

# Characterizing the physical genome

Jonathan R. Pollack<sup>1</sup> & Vishwanath R. Iyer<sup>2</sup>

doi:10.1038/ng1035

**The genome of an organism is a dynamic physical entity, comprising genomic DNA bound to many different proteins and organized into chromosomes. A thorough characterization of the physical genome is relevant to our understanding of processes such as the regulation of gene expression, DNA replication and repair, recombination, chromosome segregation, epigenetic inheritance and genomic instability. Methods based on microarrays are beginning to provide a detailed picture of this physical genome, and they complement the genome-wide studies of mRNA expression profiling that have previously been so successful.**

We live in an age in which whole-genome sequences are commonplace on our hard drives and on other storage media and servers throughout the Internet. It is easy to think of a genome in purely digital terms, that is, as an abstract string of four nucleotides whose secrets can be uncovered largely through the use of software algorithms. But in the nucleus of a cell, the genome has a larger meaning—it is a discrete physical entity, organized into chromosomes comprising genomic DNA bound to proteins in a systematic way. This structure is dynamic and participates in many different fundamental processes such as transcription, DNA replication and repair, recombination and chromosome segregation. Genomic DNA may be modified specifically by processes such as methylation, and these modifications can further define functional features of the genome. The genome can be unstable, and this instability underlies the development of genetic lesions that can lead to cancer.

How do transcription factors, DNA modifications and chromatin structure function in concert to specify the expression of genomic information? How does the genome undergo replication and recombination? How do amplifications and deletions of genomic DNA contribute to different pathological states of cells? Knowing the details of several aspects of genomic structure is clearly important for understanding the functional behavior of the genome. The phenomenal popularity of DNA microarrays has been fueled by their ability to determine global gene expression profiles of RNA. Recently, however, exciting new approaches using DNA microarrays are beginning to provide us with high-resolution views of the physical genome and to clarify how it participates in diverse cellular phenomena. Here we review the use of DNA microarrays in determining the binding-distribution of proteins on the genome, in mapping chemical modifications to chromatin and DNA, in studying DNA replication and repair, and in determining differences in DNA copy number by comparative genomic hybridization to microarrays (array CGH). We also discuss how these and other applications of DNA microarrays may further our understanding of genome dynamics in the future.

## Binding distribution of transcription factors

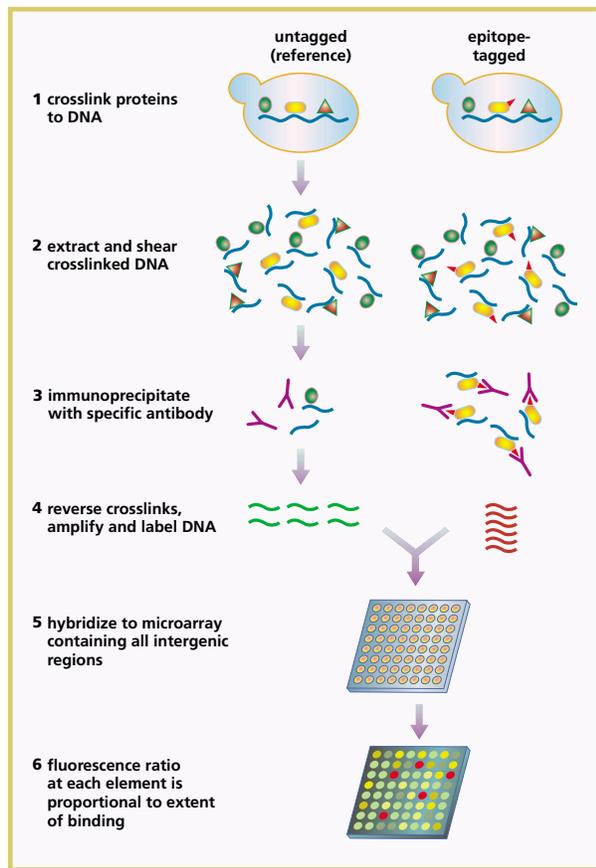
Transcription factors are important links in the transduction of signals from extracellular and intracellular stimuli to several bio-

logical processes, including cell growth, regulation of the cell-division cycle, embryonic development, differentiation, apoptosis and response to environmental stresses. Thus, a genome is brought to life by the interactions of transcription factors with DNA. The use of DNA microarrays has begun to provide insights into the genome-wide transcriptional programs of cells; in conjunction with genetic and molecular perturbations of individual transcription factors, such studies permit a dissection of the roles of individual pathways and regulatory molecules to the elaborate molecular circuitry of cells. To properly understand the roles of individual transcription factors in mediating global gene expression programs, however, it is necessary to determine the *in vivo* physical interactions of transcription factors with their chromosomal targets on the genome.

The strategy for doing this on a genome-wide scale, a version of which is shown in Fig. 1, is based on combining the technique of chromatin immunoprecipitation (ChIP) with hybridization to a DNA microarray of promoter sequences. Several variations are possible, and obviously this strategy can be applied to study the interaction of any protein that interacts with genomic DNA, not only transcription factors. In ChIP, formaldehyde is first used to crosslink proteins to their cognate binding sites on DNA; the protein of interest is then immunoprecipitated using a specific antibody, which also recovers the DNA to which the protein is bound. After reversal of the crosslinks, amplification of the DNA and fluorescent labeling, all of the enriched DNA fragments are identified simultaneously by hybridization to a microarray of promoter sequences in conjunction with an appropriate reference probe labeled with a different fluorophore. Thus, the ratio of fluorescence intensities at an element represents a measure of binding of the protein to that locus. The approach has been used most successfully in yeast, where the compact nature of the genome means that whole intergenic regions containing *cis*-regulatory information can be amplified easily from the whole genome and represented on a microarray. This strategy has been applied to identify the genomic targets of many transcription factors and other proteins that interact with the genome<sup>1–10</sup>.

Determining the genome-wide binding distribution of transcription factors is valuable for many reasons. First, it is a crucial piece of information for determining the direct downstream tar-

<sup>1</sup>Department of Pathology, Stanford University School of Medicine, CCSR Building, Room 3245A, 269 Campus Drive, Stanford, California 94305-5176, USA (e-mail: jpollack1@stanford.edu). <sup>2</sup>Institute for Cellular and Molecular Biology and Section of Molecular Genetics and Microbiology, University of Texas at Austin, 1 University Station A4800, Austin, Texas 78712-0159, USA (e-mail: vishy@mail.utexas.edu).



Bob Crimi

**Fig. 1** Strategy for mapping genome-wide DNA-protein interactions using intergenic microarrays. In the approach shown, DNA recovered after immunoprecipitation (IP) by an antibody against an epitope tag fused to the protein of interest is labeled and hybridized to a microarray, in conjunction with a reference probe from a parallel immunoprecipitation done in the absence of a tag. Alternative approaches include using an antibody against the native protein for the immunoprecipitation, using non-immunoprecipitated genomic DNA as the reference probe, using immunoprecipitated DNA derived from a strain deleted for the gene of interest as the reference probe (for non-essential genes) and directly comparing the binding distribution of the same factor under two different growth conditions by labeling immunoprecipitated DNA from the two different conditions with different fluorophores.

gets of transcription factors *in vivo*. Attempting to identify targets by mRNA expression analysis in strains lacking or overexpressing the gene of interest may be confounded by indirect secondary or pleiotropic effects. In the case of genes that are essential for viability, deletions are not possible and conditional alleles are not always available or may be complicated to interpret. One of the earliest successful studies to use this approach identified the downstream targets of Esa1, the only histone acetyltransferase that is essential for viability in yeast. Before this genomic binding analysis on microarrays, nothing was known about the *in vivo* function of Esa1; the genomic binding distribution provided the first evidence of its function in regulating ribosomal protein gene transcription<sup>1</sup>.

Second, the genomic binding distribution of a protein provides a rich source of data on the specificity of *in vivo* DNA-protein interactions. It is possible to compare the promoter sequences to which a given protein binds *in vivo* and to identify the sequence determinants and features of chromatin structure that determine *in vivo* binding specificity. For the cell-cycle regulatory factors SBF and MBF and the multifunctional Rap1 protein, consensus binding sites have been identified by applying pattern discovery algorithms to the intergenic sequences that are bound by the proteins<sup>2,5</sup>. These proteins bind preferentially to intergenic promoter sequences even though their consensus binding sites are frequently present in coding sequences; this suggests that other factors, such as chromatin structure, influence the interactions of a protein with its chromosomal targets. Even for a protein that does not interact directly with DNA, this type of sequence analysis can detect the presence of consensus sites for other proteins that can recruit it to the promoter<sup>7</sup>.

Third, genome-wide maps of the binding distribution of transcription factors and other proteins can identify details of global chromosomal features, such as clustering of binding in certain

chromosomal regions or a higher order periodicity during interaction with the genome. Such interaction maps have detected, for example, the clustering of Rap1 and the silencing proteins Sir2, Sir3 and Sir4 at subtelomeric loci<sup>5</sup>. The microarray format allows one to observe changes in these global chromosomal features in response to different perturbations.

Last, data on the binding targets of a transcription factor, when combined with genome-wide analysis of mRNA expression, provides a more reliable means of deciphering the hierarchies and networks of regulatory interactions that underlie most complex cellular responses. Genes whose promoters are bound and regulated directly by a transcription factor may themselves encode transcriptional activators or repressors, thus making it possible to have cascading responses as the outcome of an initial signal. A notable example is the serial transcriptional regulation of genes expressed periodically during the yeast cell cycle. Data on the genomic binding distribution for several transcription factors that regulate the yeast cell cycle<sup>4</sup> have been combined with previously determined mRNA expression profiling data over the course of the cell cycle<sup>11</sup>. This analysis showed that activators functioning during the G1/S transition (SBF and MBF) bind to the promoters of activators that function during G2/M. The G2/M activators in turn bind and activate the promoters of M/G1 genes. The M/G1 transcription factors then cause the activation of the G1/S transcription factors in the next round of the cell cycle, thus closing this loop of interactions<sup>4</sup>.

### Histone modifications and chromatin remodeling

Gene expression is strongly influenced by the modification of histones in nucleosomes that package and organize the genome. These post-translational modifications, including acetylation, methylation, phosphorylation and ubiquitination, are thought to be highly specific, and certain types of modification can influence one another<sup>12,13</sup>. Microarrays provide a way to map the location across the genome of such modifications, the protein factors that carry out these modifications, and the proteins that participate in other, less-understood aspects of chromatin remodeling.

The specificity of histone acetylation has been used to map the genomic distribution of acetylation loci<sup>8</sup>. In this study, immunoprecipitations were carried out using antibodies against specific acetylated lysine residues on histones H3 and H4 in yeast strains deleted for each of the different histone deacetylases, including *RPD3*, *HDA1* and several others. Although this is an indirect way to map the sites of action of individual deacetylases, this information was complemented by independent direct measurements of the binding distribution of Rpd3 (ref. 7). There are numerous differences between the directly measured binding distribution of Rpd3 and the map of *RPD3*-dependent histone deacetylation, probably owing to redundancy among the deacetylases and to possible indirect secondary effects in the latter type of experiment. This underscores the need for integrating information from different genomic approaches in order to build a clear and detailed picture of chromatin modifications.

A modification to the ChIP procedure, namely the use (in addition to formaldehyde) of the protein crosslinker dimethyl adipimate, was found to significantly enhance the ability to detect the interaction of Rpd3 with DNA<sup>7</sup>. This bodes well for future binding studies of proteins and complexes that do not interact with DNA directly, but are instead recruited to promoters by other sequence-specific DNA-binding proteins. The same general approach has been used to map the interactions of the RSC chromatin remodeling complex with the genome and its redistribution under stress conditions<sup>9,10</sup>.

An alternative microarray-based approach for mapping the interactions of chromatin proteins with their target loci uses the *Escherichia coli* Dam methylase<sup>14</sup>. In this strategy, a fusion of the protein of interest and *E. coli* Dam methylase is expressed in cells, resulting in the targeting of Dam methylase to the sites of interaction of the DNA-binding protein of interest. Consequently, DNA near these interaction loci becomes specifically methylated. These methylation sites, which thus mark the interaction loci of the protein of interest, can be identified by digesting the genomic DNA with a methylation-specific restriction endonuclease, isolating the digested fragments by size fractionation, and hybridizing them to a DNA microarray<sup>14</sup>. A distinguishing feature of this strategy is that it does not require immunoprecipitation with an antibody, which may be advantageous in higher eukaryotes.

#### DNA replication and recombination

DNA replication has been actively studied ever since the classic experiments of Meselson and Stahl<sup>15</sup> demonstrated its semiconservative nature. Although functional replication origins have been identified in yeast, the mechanism by which they are chosen for initiation is not well understood, and the timing of initiation and the rate of progression of replication forks are not well known.

Microarray hybridization has now been used to identify the distribution and timing of origins in the yeast genome. In this contemporary version of Meselson and Stahl's experiment, replicated and unreplicated DNA are distinguished by differential labeling with the heavy isotopes <sup>13</sup>C and <sup>15</sup>N, and replicating DNA is mapped by hybridization to a high-density oligonucleotide array<sup>16</sup>. Sampling the replicating DNA at different time intervals after synchronized cells are released into S phase allows the timing of replication initiation at different origins to be determined. This procedure was used to map 332 functional origins that fire at different times in S phase<sup>16</sup>. It also showed that the rate of replication fork progression varies across the genome.

An alternative approach to mapping genomic origins of replication has used the 'ChIP-on-microarray' strategy described above for transcription factors to map the binding distribution of the ORC and MCM proteins, which recognize origins and initiate replication<sup>4</sup>. This study identified 429 origins, most of which overlapped with the former study, and showed that several origins are associated with repetitive DNA elements in the yeast genome.

In *E. coli*, which has a circular genome and a unique origin, it has been possible to monitor the progression of DNA replication forks at single-gene resolution simply by simultaneously hybridizing replicating DNA labeled with one fluorophore and unreplicated DNA labeled with another fluorophore to a microarray of all *E. coli* genes<sup>17</sup>. This approach has elucidated the roles of different topoisomerases in facilitating replication.

Other aspects of chromosome dynamics such as recombination can also be studied on a global scale with microarrays. In one approach, DNA attached to the yeast Spo11 protein, which marks double-strand breaks that initiate recombination, was enriched and hybridized to genome microarrays to identify recombination hotspots and coldspots<sup>18</sup>. Although this

approach does not involve formaldehyde crosslinking, it is likely that the ChIP strategy will be useful in mapping the genomic interactions of other proteins involved in chromosome condensation, recombination and segregation.

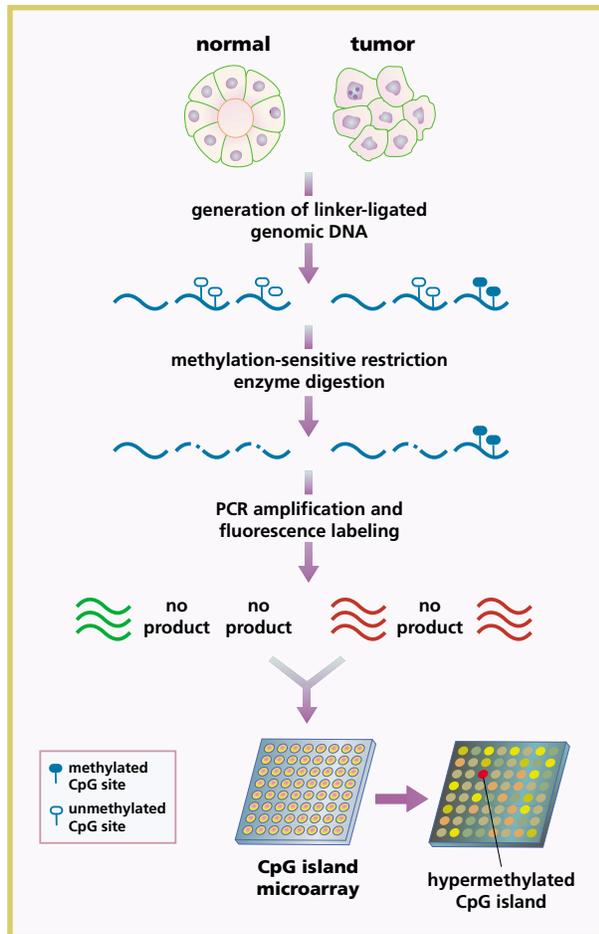
#### DNA methylation

In many eukaryotes, DNA methylation is important for regulating gene expression. Mammalian genomic DNA is methylated at cytosines in most CpG dinucleotides except for the (G+C)-rich 'CpG islands' that are associated with functional gene promoters<sup>19</sup>. DNA methylation is a potent inhibitor of gene expression and functions, in part through the recruitment of proteins that bind methyl-CpG and histone deacetylase (HDAC), to create a transcriptionally repressive chromatin state. The pattern of DNA methylation established during development and differentiation is maintained by DNA cytosine-5-methyltransferase (Dnmt1), which functions on the hemimethylated products of semiconservative DNA replication. Aberrant DNA methylation frequently contributes to the development of cancer, in which focal hypermethylation of CpG islands transcriptionally silences tumor suppressor genes<sup>20</sup>. It is therefore useful to characterize aberrant patterns of DNA methylation and to identify methylation-silenced genes and pathways.

Two general approaches have used DNA microarrays to identify methylation-silenced genes in cancer. The first approach involves using microarrays to identify upregulated transcripts after demethylation, which is accomplished either by genetic inactivation<sup>21</sup> or by pharmacological inhibition of Dnmt1, most often using 5-aza-2' deoxycytidine (DAC)<sup>22-24</sup>. Limitations of this approach include its requirement for living or dividing cells, and the difficulty in discriminating between direct and indirect (that is, downstream) effects of demethylation on gene expression. Indeed, the observation that many DAC-induced genes have no associated CpG islands<sup>23</sup> suggests that considerable downstream effects are involved. The common finding of DAC-induced interferon-responsive genes<sup>22,23</sup> may also be explained as an indirect consequence of the reactivation of methylation-silenced endogenous retroviruses. Those genes whose expression is inactivated completely in untreated cells, and whose expression is upregulated with low doses of DAC followed by the HDAC inhibitor trichostatin A, but not by trichostatin A alone, may preferentially reflect genes that are directly silenced by promoter hypermethylation<sup>24</sup>. Information from other sources on gene function and chromosomal position will help to evaluate candidate methylation-silenced tumor suppressor genes.

A second general approach involves using DNA microarrays to identify methylated genomic DNA directly<sup>25,26</sup>. A successful version of this approach, termed differential methylation hybridization (Fig. 2)<sup>25,27</sup>, uses specialized DNA microarrays comprising cloned CpG islands. Tumor DNA and normal DNA are each digested with a methylation-sensitive restriction enzyme, and polymerase chain reaction (PCR) amplicons derived from each sample are then hybridized to the CpG island array. Array elements with a stronger hybridization signal in the tumor than in the normal sample represent aberrantly hypermethylated CpG islands, which are protected from methylation-sensitive restriction enzyme cleavage and therefore amplified by PCR in the tumor sample.

This approach has detected patterns of CpG island hypermethylation that correlate with histological grade<sup>28</sup> and hormone receptor status<sup>27</sup> in breast tumors, and with progression-free survival in late-stage ovarian cancer<sup>29</sup>. Limitations of this approach include a selection bias for CpG islands containing the methylation-sensitive restriction site, and the difficulty in associating hypermethylated CpG islands, once identified, with nearby genes. A



Bob Crimi

**Fig. 2** Differential methylation hybridization. Tumor and normal linker-ligated genomic DNA are each digested with a methylation-sensitive restriction enzyme, and PCR amplicons derived from each sample are labeled and hybridized to an array comprising clones of CpG islands. Array elements showing a greater hybridization signal in the tumor than in the normal sample represent CpG islands that are aberrantly hypermethylated in the tumor sample, protected from methylation-sensitive restriction enzyme cleavage and therefore amplified by PCR (see ref. 27 for details). This figure was adapted from ref. 25.

promising version of this approach uses arrays comprising transcribed CpG islands located, for example, in the 5' untranslated region and first exon of genes<sup>30</sup>. Such arrays permit a parallel analysis of CpG island methylation and DAC-induced gene expression, assuming that sufficiently processive reverse-transcription-based labeling of mRNA can be achieved.

Oligonucleotide arrays have also shown promise in directly discriminating methylated from unmethylated CpG dinucleotides<sup>31,32</sup>. Bisulphite modification of genomic DNA converts unmethylated cytosines, but not methylated cytosines, to uracil. Regions of interest are then amplified by multiplex PCR and hybridized to an oligonucleotide array with elements designed to discriminate bisulphite-converted TpG dinucleotides from methylation-protected CpG dinucleotides. This approach has been used to identify CpG methylation patterns that can discriminate between normal and tumor tissues, and among different types of tumor<sup>32</sup>.

### DNA copy number aberrations and array CGH

Genomic DNA itself can behave dynamically. The genomic instabilities that underlie cancer create gains and losses of whole chromosomes or more localized subchromosomal regions. The resultant increased expression of oncogenes and decreased expression of tumor suppressor genes provide a selective growth advantage to tumor cells that retain such aberrations. Comparative genomic hybridization<sup>33</sup> was developed to rapidly identify and map genomic DNA copy number changes across the genome. In the original CGH procedure, test (for example, tumor) and reference (for example, normal) genomic DNAs are labeled with different fluorophores and cohybridized to a normal

metaphase chromosome spread. The ratios of test fluorescence to reference fluorescence along the chromosomes provide a cytogenetic representation of alterations in genomic DNA copy number in the test sample relative to the reference sample. Although CGH has been used to catalogue alterations in DNA copy number in many tumors, its utility in locating the underlying oncogenes or tumor suppressor genes is limited ultimately by its cytogenetic mapping resolution.

Recently, DNA microarrays have provided an alternative, high-resolution platform for CGH<sup>34–36</sup>. In array CGH, differentially labeled test and reference genomic DNAs are cohybridized to a microarray containing DNA elements of known position in the human genome map (Fig. 3). For each element in the array, the fluorescence ratio represents the relative DNA copy number for its cognate genomic DNA in the test sample as compared with the reference sample. The mapping resolution of array CGH is at least an order of magnitude higher than that of chromosome CGH and is limited by only the number and the genomic distribution of arrayed elements.

In some respects, array CGH is more technically challenging than expression profiling. Genomic DNA is many times more complex than the mRNA representation of the typical mammalian cell. In array CGH, it is also essential to quantify accurately small ratio deviations from unity, such as the expected ratio of 0.75 for a single-copy loss of a tumor suppressor gene in a tumor that has a 50% contamination of normal cells. Different array CGH formats have been developed to meet these technical challenges. Arrays comprising either large genomic DNA clones (such as bacterial artificial chromosomes)<sup>35,37,38</sup> or representations of such clones made by amplifying with degenerate oligonucleotide primers by PCR (DOP-PCR)<sup>39,40</sup> provide high-precision measurements of single-copy gain or loss for individual clones.

cDNA microarrays offer a readily available alternative<sup>36,41</sup>, where averaging the fluorescence ratios for neighboring genes along the chromosome provides an accurate identification of single-copy changes. Because the same arrays can be used for both CGH and expression profiling, cDNA microarrays also facilitate the parallel analysis of DNA copy number and gene expression. Arrays comprising PCR-amplified, reduced-complexity genomic representations<sup>42,43</sup>, although not used widely as yet, should provide stronger hybridization signals and, because the test sample is also amplified by PCR, should permit the analysis of smaller quantities of genomic DNA. Finally, although oligonucleotide arrays have so far proved useful for CGH analyses of only low-complexity genomes<sup>44,45</sup>, when combined with multiplex PCR single nucleotide polymorphism arrays may be used to analyze tumor samples to detect loss of heterozygosity or allelic loss, which, like deletion, may mark the location of tumor suppressor genes<sup>46–48</sup>.

It is often desirable to evaluate specimens, such as biopsies or microdissected tissue samples, in which the quantity of genomic DNA is limited. Whereas standard array CGH protocols require 0.1–2 µg of input genomic DNA, the random primer labeling methods used in these protocols may provide sufficient DNA amplification to permit the analysis of as little as 3 ng of input genomic DNA<sup>37</sup>. For minuscule specimens, amplification by

**Fig. 3** Comparative genomic hybridization to microarrays (array CGH). **a**, Protocol for array CGH. Test DNA (tumor) and reference DNA (normal) are labeled with different fluorophores and cohybridized to a microarray containing DNA elements with a known position on the genomic map. Fluorescence ratios reflect the relative DNA copy number in the test sample as compared with the reference sample. Here, red indicates DNA amplification, and green indicates DNA deletion. Modified from ref. 36. **b**, Tumor DNA copy number profile generated by array CGH. Fluorescence ratio peaks identify regions of DNA amplification at high resolution. The sample shown here is chromosome 8 in the breast cancer cell line SKBR3. This figure was adapted from ref. 58.

DOP-PCR, or less-biased methods for efficient whole-genome amplification that do not rely on exponential amplification by PCR (see, for example, ref. 49), may also prove useful.

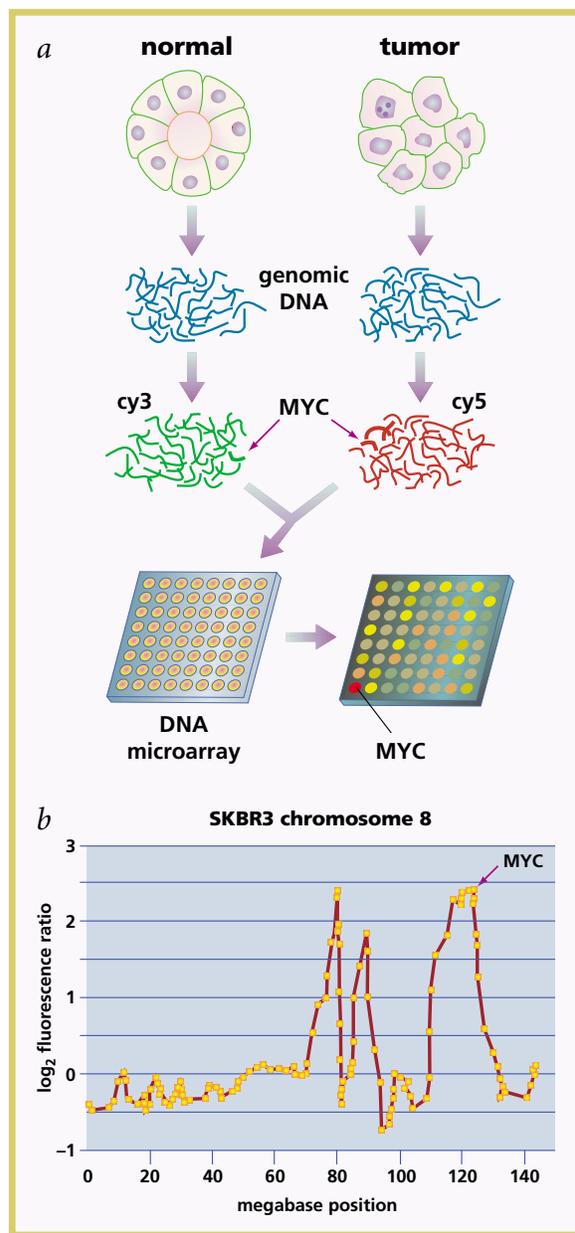
In studies of cancer, array CGH has been used to localize changes in DNA copy number that underlie the progression of mouse islet carcinoma<sup>39</sup>. Array CGH also has been useful in identifying patterns of genomic alteration that are diagnostic for subtypes of non-small cell lung cancer<sup>40</sup>, renal cell cancer<sup>50</sup> and liposarcoma<sup>51</sup>. This kind of cancer taxonomy, based on changes in DNA copy number, is particularly attractive because DNA is more stable than mRNA and is readily recovered from paraffin-embedded tissues. Array CGH has proved a facile method for detecting gene amplifications<sup>52–55</sup> and, in parallel with expression profiling, for identifying those genes whose increased expression is driven by DNA amplification in breast cancer<sup>56–58</sup> and gastric cancer<sup>59</sup>. Indeed, a parallel microarray analysis of gene copy number and expression has detected a significant impact of widespread alterations in DNA copy number on the transcriptional programs of breast tumors. This raises the possibility that global imbalances in gene expression may disrupt essential stoichiometric relationships in cell metabolism and physiology and contribute directly to tumor development or progression<sup>58</sup>.

Array CGH has also proved useful in investigations of constitutional genetic syndromes, including the high-resolution analysis of *NF2* gene deletions in individuals with neurofibromatosis type 2 (ref. 60) and the investigation of subtelomeric chromosome rearrangements in individuals with mental retardation or congenital anomalies<sup>61</sup>. Other applications of array CGH, which portend its broad utility, include physically mapping genes in somatic cell radiation hybrids<sup>62</sup>, tracing the progression of chromosomal replication forks in *E. coli*<sup>17</sup>, comparing the genomes of mycobacteria<sup>63</sup> and characterizing the genomic rearrangements that accompany experimental evolution in budding yeast (M. Dunham and D. Botstein, pers. comm.).

### Looking ahead

Mapping proteins onto the genome has been successful in yeast and will probably become a very useful adjunct to studies of gene expression profiling. Independent assessments of *in vivo* interactions by standard PCR assays for specific target loci generally corroborate the microarray results. But it is not clear that the overall approach is sufficiently quantitative; that is, the ratio of fluorescence intensities at a given microarray element may not reflect precisely the strength of the *in vivo* interaction at that locus. This limitation will need to be addressed in the future. As with all microarray techniques, the error and variability associated with each of the approaches described here need to be evaluated carefully.

Considerable effort is likely to be devoted to mapping the binding distribution of proteins on mammalian genomes. The approach of combining ChIP and microarray hybridization has been shown to work in mammalian cells in principle<sup>64,65</sup>, although the experiments, which sought to determine the binding targets of the E2F activator, assayed only a subset of poten-



tial promoter regions from the human genome. In one approach, CpG island microarrays were used, thereby enriching the representation of promoters on the array. Because these elements had not been characterized previously, binding loci for E2F were identified only after the arrayed clones that showed enrichment in the ChIP procedure were sequenced<sup>64</sup>. The other approach used a microarray of PCR fragments approximately 1 kilobasepairs in length that spanned the start site of transcription and encompassed most of the proximal promoters of about 1,500 selected genes, thus taking advantage of the fact that known E2F-binding sites are located near the transcription start site<sup>65</sup>.

Because intergenic regions in mammals are on the order of tens to hundreds of kilobases, it is currently not feasible to make comprehensive arrays covering all intergenic regions in human or mouse. This is currently one of the limitations of this approach in applications to large genomes. Computational approaches for identifying potential *cis*-regulatory elements in vast tracts of intergenic sequence, such as those based on the clustering of transcription factor binding sites<sup>66,67</sup> and on cross-

species sequence conservation in noncoding regions<sup>68</sup>, will undoubtedly facilitate the fabrication of genome-wide promoter-specific microarrays for complex genomes. For other completely sequenced genomes of model organisms, such as the fly, worm and numerous prokaryotes, the impediment to using similar approaches is smaller. For mapping other chromosomal features such as replication origins and timing, it is possible that existing microarrays of mapped cDNAs will suffice—much as they have for array CGH.

Many higher eukaryotes are also at a disadvantage in the combined ChIP and microarray approach owing to the relative difficulty, as compared with yeast, of generating exact chromosomal replacements of genes with fused epitope tags. One must then rely either on using antibodies against the native protein or on ectopically expressing epitope-tagged proteins. It is also unclear whether this approach can be used to map genomic interactions in cells in solid tissue, where a lot of interesting biology takes place. Nevertheless, the prospects seem bright for our ability to map proteins and other chromosomal features in mammalian genomes and to integrate these maps with the excellent graphical sequence annotation maps of the human and mouse genomes available on websites such as those of the University of California at Santa Cruz (<http://genome.ucsc.edu/cgi-bin/hgGateway>) and the National Center for Biotechnology Information ([http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum\\_srch](http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch)).

A formidable challenge ahead is to develop bioinformatics approaches to integrate different views of the genome, including, for example, protein localization, DNA modifications and copy number changes, with gene expression profiles to build comprehensive models for global gene regulation and other aspects of genome dynamics. Such modeling is essential for making the leap from our current, mostly cataloguing activities to global insights into cell function under normal conditions and in pathological states. For studies of cancer, integrating microarray data on DNA methylation and alterations in copy number with microarray data on gene expression will provide complementary views of the cancer cell. For example, the microarray identification of a chromosomal region showing hypermethylation, deletion, loss of heterozygosity and decreased expression across a series of tumors would provide strong presumptive evidence for the existence of an important tumor suppressor gene. Integrating varied views of the physical genome with gene expression data will also undoubtedly provide a basis for an improved molecular classification of cancer. Finally, the genomic alterations and aberrant DNA methylation observed in cancer are certain to disrupt normal patterns of chromatin and to alter the distribution of DNA-binding proteins. It will be interesting to collect and integrate ChIP and microarray data on such altered patterns of DNA-binding proteins including transcription factors (for example, the oncogenic fusion proteins observed in many leukemias). Through continued technical innovation and clever experimentation, microarrays are certain to provide an increasingly detailed understanding of dynamic genomes in the future.

1. Reid, J.L., Iyer, V.R., Brown, P.O. & Struhl, K. Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* **6**, 1297–1307 (2000).
2. Iyer, V.R. et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533–538 (2001).
3. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
4. Simon, I. et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697–708 (2001).
5. Lieb, J.D., Liu, X., Botstein, D. & Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein–DNA association. *Nature Genet.* **28**, 327–334 (2001).
6. Wyrick, J.J. et al. Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: high-resolution mapping of replication origins. *Science* **294**, 2357–2360 (2001).

7. Kurdستاني, S.K., Robyr, D., Tavazoie, S. & Grunstein, M. Genome-wide binding map of the histone deacetylase Rpd3 in yeast. *Nature Genet.* **31**, 248–254 (2002).
8. Robyr, D. et al. Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. *Cell* **109**, 437–446 (2002).
9. Damelin, M. et al. The genome-wide localization of Rsc9, a component of the RSC chromatin-remodeling complex, changes in response to stress. *Mol. Cell* **9**, 563–573 (2002).
10. Ng, H.H., Robert, F., Young, R.A. & Struhl, K. Genome-wide location and regulated recruitment of the RSC nucleosome-remodeling complex. *Genes Dev.* **16**, 806–819 (2002).
11. Spellman, P.T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
12. Zhang, Y. & Reinberg, D. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev.* **15**, 2343–2360 (2001).
13. Sun, Z.W. & Allis, C.D. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* **418**, 104–108 (2002).
14. van Steensel, B., Delrow, J. & Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nature Genet.* **27**, 304–308 (2001).
15. Meselson, M. & Stahl, F. The replication of DNA in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **44**, 671–682 (1958).
16. Raghuraman, M.K. et al. Replication dynamics of the yeast genome. *Science* **294**, 115–121 (2001).
17. Khodursky, A.B. et al. Analysis of topoisomerase function in bacterial replication fork movement: use of DNA microarrays. *Proc. Natl Acad. Sci. USA* **97**, 9419–9424 (2000).
18. Gerton, J.L. et al. Inaugural article: Global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97**, 11383–11390 (2000).
19. Ng, H.H. & Bird, A. DNA methylation and chromatin modification. *Curr. Opin. Genet. Dev.* **9**, 158–163 (1999).
20. Jones, P.A. & Laird, P.W. Cancer epigenetics comes of age. *Nature Genet.* **21**, 163–167 (1999).
21. Jackson-Grusby, L. et al. Loss of genomic methylation causes p53-dependent apoptosis and epigenetic deregulation. *Nature Genet.* **27**, 31–39 (2001).
22. Karpf, A.R. et al. Inhibition of DNA methyltransferase stimulates the expression of signal transducer and activator of transcription 1, 2, and 3 genes in colon tumor cells. *Proc. Natl Acad. Sci. USA* **96**, 14007–14012 (1999).
23. Liang, G., Gonzales, F.A., Jones, P.A., Orntoft, T.F. & Thykjaer, T. Analysis of gene induction in human fibroblasts and bladder cancer cells exposed to the methylation inhibitor 5-aza-2'-deoxycytidine. *Cancer Res.* **62**, 961–966 (2002).
24. Suzuki, H. et al. A genomic screen for genes upregulated by demethylation and histone deacetylase inhibition in human colorectal cancer. *Nature Genet.* **31**, 141–149 (2002).
25. Huang, T.H., Perry, M.R. & Laux, D.E. Methylation profiling of CpG islands in human breast cancer cells. *Hum. Mol. Genet.* **8**, 459–470 (1999).
26. Tompa, R. et al. Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. *Curr. Biol.* **12**, 65–68 (2002).
27. Yan, P.S. et al. Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays. *Cancer Res.* **61**, 8375–8380 (2001).
28. Yan, P.S. et al. CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer. *Clin. Cancer Res.* **6**, 1432–1438 (2000).
29. Wei, S.H. et al. Methylation microarray analysis of late-stage ovarian carcinomas distinguishes progression-free survival in patients and identifies candidate epigenetic markers. *Clin. Cancer Res.* **8**, 2246–2252 (2002).
30. Shi, H. et al. Expressed CpG island sequence tag microarray for dual screening of DNA hypermethylation and gene silencing in cancer cells. *Cancer Res.* **62**, 3214–3220 (2002).
31. Gitan, R.S., Shi, H., Chen, C.M., Yan, P.S. & Huang, T.H. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. *Genome Res.* **12**, 158–164 (2002).
32. Adorjan, P. et al. Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.* **30**, e21 (2002).
33. Kallioniemi, A. et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821 (1992).
34. Solinas-Toldo, S. et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosom. Cancer* **20**, 399–407 (1997).
35. Pinkel, D. et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20**, 207–211 (1998).
36. Pollack, J.R. et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet.* **23**, 41–46 (1999).
37. Snijders, A.M. et al. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet.* **29**, 263–264 (2001).
38. Cai, W.W. et al. Genome-wide detection of chromosomal imbalances in tumors using BAC microarrays. *Nature Biotechnol.* **20**, 393–396 (2002).
39. Hodgson, G. et al. Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genet.* **29**, 459–464 (2001).
40. Massion, P.P. et al. Genomic copy number analysis of non-small cell lung cancer using array comparative genomic hybridization: implications of the phosphatidylinositol 3-kinase pathway. *Cancer Res.* **62**, 3636–3640 (2002).
41. Heiskanen, M.A. et al. Detection of gene amplification by genomic hybridization to cDNA microarrays. *Cancer Res.* **60**, 799–802 (2000).
42. Geschwind, D.H. et al. Klinefelter's syndrome as a model of anomalous cerebral laterality: testing gene dosage in the X chromosome pseudoautosomal region using a DNA microarray. *Dev. Genet.* **23**, 215–229 (1998).
43. Lucito, R. et al. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. *Genome Res.* **10**, 1726–1736 (2000).
44. Kato-Maeda, M. et al. Comparing genomes within the species *Mycobacterium tuberculosis*. *Genome Res.* **11**, 547–554 (2001).
45. Steinmetz, L.M. et al. Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**, 326–330 (2002).
46. Lindblad-Toh, K. et al. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnol.* **18**, 1001–1005 (2000).

47. Mei, R. *et al.* Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.* **10**, 1126–1137 (2000).
48. Schubert, E.L. *et al.* Single nucleotide polymorphism array analysis of flow-sorted epithelial cells from frozen versus fixed tissues for whole genome analysis of allelic loss in breast cancer. *Am. J. Pathol.* **160**, 73–79 (2002).
49. Dean, F.B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99**, 5261–5266 (2002).
50. Wilhelm, M. *et al.* Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Res.* **62**, 957–960 (2002).
51. Fritz, B. *et al.* Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma. *Cancer Res.* **62**, 2993–2998 (2002).
52. Takeo, S. *et al.* Examination of oncogene amplification by genomic DNA microarray in hepatocellular carcinomas: comparison with comparative genomic hybridization analysis. *Cancer. Genet. Cytogenet.* **130**, 127–132 (2001).
53. Hui, A.B., Lo, K.W., Yin, X.L., Poon, W.S. & Ng, H.K. Detection of multiple gene amplifications in glioblastoma multiforme using array-based comparative genomic hybridization. *Lab. Invest.* **81**, 717–723 (2001).
54. Hui, A.B., Lo, K.W., Teo, P.M., To, K.F. & Huang, D.P. Genome wide detection of oncogene amplifications in nasopharyngeal carcinoma by array based comparative genomic hybridization. *Int. J. Oncol.* **20**, 467–473 (2002).
55. Zhao, J. *et al.* Combined comparative genomic hybridization and genomic microarray for detection of gene amplifications in pulmonary artery intimal sarcomas and adrenocortical tumors. *Genes Chromosom. Cancer* **34**, 48–57 (2002).
56. Kauraniemi, P., Barlund, M., Monni, O. & Kallioniemi, A. New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. *Cancer Res.* **61**, 8235–8240 (2001).
57. Clark, J. *et al.* Identification of amplified and expressed genes in breast cancer by comparative hybridization onto microarrays of randomly selected cDNA clones. *Genes Chromosom. Cancer* **34**, 104–114 (2002).
58. 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).
59. Varis, A. *et al.* Targets of gene amplification and overexpression at 17q in gastric cancer. *Cancer Res.* **62**, 2625–2629 (2002).
60. Bruder, C.E. *et al.* High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH. *Hum. Mol. Genet.* **10**, 271–282 (2001).
61. Veltman, J.A. *et al.* High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am. J. Hum. Genet.* **70**, 1269–1276 (2002).
62. Lin, J.Y. *et al.* Physical mapping of genes in somatic cell radiation hybrids by comparative genomic hybridization to cDNA microarrays. *Genome Biol.* **3**, research0026.1–research0026.7 (2002).
63. Behr, M.A. *et al.* Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284**, 1520–1523 (1999).
64. Weinmann, A.S., Yan, P.S., Oberley, M.J., Huang, T.H. & Farnham, P.J. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* **16**, 235–244 (2002).
65. Ren, B. *et al.* E2F integrates cell cycle progression with DNA repair, replication, and G<sub>2</sub>/M checkpoints. *Genes Dev.* **16**, 245–256 (2002).
66. Berman, B.P. *et al.* Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **99**, 757–762 (2002).
67. Markstein, M., Markstein, P., Markstein, V. & Levine, M.S. Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the *Drosophila* embryo. *Proc. Natl Acad. Sci. USA* **99**, 763–768 (2002).
68. Levy, S., Hannehalli, S. & Workman, C. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **17**, 871–877 (2001).