situ* hybridization; CNV, copy number variation/polymorphism; ESP, end sequence profiling; FISH, fluorescent *in situ* hybridization; FFPE, formalin-fixed, paraffin-embedded; PCR, polymerase chain reaction; siRNA, small interference RNA; SISH, silver *in situ* hybridization; SNP, single-nucleotide polymorphism.

predictive significance of these proteins will also provide a further layer of evidence with regards to phenotypic relevance. From this shortlist, selected genes which are amplified, overexpressed and shown to be of prognostic significance can be investigated to *in vitro*, for example, using short interfering RNA (siRNA) to knock them down in cell lines harbouring this specific gene amplification and overexpression, to derive their functional significance.[9] Tumour DNA can then be subjected to gene sequencing to look for mutations of functionally significant genes. Biologically and clinically relevant oncogenes thus identified are likely to represent key amplicon drivers and potential therapeutic targets, and, by mapping *a posteriori* to known gene networks, serve as a basis for further investigation of other key upstream and downstream regulatory molecules, pathways and networks in cancer development, and progression as well.

FUTURE OF ARRAY-BASED CGH

Despite the rapid advances in aCGH over last 10 years, the current technology and analysis tools only provide a rough map of genomic aberrations in the genome. Even if the resolution, sensitivity, mapping accuracy, and reproducibility of these arrays improve to the extent that microamplifications and -deletions of less than a kb can be confidently called, they will still reveal little definitive information about the complex modification of various regulatory and epigenetic elements, or the presence of translocations and gene fusions, that are represented by these changes. Indeed, if the ultimate goal of sequencing the entire human genome at the cost of US$1000 becomes a reality, aCGH might eventually be rendered obsolete. Highly parallel sequencing technologies that can provide both quantitative and qualitative assays of the human genome sequence are being developed[13] and are now commercially available (eg www.illumina.com).

The fact is that aCGH currently remains a crude, albeit powerful, screening tool, and should be used as an adjunct to other molecular techniques. Although aCGH has definitely expedited the identification of novel amplicons and tumour suppressor genes,[57,58,91,97] this is only a step forward towards our understanding of the complexity of cancer. Indeed, the vogue is an integrative, systems biology approach[43,96] where results from aCGH, gene-expression analysis and functional assays are combined to develop models that facilitate the understanding of complex biological systems such as cancer, and can serve as a basis for hypothesis generation, testing, and validation. This requires a huge emphasis on sound bioinformatics support and there is an urgent need for concurrent development of expertise in this field if this nascent technology is to become established as a reliable and user-friendly assay. This can only happen in a timely fashion with the development of multi-institutional, pan-national, or international collaborative efforts of data sharing and validation, and bioinformatic tools development, such as the Cancer Biomedical Informatics Grid (caBIG™) initiative pioneered by National Cancer Institute (http://cabig.cancer.gov). In addition, we should strive for the development of guidelines for aCGH studies similar to but with greater emphasis on study design and data analysis than the minimum information about a microarray experiment (MIAME) guidelines (http://www.mged.org/Workgroups/MIAME/MIAMEchecklist_cgh.doc) and the implementation of more user friendly repository websites with meta-analysis functions, similar to websites that allow for meta-analysis of expression array data (eg Oncomine—http://www.oncomine.org). Ultimately, the successful application and technological development of aCGH approaches in cancer research can only be achieved with a full understanding of its current limitations (Table 6) before careful development of strategies to circumvent them. One might even say that designing an aCGH study is akin to laying the foundations of a building: you only get it right at the end if you got it right at the beginning.

ACKNOWLEDGEMENTS

We thank Dr C Westbury, Mrs M Iravani, and the other members of the Molecular Pathology Laboratory at The Breakthrough Breast Cancer Research Centre for their insightful comments in the preparation of this manuscript. RN, ML, and JSR-F are funded by Breakthrough Breast Cancer. DT is funded in part by a grant from Cancer Research UK.

CONFLICT OF INTEREST

No authors have any conflict of interest with regards to the information on commercial or potentially commercial products and devices.

1. Kallioniemi A, Kallioniemi OP, Sudar D, *et al*. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992;258:818–821.
2. Oostlander AE, Meijer GA, Ylstra B. Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet 2004;66:488–495.
3. Ylstra B, van den Ijssel P, Carvalho B, *et al*. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). Nucleic Acids Res 2006;34:445–450.
4. Jong K, Marchiori E, van der Vaart A, *et al*. Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors. Oncogene 2007;26:1499–1506.
5. Bergamaschi A, Kim YH, Wang P, *et al*. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer 2006;45:1033–1040.
6. Tanami H, Tsuda H, Okabe S, *et al*. Involvement of cyclin D3 in liver metastasis of colorectal cancer, revealed by genome-wide copy-number analysis. Lab Invest 2005;85:1118–1129.
7. Reis-Filho JS, Simpson PT, Gale T, *et al*. The molecular genetics of breast cancer: the contribution of comparative genomic hybridization. Pathol Res Pract 2005;201:713–725.
8. Reis-Filho JS, Simpson PT, Jones C, *et al*. Pleomorphic lobular carcinoma of the breast: role of comprehensive molecular pathology in characterization of an entity. J Pathol 2005;207:1–13.
9. Reis-Filho JS, Simpson PT, Turner NC, *et al*. FGFR1 emerges as a potential therapeutic target for lobular breast carcinomas. Clin Cancer Res 2006;12:6652–6662.
10. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet 2005;37(Suppl):S11–S17.
11. Simon R, Radmacher MD, Dobbin K, *et al*. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–18.

12. Lockwood WW, Chari R, Chi B, *et al*. Recent advances in array comparative genomic hybridization technologies and their applications in human genetics. Eur J Hum Genet 2006;14: 139–148.
13. Fan JB, Chee MS, Gunderson KL. Highly parallel genomic assays. Nat Rev Genet 2006;7:632–644.
14. van Beers EH, Joosse SA, Ligtenberg MJ, *et al*. A multiplex PCR predictor for aCGH success of FFPE samples. Br J Cancer 2006;94: 333–337.
15. Reis-Filho JS, Pinheiro C, Lambros MB, *et al*. EGFR amplification and lack of activating mutations in metaplastic breast carcinomas. J Pathol 2006;209:445–453.
16. Baldwin C, Garnis C, Zhang L, *et al*. Multiple microalterations detected at high frequency in oral cancer. Cancer Res 2005;65: 7561–7567.
17. Nessling M, Richter K, Schwaenen C, *et al*. Candidate genes in breast cancer revealed by microarray-based comparative genomic hybridization of archived tissue. Cancer Res 2005;65:439–447.
18. van Dekken H, Paris PL, Albertson DG, *et al*. Evaluation of genetic patterns in different tumor areas of intermediate-grade prostatic adenocarcinomas by high-resolution genomic array analysis. Genes Chromosomes Cancer 2004;39:249–256.
19. Harvell JD, Kohler S, Zhu S, *et al*. High-resolution array-based comparative genomic hybridization for distinguishing paraffin-embedded Spitz nevi and melanomas. Diagn Mol Pathol 2004;13: 22–25.
20. Linn SC, West RB, Pollack JR, *et al*. Gene expression patterns and gene copy number changes in dermatofibrosarcoma protuberans. Am J Pathol 2003;163:2383–2395.
21. Thompson ER, Herbert SC, Forrest SM, *et al*. Whole genome SNP arrays using DNA derived from formalin-fixed, paraffin-embedded ovarian tumor tissue. Hum Mutat 2005;26:384–389.
22. Oosting J, Lips EH, van Eijk R, *et al*. High-resolution copy number analysis of paraffin-embedded archival tissue using SNP BeadArrays. Genome Res 2007;17:368–376.
23. Weiss MM, Hermsen MA, Meijer GA, *et al*. Comparative genomic hybridization. Mol Pathol 1999;52:243–251.
24. Arriola E, Lambros MB, Jones C, *et al*. Evaluation of Phi29-based whole-genome amplification for microarray-based comparative genomic hybridization. Lab Invest 2007;87:75–83.
25. Raponi M, Zhang Y, Yu J, *et al*. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006;66:7466–7472.
26. Pirker C, Raidl M, Steiner E, *et al*. Whole genome amplification for CGH analysis: linker-adapter PCR as the method of choice for difficult and limited samples. Cytometry A 2004;61:26–34.
27. Tanabe C, Aoyagi K, Sakiyama T, *et al*. Evaluation of a whole-genome amplification method based on adaptor-ligation PCR of randomly sheared genomic DNA. Genes Chromosomes Cancer 2003; 38:168–176.
28. Fiegler H, Geigl JB, Langer S, *et al*. High resolution array-CGH analysis of single cells. Nucleic Acids Res 2007;35:e15.
29. Aviel-Ronen S, Qi Zhu C, Coe BP, *et al*. Large fragment Bst DNA polymerase for whole genome amplification of DNA from formalin-fixed paraffin-embedded tissues. BMC Genomics 2006; 7:312.
30. Little SE, Vuononvirta R, Reis-Filho JS, *et al*. Array CGH using whole genome amplification of fresh-frozen and formalin-fixed, paraffin-embedded tumor DNA. Genomics 2006;87:298–306.
31. Lovmar L, Syvanen AC. Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. Hum Mutat 2006;27:603–614.
32. Lage JM, Leamon JH, Pejovic T, *et al*. Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. Genome Res 2003;13: 294–307.
33. Fredman D, White SJ, Potter S, *et al*. Complex SNP-related sequence variation in segmental genome duplications. Nat Genet 2004;36: 861–866.
34. Bredel M, Bredel C, Juric D, *et al*. Amplification of whole tumor genomes and gene-by-gene mapping of genomic aberrations from limited sources of fresh-frozen and paraffin-embedded DNA. J Mol Diagn 2005;7:171–182.
35. Redon R, Ishikawa S, Fitch KR, *et al*. Global variation in copy number in the human genome. Nature 2006;444:444–454.
36. Iafrate AJ, Feuk L, Rivera MN, *et al*. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949–951.
37. Sebat J, Lakshmi B, Troge J, *et al*. Large-scale copy number polymorphism in the human genome. Science 2004;305:525–528.
38. Freeman JL, Perry GH, Feuk L, *et al*. Copy number variation: new insights in genome diversity. Genome Res 2006;16:949–961.
39. McCarroll SA, Hadnott TN, Perry GH, *et al*. Common deletion polymorphisms in the human genome. Nat Genet 2006;38:86–92.
40. Khalique L, Ayhan A, Weale ME, *et al*. Genetic intra-tumor heterogeneity in epithelial ovarian cancer and its implications for molecular diagnosis of tumors. J Pathol 2007;211:286–295.
41. Pierga JY, Reis-Filho JS, Cleator SJ, *et al*. Microarray-based comparative genomic hybridization of breast cancer patients receiving neoadjuvant chemotherapy. Br J Cancer 2007;96:341–351.
42. Crawford JM, Tykocinski ML. Pathology as the enabler of human research. Lab Invest 2005;85:1058–1064.
43. Hornberg JJ, Bruggeman FJ, Westerhoff HV, *et al*. Cancer: a systems biology disease. Biosystems 2006;83:81–90.
44. Loo LW, Grove DI, Williams EM, *et al*. Array comparative genomic hybridization analysis of genomic alterations in breast cancer subtypes. Cancer Res 2004;64:8541–8549.
45. Stange DE, Radlwimmer B, Schubert F, *et al*. High-resolution genomic profiling reveals association of chromosomal aberrations on 1q and 16p with histologic and genetic subgroups of invasive breast cancer. Clin Cancer Res 2006;12:345–352.
46. Chung CH, Parker JS, Karaca G, *et al*. Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell 2004;5:489–500.
47. Perou CM, Sorlie T, Eisen MB, *et al*. Molecular portraits of human breast tumors. Nature 2000;406:747–752.
48. van de Vijver MJ, He YD, van't Veer LJ, *et al*. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
49. Ioannidis JP. Microarrays and molecular research: noise discovery? Lancet 2005;365:454–455.
50. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA 2006;103:5923–5928.
51. Simpson PT, Reis-Filho JS, Gale T, *et al*. Molecular evolution of breast cancer. J Pathol 2005;205:248–254.
52. Buerger H, Mommers EC, Littmann R, *et al*. Ductal invasive G2 and G3 carcinomas of the breast are the end stages of at least two different lines of genetic evolution. J Pathol 2001;194: 165–170.
53. Buerger H, Otterbach F, Simon R, *et al*. Comparative genomic hybridization of ductal carcinoma *in situ* of the breast-evidence of multiple genetic pathways. J Pathol 1999;187:396–402.
54. Buerger H, Otterbach F, Simon R, *et al*. Different genetic pathways in the evolution of invasive breast cancer are associated with distinct morphological subtypes. J Pathol 1999;189:521–526.
55. Roylance R, Gorman P, Harris W, *et al*. Comparative genomic hybridization of breast tumors stratified by histological grade reveals new insights into the biological progression of breast cancer. Cancer Res 1999;59:1433–1436.
56. Feber A, Clark J, Goodwin G, *et al*. Amplification and overexpression of E2F3 in human bladder cancer. Oncogene 2004;23:1627–1630.
57. Cheng KW, Lahad JP, Kuo WL, *et al*. The RAB25 small GTPase determines aggressiveness of ovarian and breast cancers. Nat Med 2004;10:1251–1256.
58. Natrajan R, Reis-Filho JS, Little SE, *et al*. Blastemal expression of type I insulin-like growth factor receptor in Wilms' tumors is driven by increased copy number and correlates with relapse. Cancer Res 2006;66:11148–11155.
59. Chin K, DeVries S, Fridlyand J, *et al*. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 2006;10:529–541.
60. Neve RM, Chin K, Fridlyand J, *et al*. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 2006;10:515–527.
61. Weinstein IB. Cancer. Addiction to oncogenes—the Achilles heal of cancer. Science 2002;297:63–64.

62. Weinstein IB, Joe AK. Mechanisms of disease: oncogene addiction—a rationale for molecular targeting in cancer therapy. Nat Clin Pract Oncol 2006;3:448–457.
63. Gao X, LeProust E, Zhang H, *et al*. A flexible light-directed DNA chip synthesis gated by deprotection using solution photogenerated acids. Nucleic Acids Res 2001;29:4744–4750.
64. Pollack JR, Perou CM, Alizadeh AA, *et al*. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 1999;23: 41–46.
65. Johnson NA, Hamoudi RA, Ichimura K, *et al*. Application of array CGH on archival formalin-fixed paraffin-embedded tissues including small numbers of microdissected cells. Lab Invest 2006;86:968–978.
66. Vincent-Salomon A, Gruel N, Lucchesi C, *et al*. Identification of typical medullary breast carcinoma as a genomic sub-group of basal-like carcinomas, a heterogeneous new molecular entity. Breast Cancer Res 2007;9:R24.
67. Lambros MB, Simpson PT, Jones C, *et al*. Unlocking pathology archives for molecular genetic studies: a reliable method to generate probes for chromogenic and fluorescent *in situ* hybridization. Lab Invest 2006;86:398–408.
68. Di Palma S, Lambros MB, Savage K, *et al*. Oncocytic change in pleomorphic adenoma: molecular evidence in support of an origin in neoplastic cells. J Clin Pathol 2006 [E-pub ahead of print; doi:10.1136/jcp.2005.031369].
69. Lucito R, Healy J, Alexander J, *et al*. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res 2003;13:2291–2305.
70. Hicks J, Krasnitz A, Lakshmi B, *et al*. Novel patterns of genome rearrangement and their association with survival in breast cancer. Genome Res 2006;16:1465–1479.
71. Hupe P, Stransky N, Thiery JP, *et al*. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics 2004;20:3413–3422.
72. Shah SP, Xuan X, DeLeeuw RJ, *et al*. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 2006;22:e431–e439.
73. Yang YH, Dudoit S, Luu P, *et al*. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30:e15.
74. van Beers EH, Nederlof PM. Array-CGH and breast cancer. Breast Cancer Res 2006;8:210.
75. Ji H, Kumm J, Zhang M, *et al*. Molecular inversion probe analysis of gene copy alterations reveals distinct categories of colorectal carcinoma. Cancer Res 2006;66:7910–7919.
76. Wang Y, Moorhead M, Karlin-Neumann G, *et al*. Allele quantification using molecular inversion probes (MIP). Nucleic Acids Res 2005;33:e183.
77. Greshock J, Nathanson K, Martin AM, *et al*. Cancer cell lines as genetic models of their parent histology: analyses based on array comparative genomic hybridization. Cancer Res 2007;67:3594–3600.
78. Courjal F, Cuny M, Simony-Lafontaine J, *et al*. Mapping of DNA amplifications at 15 chromosomal localizations in 1875 breast tumors: definition of phenotypic groups. Cancer Res 1997;57: 4360–4367.
79. Gelsi-Boyer V, Orsetti B, Cervera N, *et al*. Comprehensive profiling of 8p11–12 amplification in breast cancer. Mol Cancer Res 2005;3: 655–667.
80. Ginestier C, Cervera N, Finetti P, *et al*. Prognosis and gene expression profiling of 20q13-amplified breast cancers. Clin Cancer Res 2006;12:4533–4544.
81. Natrajan R, Williams RD, Grigoriadis A, *et al*. Delineation of a 1 Mb breakpoint region at 1p13 in Wilms tumors by fine-tiling oligonucleotide array CGH. Genes Chromosomes Cancer 2007;46: 607–615.
82. Gribble SM, Fiegler H, Burford DC, *et al*. Applications of combined DNA microarray and chromosome sorting technologies. Chromosome Res 2004;12:35–43.
83. Gribble SM, Kalaitzopoulos D, Burford DC, *et al*. Ultra-high resolution array painting facilitates breakpoint sequencing. J Med Genet 2007;44:51–58.
84. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 2005;23:7332–7341.
85. Pawitan Y, Murthy KR, Michiels S, *et al*. Bias in the estimation of false discovery rate in microarray studies. Bioinformatics 2005;21: 3865–3872.
86. Paris PL, Andaya A, Fridlyand J, *et al*. Whole genome scanning identifies genotypes associated with recurrence and metastasis in prostate tumors. Hum Mol Genet 2004;13:1303–1313.
87. Colella S, Yau C, Taylor JM, *et al*. QuantiSNP: an objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35: 2013–2025.
88. Buckley PG, Mantripragada KK, Piotrowski A, *et al*. Copy-number polymorphisms: mining the tip of an iceberg. Trends Genet 2005;21:315–317.
89. Ng G, Huang J, Roberts I, *et al*. Defining ploidy-specific thresholds in array comparative genomic hybridization to improve the sensitivity of detection of single copy alterations in cell lines. J Mol Diagn 2006;8:449–458.
90. Lambros MB, Fiegler H, Jones A, *et al*. Analysis of ovarian cancer cell lines using array-based comparative genomic hybridization. J Pathol 2005;205:29–40.
91. Natrajan R, Little SE, Reis-Filho JS, *et al*. Amplification and overexpression of CACNA1E correlates with relapse in favorable histology Wilms' tumors. Clin Cancer Res 2006;12:7284–7293.
92. Natrajan R, Williams RD, Hing SN, *et al*. Array CGH profiling of favourable histology Wilms tumors reveals novel gains and losses associated with relapse. J Pathol 2006;210:49–58.
93. Ginzinger DG. Gene quantification using real-time quantitative PCR: an emerging technology hits the mainstream. Exp Hematol 2002;30: 503–512.
94. Greenbaum D, Jansen R, Gerstein M. Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. Bioinformatics 2002;18:585–596.
95. Reif DM, White BC, Moore JH. Integrated analysis of genetic, genomic and proteomic data. Expert Rev Proteomics 2004;1: 67–75.
96. Bosl WJ. Systems biology by the rules: hybrid intelligent systems for pathway modeling and discovery. BMC Syst Biol 2007; 1:13.
97. Rivera MN, Kim WJ, Wells J, *et al*. An X chromosome gene, WTX, is commonly inactivated in Wilms tumor. Science 2007;315: 642–645.