SAGE and related approaches for cancer target identification

Dale Porter[1], Jun Yao[2] and Kornelia Polyak[2]

[1]Oncology, Novartis Institutes for Biomedical Research, Cambridge, MA 02139, USA
[2]Department of Medical Oncology, Dana Farber Cancer Institute, 44 Binney Street Harvard Medical School, Boston, MA 02115, USA

Comprehensive genetic, epigenetic and transcriptional analyses of normal and cancerous tissues and cells have yielded many candidate diagnostic, predictive, and prognostic markers and therapeutic targets in human cancer. This article provides a brief overview of SAGE and SAGE-like techniques, highlighting their utility and advantages relative to other genomic technologies for the discovery of drug targets. We also summarize the results of recent comprehensive profiling studies that utilize these methods to provide insights into mechanisms of tumor initiation and progression, to improve our molecular understanding of the tumor microenvironment and to reveal new targets and avenues for therapeutic interventions.

An approach to the identification of novel molecular tumor markers and therapeutic targets is to analyze normal and cancerous tissue in an unbiased way by using comprehensive genomic technologies. In the past few years, several such methods have been developed for the analysis of genetic, epigenetic, gene expression, protein and enzymatic activity profiles. Using these approaches, the molecular classification of human tumors has become a reality and molecular signatures correlating with metastatic behavior and clinical outcome have been identified in various cancer types [1–9]. Among other methods, gene expression profiling and hierarchical clustering of breast cancers have identified classes of tumors with poor clinical prognosis [1–4,10], and studies have found distinct gene expression patterns depending on breast cancer 1 and 2 (*BRCA1* and *BRCA2*) gene status [11,12]. Gene expression profiling has also been used for the identification of candidate therapeutic targets, although the validation of drug target candidates emerging from these types of studies requires significant follow-up work. Methods developed for genome-wide expression profiling can be divided into two main categories: closed and open platform-based approaches. DNA microarrays and SAGE (serial analysis of gene expression) are key technologies representative of closed and open platform-based genome-wide expression profiling approaches, respectively.

SAGE and related techniques

SAGE is a gene expression profiling method that allows for global, unbiased and quantitative characterization of transcriptomes [13]. Following the development of SAGE, numerous additional methods based on the use of short sequencing tags were derived and used for transcriptome and genome analyses [14]. The SAGE protocol is outlined in Box 1 and Figure 1. Detailed descriptions of the SAGE expression profiling technique and protocols, as well as access to SAGE data and data analysis tools are available on the world wide web (Box 2). The SAGE expression profiling technique is based on two principles: (i) a short, 10–27 bp sequence, referred to as a SAGE 'tag', derived from a defined position of the mRNA is sufficient to uniquely identify a transcript, and (ii) the concatemerization of these SAGE tags increases the efficiency of sequence-based transcriptome analyses [13]. A collection of SAGE tags derived from a single cell or tissue sample comprises a 'SAGE library', reflecting the identity and abundance of all transcripts in a given sample at a given time. Advantages of SAGE include that it does not require *a priori* knowledge of gene sequences and, thus, is useful for identification of novel transcripts (transcripts from previously undiscovered genes and not predicted by *in silico* analyses), that the data exists in 'digital format' (i.e. tag numbers reflect actual transcript copy numbers) and that relatively small amounts of tissue (10,000–50,000 cells) are required for the generation of comprehensive expression profiles, without requirements for an

Corresponding author: Polyak, K. (Kornelia_Polyak@dfci.harvard.edu).

1359-6446/06/$ – see front matter PII: S1359-6446(05)03694-9

BOX 1

Outline of the SAGE method

1. Extract RNA from tissue or cells of interest and capture onto oligo(dT)-magnetic beads.
2. Synthesize double stranded cDNA.
3. Digest anchored cDNAs with a frequent cutting 'anchoring' restriction enzyme. The enzyme used in this step is called an 'anchoring enzyme' because it will determine the location of the SAGE tag in the cDNA, thus, 'anchors' the tag.
   3.1 Nla III is well suited for this step because its short 4 bp recognition sequence, CATG, occurs commonly in the genome and will be present in almost all cDNAs. In addition, the Nla III site generates a 4 bp overhang making it ideal for ligation. After Nla III digestion, only 3′ cDNA fragments remain attached to the magnetic beads, and the 3′-most Nla III site defines a unique position in each cDNA transcript for subsequent mapping to the transcriptome.
4. Ligate linkers containing a type IIs 'tagging' restriction enzyme recognition site and one of two PCR primer sequences. The enzyme used in this step is called a 'tagging enzyme' because cleavage with this enzyme leads to the generation of SAGE tags (described below).
   4.1 The bound cDNA pool is divided into two fractions for ligation with one of two linkers, each having the same type IIs restriction enzyme recognition sequence but with unique PCR primer sites (i.e. primer sites A or B).
5. Digest linker-adapted cDNA with a type IIs 'tagging' restriction enzyme that cleaves the cDNA a defined distance from its recognition sequence and releases short cDNA fragments 'tags' from the beads:
   5.1 Bsmf I tagging enzyme cleaves 14 bp downstream of its recognition sequence generating a 14 bp tag (regular SAGE).
   5.2 Mme I tagging enzyme cleaves 21 bp downstream of its recognition sequence generating a 21 bp tag (LongSAGE).
6. Blunt end released cDNA tags with Klenow DNA polymerase (this step is omitted in LongSAGE to maximize tag length).
7. Ligate tags to form ditags.
   7.1 Tags with PCR primer A and B linkers are mixed and ligated with T4 DNA ligase to form ditags.
8. PCR amplification of linker-adapted ditags with primers A and B.
9. Release of ditags from linkers by digestion with Nla III.
10. Ligation of ditags to form concatemers.
    10.1 T4 DNA ligase is used to concatemerize ditags with sticky Nla III overhangs. Resulting concatemers are size-selected before cloning.
11. Clone concatemers into plasmid vectors. SphI cut pZERO1.0 is routinely used for this step.
12. Sequence clones to determine the sequence identity and number of tags representing mRNA transcript identity and abundance. Each tag will have the same length because of the restriction enzymes used for their generation.
13. The collection of all tags from one sample is called a 'SAGE library'.

amplification step. However, if the examination of an even smaller number of cells is necessary, for example for the analysis of a unique cell type, a linear RNA amplification step can be performed before the SAGE tag generation steps and the resulting SAGE libraries have faithful representation of the transcriptome [15].

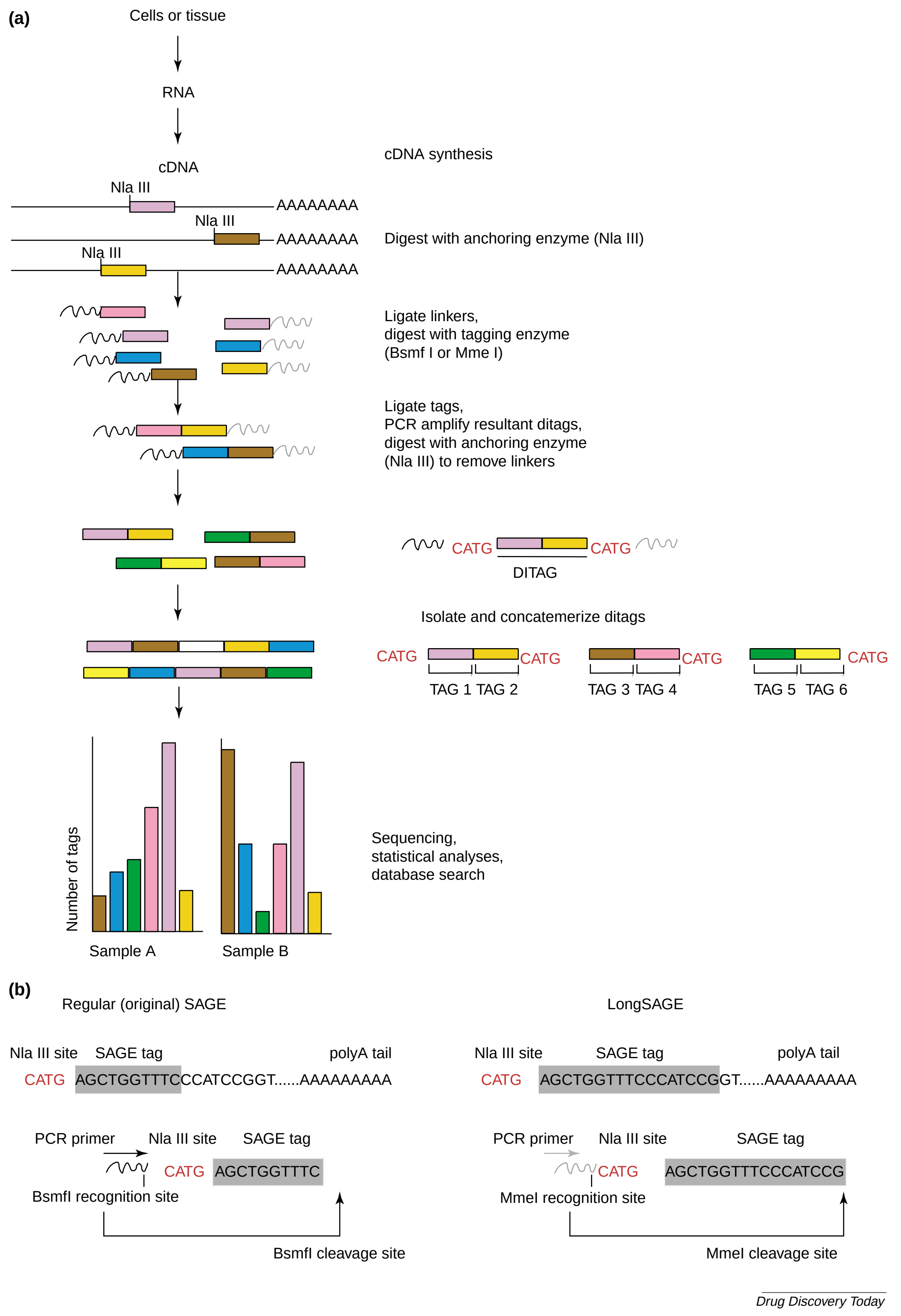
Various modifications of the original SAGE protocol have been described. Modifications that result in longer SAGE tags are advantageous because resulting tags can more frequently be uniquely matched to genomic sequence data (e.g. 17 bp-LongSAGE [16], 27 bp-SuperSAGE [17] and RECORD (representation by concatenation of restriction digests) [18]). MPSS (massively parallel signature sequencing) is a proprietary method requiring sophisticated instruments that accelerate data acquisition by sequencing the tags on the surface of microbeads [19]. In addition, techniques to sequence tags derived from the 5′-end of transcripts and useful for identification of mRNA start sites include CAGE (cap analysis gene expression) [20], TEC-RED (trans-spliced exon-coupled RNA end determination) [21] and 5′ SAGE [22]. GIS (gene identification signature) analysis extracts tag sequence information for the 5′- and 3′-ends of all full-length cDNAs to create pair-ended ditags (PETs), which are then concatenated for efficient sequencing and mapping to genome sequences to annotate the transcription start and stop sites for all genes [23]. Therefore, the GIS technique is particularly useful for unveiling differential expression of alternatively spliced transcripts between cancer and normal cells. Recent studies using SAGE and related techniques to analyze genetic, epigenetic and transcriptional changes during cancer progression have revealed mechanistic insights and therapeutic opportunities in cancer research.

Comparison of SAGE with array-based methods

Gene expression profiling is commonly performed using cDNA or oligonucleotide arrays. Compared with SAGE, array-based expression profiling, involving the hybridization of cDNA samples to arrays of oligonucleotides or cDNA sequences, can be performed more rapidly on large numbers of samples [24]. Given that array-based profiling has these advantages of economy and scale and that SAGE has advantages of being unbiased and digitally quantitative, experimental needs often dictate which technique should be employed (Table 1). In general, comparison of array-based profiling with SAGE suggests that agreement between platforms is fairly good for the quantitation of highly abundant genes and large differences in expression levels, but only modest when comparing expression of low copy number transcripts and detecting small differences in mRNA levels [25–27]. This emphasizes the crucial need to perform follow-up validation studies using independent techniques, such as quantitative RT–PCR or northern blot analysis. Additionally, the utility of SAGE for the discovery and characterization of novel transcripts and its use to analyze poorly characterized transcriptomes of model organisms is clear [26].

Combining SAGE with array comparative genomic hybridization (aCGH) for amplicon target identification

SAGE data have proven to be invaluable for the identification of genetic amplification targets in breast tumors. The target of an amplicon is a gene with increased copy number and overexpression that provide a selective advantage for cells containing this amplicon. Many oncogenes (e.g. *ERBB2* and *MYC*) are amplified in breast cancer and other cancers. Chromosomal copy number changes can be analyzed using various methods, including array comparative genomic hybridization (aCGH), which involves the competitive hybridization of genomic DNA isolated from normal and tumor cells to an array containing cDNA, oligo or BAC (bacterial artificial chromosome) clones [28]. This data will provide information on the relative copy number of the clones present on the arrays in normal and tumor cells. Identification of amplification targets, however, is notoriously difficult because the real



Drug Discovery Today

target is usually masked inside an amplicon, with many of the neighboring genes being amplified to similar extent. Even the minimum common amplified region, defined after alignment of the same amplified area from different tumors, might still have too many genes left to analyze. To facilitate this process, an algorithm was developed allowing the combined analysis of SAGE and aCGH data for the identification of candidate targets of amplicons in commonly amplified regions (Yao *et al.*, manuscript submitted). This approach assigns genetic amplification (based on aCGH data) and mRNA overexpression (relative to normal breast tissue, based on SAGE data) values for each gene within an amplicon and ranks the candidates based on statistical analysis. Applying this approach for

FIGURE 1

Outline of the SAGE method. (a) Schematic outline of SAGE library generation and analysis. Double stranded cDNA is synthesized from mRNA isolated from cells or tissues and immobilized to oligo(dT)-magnetic beads, and then digested using the anchoring enzyme (commonly Nla III). Following Nla III digestion, linkers that contain a recognition site for the tagging enzyme (Bsmf I for regular SAGE or Mme I for LongSAGE) are ligated to the 3′ cDNA ends. Linker-tag fragments are then released from the cDNA following digestion with the tagging enzyme. Resulting free linker-tag fragments are ligated together into 'ditags', PCR amplified, concatemerized, subcloned into a vector and finally sequenced as one long fragment of DNA. Each 14 bp (regular SAGE) or 21 bp (LongSAGE) tag should uniquely identify a specific gene transcript and the abundance of tags sequenced in a given library reflects the absolute transcript level within the sample analyzed. SAGE tags are indicated with differently colored bars, whereas wavy black and gray lines denote linkers. **(b)** Comparison of regular (original) and LongSAGE. Only the step that is different between the two procedures is highlighted. In the case of regular SAGE, the tagging enzyme is Bsmf I that will lead to the generation of 14 bp tags (CATG + 10 nucleotides), whereas in LongSAGE, the use of Mme I as tagging enzyme will result in 21 bp tags (CATG + 17 nucleotides). Location of the Nla III sites (CATG), tagging enzyme recognition sites (in the linkers) and cleavage sites, PCR primers within the linkers (wavy black and gray lines) and sequence of potential tags (highlighted in gray) are shown.

the analysis of the *ERBB2* amplicon, which contains more than 20 genes, identified only two candidate targets (*ERBB2* and the neighboring gene *PERLD1*). Compared with using microarrays for the purpose of assigning an mRNA overexpression value, SAGE has advantages and disadvantages. The simplicity of SAGE data structure, which is made up of stand-alone tag libraries, allows fast and accurate comparisons among different samples. Most microarrays only detect relative changes in gene expression between samples, thus the results are not portable between different studies and data analysis is fairly complex and more error prone [29,30].

Transcriptome changes during breast cancer progression

Analyses of several SAGE libraries generated from tissues and isolated cells, derived from primary normal breast and breast cancers of various stages, demonstrated that the most dramatic transcriptome changes occur at the earliest stages of cancer development – the transition from normal epithelia to ductal carcinomas *in situ* (DCIS) [31,32]. DCIS is an early-stage, pre-invasive tumor that is thought to be an obligate precursor of invasive cancer. Because of the increased use of mammograms, a significant fraction (~25%) of new breast cancer patients is diagnosed with DCIS. However, because of our lack of understanding of the molecular pathophysiology of DCIS, the clinical management of these patients is still not well defined. The finding that the gene expression profile of DCIS cells was dramatically different from normal cells, but very similar to that of invasive tumors, emphasizes the importance of examining early lesions to decipher pathways involved in tumorigenesis. Interestingly, several of the genes dramatically down-regulated in tumor epithelial cells compared with normal cells, demonstrated by SAGE analysis, encode secreted proteins: chemokines GROα and β (melanoma growth stimulating activity, α and β), LIF (leukemia inhibitory factor) and HIN-1 (high in normal-1), among others [31,32]. Chemokines have traditionally been thought to regulate immune function but recent evidence suggests that they might influence tumorigenesis by directly acting on tumor cells [33,34]. The differential expression of several secreted proteins and receptors between normal and cancerous mammary epithelial cells suggests potential targets involved in autocrine and/or paracrine signaling.

BOX 2

URL links related to SAGE

- SAGEnet: www.sagenet.org
- CGAP: http://cgap.nci.nih.gov
- NCBI SAGEmap: www.ncbi.nlm.nih.gov/sage
- Gene Expression Omnibus: www.ncbi.nlm.nih.gov/geo
- SAGE genie: http://cgap.nci.nih.gov/SAGE/
- Digital Karyotyping: http://cgap-stage.nci.nih.gov/SAGE/DKViewHome
- WebSAGE: http://bioserv.rpbs.jussieu.fr/websage/index.php
- Array data on breast cancer: http://genome-www.stanford.edu/breast_cancer/
- Oncomine: http://www.oncomine.org/main/index.jsp
- MD Anderson SAGE site: http://sciencepark.mdanderson.org/ggeg/default.html
- Tagsorter: www.tagsorter.com
- Labonweb: www.labonweb.com
- SADE: http://www-dsv.cea.fr/thema/get/sade.html
- 5′ SAGE: http://5sage.gi.k.u-tokyo.ac.jp

Further analysis of all publicly available SAGE data identified 149 genes differentially expressed between *in situ* [35] and invasive breast cancer. These genes were classified into functional categories, including extracellular matrix (ECM) or secreted proteins (13%), cell cycle (12%), cell adhesion and motility (6%) and signal transduction (6%). Collagens were among the genes most significantly overexpressed in invasive breast carcinoma. This observation is particularly interesting in light of work demonstrating by SAGE analysis that *COL6A3* (collagen type 6, α 3) is one of the most highly upregulated genes in cisplatin resistant ovarian cancer cells [36]. Thus, changes in ECM might contribute to resistance to antitumor drugs and a greater understanding of tumor cell – ECM interactions could provide insight on how to treat drug-resistant tumors. A SAGE database analysis of chemosensitivity comparing cytotoxin-sensitive cell lines with their less-sensitive corresponding solid tumors revealed that ECM proteins were significantly and coordinately overexpressed in solid tumors when compared with cultured cancer cells [37]. In addition, poor 5-year survival correlated with expression of ECM genes in solid tumors from patients diagnosed with metastatic cancers [37]. ECM proteins could, therefore, serve as prognostic and predictive markers and provide possible targets for anticancer therapy.

Gene expression changes in the tumor microenvironment

Validation of genes that were found to be differentially expressed between normal and cancerous breast and pancreatic tissues by using mRNA *in situ* hybridization and/or immunohistochemical techniques demonstrated that differential gene expression is not restricted to tumor epithelial cells [38–40]. Several genes were found to be differentially expressed by tumor-associated stromal cells, including fibroblasts, endothelial cells and inflammatory cells [38–40]. For example, osteonectin gene overexpression is localized to the juxtatumoral stroma in breast cancer and MMP11 (matrix metallopeptidase 11) gene overexpression is localized to juxtatumoral stroma in breast and pancreatic cancer [38–40]. The restriction of osteonectin and MMP11 overexpression to stromal cells most proximal to tumor cells suggests these genes might be

TABLE 1

Comparison of SAGE and array-based methods

Feature	SAGE (open platforms)	Arrays (closed platforms)
Throughput	Low	High
Novel transcript discovery	Often	Never
Quantitation of expression of multiple transcripts (including anti-sense) from the same gene	Often	Possible only using specially designed arrays and knowing the transcript in advance
Transcriptome analysis in uncharacterized organisms	Yes	No
Data Analysis	Digital data allows simple uniform analysis	Varied, difficult to normalize, and often complex
Inter-experimental comparisons	Easily done	Often difficult or not possible
Technical demand	Demanding but kits are commercially available	Relatively simple
Determination of absolute mRNA levels	SAGE data reflects absolute transcript levels	Difficult and less accurate
Cost	Higher than arrays but cost will drop as sequencing expenses are reduced	Relatively low

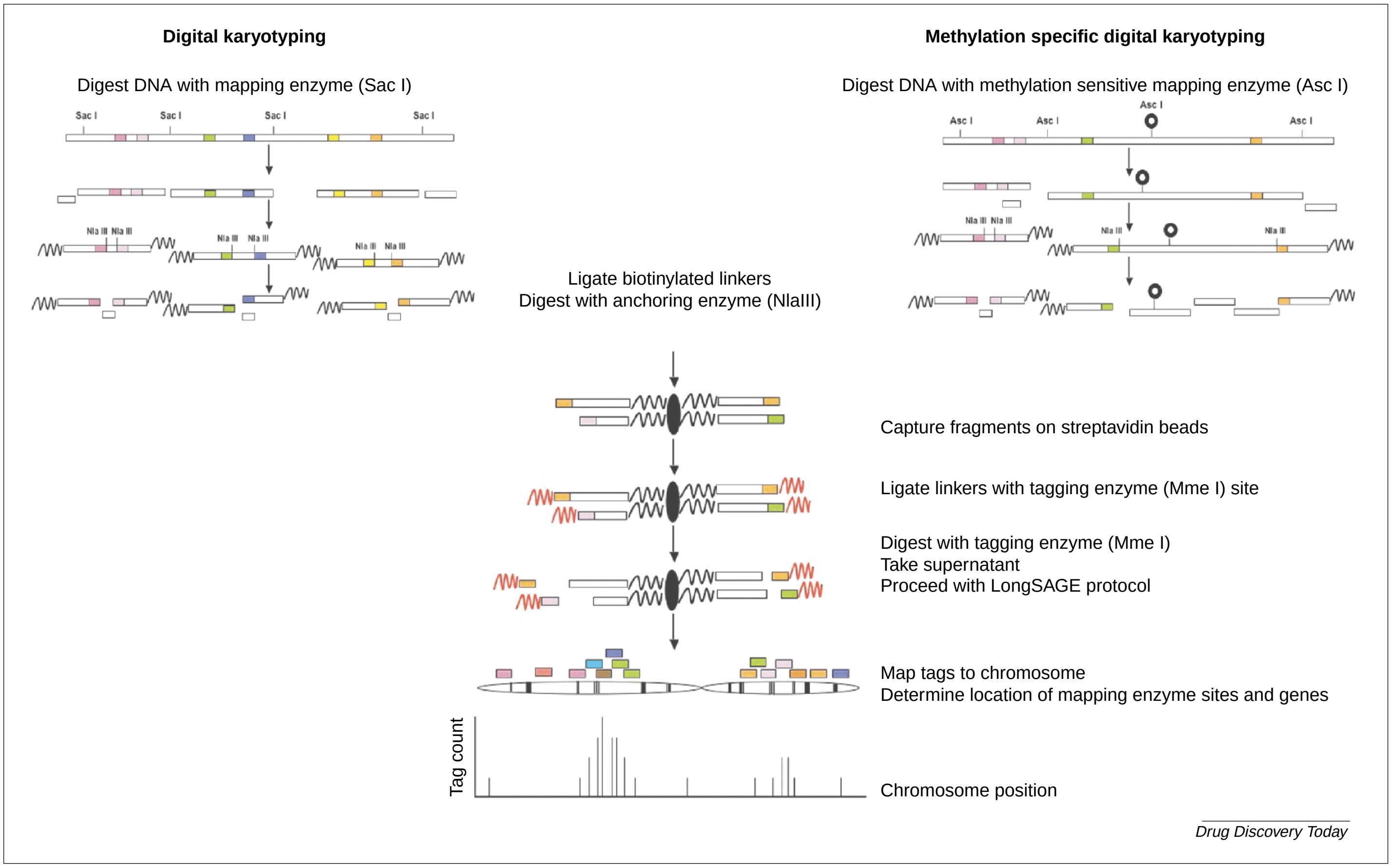
involved in the invasive process and might represent therapeutic targets in genetically stable stromal cells.

Targeting the growth of tumor endothelial cells is an attractive therapeutic avenue. Bevacizumab (Avastin™), a monoclonal antibody against vascular endothelial growth factor (VEGF), has shown efficacy as a combination therapeutic in clinical trials investigating the treatment of metastatic colorectal cancer [41]. When SAGE was used to profile gene expression profiles of endothelial cells derived from normal and malignant colorectal tissues, tumor-specific endothelial markers were discovered [42]. The finding that tumor endothelium is molecularly distinct from endothelium derived from normal tissue might have important implications for the development of antiangiogenic therapies. For breast cancer, a database of gene expression changes accompanying vascular proliferation during tumor progression has been generated [43]. Vascular endothelial-cadherin (VE-cadherin) and osteonectin were among the genes validated as upregulated in breast tumor vasculature. In addition, the protein tyrosine phosphatase PTP4A3 (PRL-3), which is overexpressed in metastatic colon cancer, was found to be overexpressed in the vasculature of invasive breast cancers and shown to stimulate the migration of endothelial cells *in vitro* [43]. These results provide insights into vascular regulation and suggest potential roles for VE-cadherin, osteonectin, and PRL-3 in driving tumor angiogenesis.

Comprehensive gene expression profiles of each cell type (epithelial, myoepithelial and endothelial cells, as well as leukocytes, myofibroblasts and stromal fibroblasts) composing normal breast tissue and *in situ* and invasive breast carcinomas have been generated using SAGE [44]. Extensive gene expression changes were observed in all cell types during cancer progression and these changes occurred in the absence of stromal cell genetic alterations. Chemokines CXCL14 and CXCL12 (SDF-1) were among the transcripts overexpressed in tumor myoepithelial cells and myofibroblasts, respectively [44]. CXCL12 has previously been implicated in breast cancer metastasis to bone and was shown to bind receptors on epithelial cells and enhance proliferation and invasion [45,46]. More recent data suggest that CXCL12 might also play a role in the recruitment of endothelial progenitors to the tumor, thus, promoting angiogenesis [47]. These observations taken together with the noted changes in expression of secreted proteins by cancer cells during breast tumor progression (described above) suggest that chemokines and other cell non-autonomous factors might act as paracrine mediators of epithelial–stromal cell interactions. Antagonizing these interactions might provide an avenue for therapeutic intervention. In accordance with this, several studies reported that downregulation or pharmacologic inhibition of CXCR4, the receptor for CXCL12, leads to decreased tumor cell growth, migration, invasion and metastasis [46,48–50]. Because several chemokine inhibitors are already in clinical trials for the treatment of various immune related disorders, it is fairly straightforward to evaluate them for the treatment of selected cancer types and a few such studies are already ongoing for breast cancer and glioblastoma.

Analysis of genetic changes using digital karyotyping

In addition to transcriptome analysis, several 'SAGE-like' methods have recently been developed for the analysis of changes in DNA copy number, methylation or chromatin structure status. All these methods are similar to SAGE and are based on two concepts: short sequence tags can be derived from specific locations in the human genome and these tags can directly be matched to the human genome sequence and consecutively arranged along the chromosomes. The 21–27 bp genomic tags usually contain sufficient information to uniquely match them to the genome and identify the loci that they were derived from. Digital karyotyping (DK) is a recently developed technology that allows the analysis of DNA copy number in a quantitative manner at genome-wide scale [51] (Figure 2). Similarly, the RECORD method generating 27 bp genomic tags can also be used for genomic copy number analysis [18]. Digital karyotyping has been used for the identification of gene amplifications, homozygous deletions and gross chromosomal alterations [51]. By comparing DK libraries generated from matched normal and tumor sample, homozygous deletions are identified as tags present only in normal cells, whereas chromosomal copy number gains are detected as tags significantly more abundant in tumor cells. Digital karyotyping was used for the analysis of colon cancer metastases resistant to 5-fluorouracil (5FU) therapy, with the aim of identifying putative genetic alterations responsible for chemotherapeutic resistance [52]. Interestingly, one of the most common amplicons in tumors of 5FU-treated patients was located on 18p11.32 and included the thymidylate synthase gene (*TYMS*), the pharmacologic target of 5FU. Furthermore, patients with amplification of


Digital karyotyping
Methylation specific digital karyotyping
Digest DNA with mapping enzyme (Sac I)
Digest DNA with methylation sensitive mapping enzyme (Asc I)
Sac I
Asc I
Nla III
Ligate biotinylated linkers
Digest with anchoring enzyme (NlaIII)
Capture fragments on streptavidin beads
Ligate linkers with tagging enzyme (Mme I) site
Digest with tagging enzyme (Mme I)
Take supernatant
Proceed with LongSAGE protocol
Map tags to chromosome
Determine location of mapping enzyme sites and genes
Tag count
Chromosome position

Drug Discovery Today

FIGURE 2

Outline of DK and MSDK protocols. DNA is digested with mapping enzyme (Sac I for DK and methylation sensitive enzyme Asc I in the case of MSDK), ligated to biotinylated linkers and then digested with Nla III as anchoring enzyme. DNA fragments with biotinylated linkers are then captured on streptavidin coated magnetic beads and the remaining steps are essentially the same as for LongSAGE. Following their extraction using the SAGE software, the tags are matched to a virtual tag library based on the predicted location of the restriction enzyme sites in the human genome. DK or MSDK tags are indicated with differently colored bars, location of restriction enzyme sites (Asc I and Nla III) are marked with bars or with pinheads (methylated Asc I site that will not be cut by the enzyme), whereas wavy black and red lines denote linkers.

TYMS had significantly shorter overall survival than patients without amplification, implicating *TYMS* amplification as the molecular basis of therapeutic resistance to 5FU in metastatic colon carcinomas [52].

Digital karyotyping analysis of medulloblastoma cell lines performed by two independent groups demonstrated that the *OTX2* homeogene is highly amplified in a subset of tumors [53,54]. *OTX2* is one of the important regulators of cerebellar development and differentiation. Subsequently, *OTX2* was found to be overexpressed in the majority of anaplastic medulloblastomas, downregulation of *OTX2* using siRNA inhibited medulloblastoma cell growth *in vitro* [53] and treatment of cells with retinoic acid, a differentiation inducing agent, downregulated *OTX2* mRNA levels resulting in apoptosis in cells overexpressing *OTX2* [53]. These findings suggest that *OTX2* expressing medulloblastomas might be amenable to therapy with all-*trans* retinoic acid.

Analysis of DNA methylation using methylation specific digital karyotyping

The MSDK (methylation specific digital karyotyping) method is very similar to the original digital karyotyping procedure. The only difference is that, instead of Sac I (a frequent cutter enzyme not affected by DNA methylation), a methylation sensitive enzyme (Asc I) is used as mapping enzyme (Figure 2). In theory, any methylation sensitive restriction enzyme can be used as mapping enzyme for MSDK [55]. However, Asc I is a particularly good choice for mapping enzyme based on the fact that (i) the Asc I recognition sequence (GGCGCGCC) contains two possible methylated CG sequences, (ii) its recognition site is preferentially located in CpG islands associated with coding genes versus repetitive elements, and (iii) it is a rare cutter enzyme (~5,000 Asc I sites per human genome), allowing fairly comprehensive methylation analysis at reasonable sequencing costs. The feasibility of this method was demonstrated by preparing MSDK libraries (MSDK tags obtained from one sample) using genomic DNA isolated from the HCT116 human colon cancer cell line and its derivative in which the DNMT1 and DNMT3b DNA methyltransferases have been homozygously deleted [56]. As a result of the deletion of these two DNA methyltransferases, the methylation of the genomic DNA in the knockout cells is reduced by >95% compared with the parental HCT116 cells, which resulted in the upregulation of several genes including *INK4a* (*p16*), *TIMP3*, *FAT* and *DUX4* [56,57]. MSDK analysis of these cells identified 261 tags that were statistically significantly (p<0.05) differentially present in the two libraries, with the majority

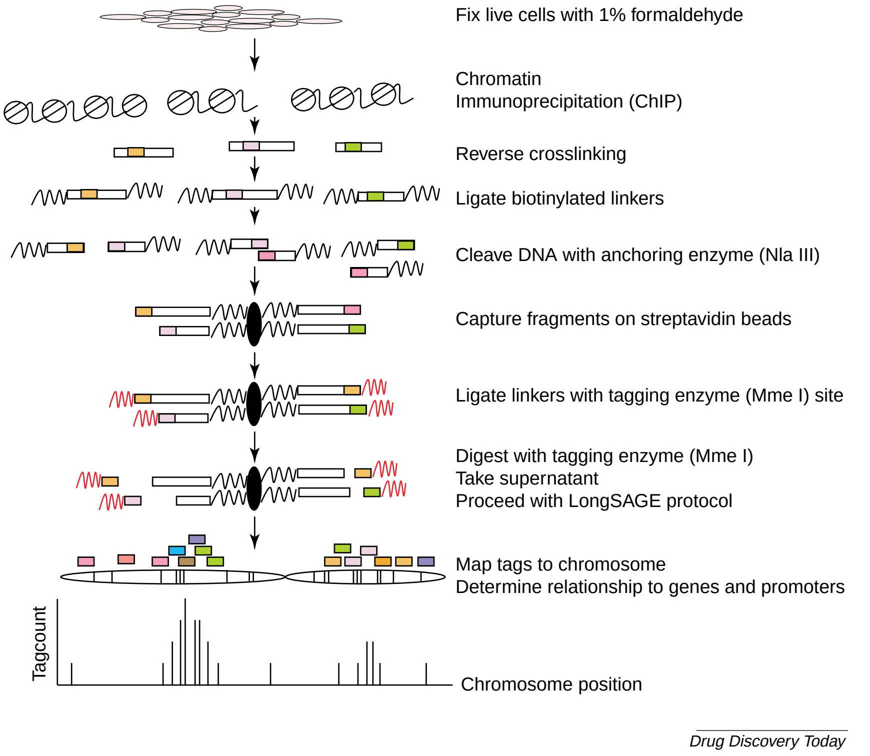

Fix live cells with 1% formaldehyde
Chromatin
Immunoprecipitation (ChIP)
Reverse crosslinking
Ligate biotinylated linkers
Cleave DNA with anchoring enzyme (Nla III)
Capture fragments on streptavidin beads
Ligate linkers with tagging enzyme (Mme I) site
Digest with tagging enzyme (Mme I)
Take supernatant
Proceed with LongSAGE protocol
Map tags to chromosome
Determine relationship to genes and promoters
Tagcount
Chromosome position
Drug Discovery Today

FIGURE 3

Schematic outline of the GMAT and SACO methods. Cells are fixed with 1% formaldehyde to immobilize chromatin and transcription factors, followed by fragmentation by sonication to generate 300–500 bp fragments. ChIP assays are performed using the appropriate specific and control antibodies, followed by reverse crosslinking and ligation of biotinylated linkers. Digestion with a frequent cutter enzyme, such as Nla III, cleaves the majority of the ChIP DNA; resulting fragments are then immobilized on streptavidin beads and linkers containing the recognition sequence of Mme I tagging enzyme are ligated. Digest with tagging enzyme Mme I releases the linker–tag fragments. From this point on, the steps of the LongSAGE protocol are followed for the generation of the libraries. Resulting tags are mapped to the human genome and their relationship to genes, promoter areas and transcriptional factor recognition sites are analyzed. GMAT and SACO tags are indicated with differently colored bars, whereas wavy black and red lines denote linkers.

of the tags being more abundant in the DKO library, correlating with the hypomethylation of these cells. Interestingly, a significant fraction of the differentially methylated genes encoded transcription factors, particularly homeogenes and other regulators of developmental processes, suggesting that cellular differentiation might be controlled by epigenetic mechanisms. However, it is not known what fraction of the differentially methylated genes is also differentially expressed between the two cell types. Thus, demonstrating a role for epigenetic changes as a regulator of cellular differentiation requires further studies.

Another application of MSDK was for the analysis of epigenetic changes in epithelial and myoepithelial cells, and stromal fibroblasts in normal breast tissue and breast carcinomas [55]. All prior methylation studies have focused only on the cancer cells themselves, whereas cells composing the tumor microenvironment have not been analyzed. Thus, epigenetic changes in stromal cells have not been previously reported. Interestingly, methylation changes were detected not only in the cancer epithelial, but also in myoepithelial cells and stromal fibroblasts. Similar to the colon cancer study, a significant fraction of differentially methylated genes encoded transcription factors. Analysis of the effects of methylation changes on gene expression revealed that, whereas promoter area methylation inversely correlates with mRNA levels (hypermethylation leads to decreased expression), methylation in introns and 3′ untranslated regions can result in increased mRNA levels. These results imply a role for epigenetic alterations in the maintenance of the abnormal tumor microenvironment in breast cancer. In addition, using MSDK, hundreds of novel loci aberrantly methylated in cancer were identified, many of which could potentially be used for the design of molecular based cancer diagnostic tests. The results of this study also highlight the complexity of epigenetic changes in tumorigenesis and raise issues about the applicability of agents that have generalized effects on DNA methylation (5-azacitidine) or chromatin structure (histone deacetylase inhibitors) for cancer therapy.

Analysis of chromatin structure and targets of transcription factors

One way of analyzing overall chromatin structure is the localization of DNase I-hypersensitive sites and associated *cis*-regulatory elements. DACS (digital analysis of chromatin structure) is a methodology that was developed to allow the quantitative, digital analysis of such DNase I-hypersensitive sites in a high-throughput manner. DACS was used for the analysis of erythroid cells and led to the discovery of large differences in the accessibility of distant regulatory sequences, suggesting a hierarchy of chromatin organization not detected by conventional assays [58]. For the analysis of targets of specific DNA-binding proteins several groups independently developed a technique that combines chromatin immunoprecipitation (ChIP) with a SAGE-like method, outlined in Figure 3. One group called it SACO (serial analysis of chromatin occupancy) and used it for the identification of transcriptional targets of CREB in rat PC12 cells [59]. Another group named it STAGE (sequence tag analysis of genomic enrichment) and identified targets of the yeast TATA-box binding protein with it [58], whereas a third group coined the term GMAT (genome-wide mapping technique), and used this technique for the analysis of active (acetylated) chromatin in yeast and in resting and activated human T cells [60,61]. By using SACO, Impey *et al.* [59] performed the most comprehensive identification of transcription factor binding sites in a metazoan species. They identified 6302 loci supported by multiple GSTs (genomic sequence tags) and confirmed CREB binding and functional relevance of many of these. Specifically, in all the cases analyzed, binding of CREB to the identified sites influenced the transcription of the downstream gene. Similarly, by using the GMAT method, Roh *et al.* [60,61] identified 4045 acetylation loci that might mediate global chromatin remodeling in response to T cell activation. Furthermore, many of these sites were localized in known regulatory elements in T cells, confirming the validity of the approach. This technique shows great promise for genome-wide mapping of transcription factor binding sites in cancer cells. Although the SACO and GMAT SAGE-like techniques have not been employed to map transcription factor binding in cancer cells to date, a related technique combining ChIP with tiled microarrays suggests that such an analysis will provide important insights. The association of estrogen receptor (ER) with the complete sequence of human chromosomes 21 and 22, combining ChIP with tiled microarrays, has revealed that ER selectively binds to a limited number of sites in MCF-7 breast cancer cells [62]. These ER binding sites are often distant from transcription start sites of regulated genes, and direct ER binding requires Forkhead factor (FoxA1) binding in close proximity. In addition, knockdown of FoxA1 expression blocks the association of ER with chromatin and estrogen-induced gene expression, suggesting that proximal Fox1A binding is required in mediating an ER-mediated response in breast cancer cells. The SACO and GMAT technique described above could be used to extend these observations across the entire genome and can be similarly used to interrogate the genome to find genuine direct binding sites for additional transcription factors known to play roles in tumor cell growth and survival.

Concluding remarks

In the past ten years since it was developed, SAGE has proven to be useful for the identification of candidate diagnostic markers and therapeutic targets in various cancer types. Modification of the original method, such as its application for genomic DNA analysis, further enhanced its utility for cancer target identification. One of the limitations of SAGE is the associated high cost for sequencing. However, several recent technical advances are likely to revolutionize DNA sequencing, leading to fast high-volume sequencing at low cost. This would allow studies using SAGE and SAGE-like methods to be performed at the same or lower cost and throughput as those using commercial DNA microarrays.

Despite these limitations, SAGE and SAGE-like techniques have identified several useful prognostic markers and therapeutic targets in cancer. In breast cancer, SAGE profiling indicates that the most profound changes in gene expression associated with tumorigenesis occur early – at the normal-to-*in situ* carcinoma transition – and that many of the changes affect expression of secreted proteins. In addition, SAGE profiling of cellular subtypes from tumors at different stages identified extensive gene expression changes in all cell types analyzed further, emphasizing the role of epithelial–stromal cell interactions during tumor progression. Profiling additional tumor types using these techniques will further elucidate mechanisms of tumorigenesis and lead to the identification of new therapeutic targets.

References

1 Sorlie, T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10869–10874

2 van 't Veer, L.J. *et al.* (2002) Expression profiling predicts outcome in breast cancer. *Breast Cancer Res.* 5, 57–58

3 van 't Veer, L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536

4 van de Vijver, M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347, 1999–2009

5 Ramaswamy, S. *et al.* (2003) A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* 33, 49–54

6 Pomeroy, S.L. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415, 436–442

7 Alizadeh, A.A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503–511

8 Bhattacharjee, A. *et al.* (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13790–13795

9 Singh, D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 203–209

10 Wessels, L.F. *et al.* (2002) Molecular classification of breast carcinomas by comparative genomic hybridization: a specific somatic genetic profile for BRCA1 tumors. *Cancer Res.* 62, 7110–7117

11 Hedenfalk, I. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* 344, 539–548

12 Hedenfalk, I. *et al.* (2003) Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* 100, 2532–2537

13 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487

14 Harbers, M. and Carninci, P. (2005) Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods* 2, 495–502

15 Heidenblut, A.M. *et al.* (2004) aRNA-longSAGE: a new approach to generate SAGE libraries from microdissected cells. *Nucleic Acids Res.* 32, e131

16 Saha, S. *et al.* (2002) Using the transcriptome to annotate the genome. *Nat. Biotechnol.* 20, 508–512

17 Matsumura, H. *et al.* (2003) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15718–15723

18 Tengs, T. *et al.* (2004) Genomic representations using concatenates of Type IIB restriction endonuclease digestion fragments. *Nucleic Acids Res.* 32, e121

19 Brenner, S. *et al.* (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* 18, 630–634

20 Shiraki, T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15776–15781

21 Hwang, B.J. *et al.* (2004) Genome annotation by high-throughput 5′ RNA end determination. *Proc. Natl. Acad. Sci. U. S. A.* 101, 1650–1655

22 Wei, C.L. *et al.* (2004) 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. *Proc. Natl. Acad. Sci. U. S. A.* 101, 11701–11706

23 Ng, P. *et al.* (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* 2, 105–111

24 Cuperlovic-Culf, M. *et al.* (2005) Determination of tumour marker genes from gene expression data. *Drug Discov. Today* 10, 429–437

25 van Ruissen, F. *et al.* (2005) Evaluation of the similarity of gene expression data estimated with SAGE and Affymetrix GeneChips. *BMC Genomics* 6, 91

26 Ibrahim, A.F. *et al.* (2005) A comparative analysis of transcript abundance using SAGE and Affymetrix arrays. *Funct. Integr. Genomics* 5, 163–174

27 Lu, J. *et al.* (2004) A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. *Genomics* 84, 631–636

28 Pinkel, D. and Albertson, D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.* 37 (Suppl), S11–S17

29 Pollack, J.R. *et al.* (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23, 41–46

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

31 Krop, I.E. *et al.* (2001) HIN-1, a putative cytokine highly expressed in normal but not cancerous mammary epithelial cells. *Proc. Natl. Acad. Sci. U. S. A.* 98, 9796–9801

32 Porter, D.A. *et al.* (2001) A SAGE (serial analysis of gene expression) view of breast tumor progression. *Cancer Res.* 61, 5697–5702

33 Zlotnik, A. (2004) Chemokines and cancer. *Ernst Schering Res Found Workshop* 45, 53–58

34 Balkwill, F. (2003) Chemokine biology in cancer. *Semin. Immunol.* 15, 49–55

35 Abba, M.C. *et al.* (2005) Gene expression signature of estrogen receptor alpha status in breast cancer. *BMC Genomics* 6, 37

36 Sherman-Baust, C.A. *et al.* (2003) Remodeling of the extracellular matrix through overexpression of collagen VI contributes to cisplatin resistance in ovarian cancer cells. *Cancer Cell* 3, 377–386

37 Stein, W.D. *et al.* (2004) A Serial Analysis of Gene Expression (SAGE) database analysis of chemosensitivity: comparing solid tumors with cell lines and comparing solid tumors from different tissue origins. *Cancer Res.* 64, 2805–2816

38 Iacobuzio-Donahue, C.A. *et al.* (2002) The desmoplastic response to infiltrating breast carcinoma: gene expression at the site of primary invasion and implications for comparisons between tumor types. *Cancer Res.* 62, 5351–5357

39 Iacobuzio-Donahue, C.A. *et al.* (2002) Exploring the host desmoplastic response to pancreatic carcinoma: gene expression of stromal and neoplastic cells at the site of primary invasion. *Am. J. Pathol.* 160, 91–99

40 Porter, D. *et al.* (2003) Molecular markers in ductal carcinoma *in situ* of the breast. *Mol. Cancer Res.* 1, 362–375

41 Alekshun, T. and Garrett, C. (2005) Targeted therapies in the treatment of colorectal cancers. *Cancer Control* 12, 105–110

42 St Croix, B. *et al.* (2000) Genes expressed in human tumor endothelium. *Science* 289, 1197–1202

43 Parker, B.S. *et al.* (2004) Alterations in vascular gene expression in invasive breast carcinoma. *Cancer Res.* 64, 7857–7866

44 Allinen, M. *et al.* (2004) Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 6, 17–32

45 Muller, A. *et al.* (2001) Involvement of chemokine receptors in breast cancer metastasis. *Nature* 410, 50–56

46 Kang, Y. *et al.* (2003) A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* 3, 537–549

47 Orimo, A. *et al.* (2005) Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated SDF-1/CXCL12 secretion. *Cell* 121, 335–348

48 Chen, Y. *et al.* (2003) Down-regulation of CXCR4 by inducible small interfering RNA inhibits breast cancer cell invasion *in vitro*. *Cancer Res.* 63, 4801–4804

49 Rubin, J.B. *et al.* (2003) A small-molecule antagonist of CXCR4 inhibits intracranial growth of primary brain tumors. *Proc. Natl. Acad. Sci. U. S. A.* 100, 13513–13518

50 Smith, M.C. *et al.* (2004) CXCR4 regulates growth of both primary and metastatic breast cancer. *Cancer Res.* 64, 8604–8612

51 Wang, T.L. *et al.* (2002) Digital karyotyping. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16156–16161

52 Wang, T.L. *et al.* (2004) Digital karyotyping identifies thymidylate synthase amplification as a mechanism of resistance to 5-fluorouracil in metastatic colorectal cancer patients. *Proc. Natl. Acad. Sci. U. S. A.* 101, 3089–3094

53 Di, C. *et al.* (2005) Identification of OTX2 as a medulloblastoma oncogene whose product can be targeted by all-trans retinoic acid. *Cancer Res.* 65, 919–924

54 Boon, K. *et al.* (2005) Genomic amplification of orthodenticle homologue 2 in medulloblastomas. *Cancer Res.* 65, 703–707

55 Hu, M. *et al.* (2005) Distinct epigenetic changes in the stromal cells of breast cancers. *Nat. Genet.* 37, 899–905

56 Rhee, I. *et al.* (2002) DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* 416, 552–556

57 Paz, M.F. *et al.* (2003) Genetic unmasking of epigenetically silenced tumor suppressor genes in colon cancer cells deficient in DNA methyltransferases. *Hum. Mol. Genet.* 12, 2209–2219

58 Kim, J. *et al.* (2005) Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. *Nat. Methods* 2, 47–53

59 Impey, S. *et al.* (2004) Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* 119, 1041–1054

60 Roh, T.Y. *et al.* (2004) High-resolution genome-wide mapping of histone modifications. *Nat. Biotechnol.* 22, 1013–1016

61 Roh, T.Y. *et al.* (2005) Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* 19, 542–552

62 Carroll, J.S. *et al.* (2005) Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 122, 33–43