Lessons from genetic profiling in soft tissue sarcomas

Nilbert, Mef; Meza-Zepeda, L. A.; Francis, Princy; Berner, J. M.; Namløs, H. M.; Fernebro, Josefin; Myklebost, O.

Published in:
Acta Orthopaedica Scandinavica. Supplementum

DOI:
10.1080/00016470410001708310

2004

Lessons from genetic profiling in soft tissue sarcomas

M. Nilbert[1], L. A. Meza-Zepeda[2], P. Francis[1], J. M. Berner[2], H. M. Namløs[2], J. Fernebro[1] and O. Myklebost[2]

[1]Dept of Oncology, University Hospital, Lund, Sweden, [2]Dept of Tumour Biology, the Norwegian Radium Hospital, Oslo, Norway
mef.nilbert@onk.lu.se

ABSTRACT **Soft tissue sarcomas represent a heterogeneous group of tumors and include over 50 histotypes. Some of these tumor types are characterized by specific chromosomal translocations, whereas other types show complex genetic aberrations. The recent developments within gene expression technologies have now been applied to studies of soft tissue sarcomas (STS) and the first results indicate that genetic signatures are useful for classification and diagnosis. Distinctive expression profiles have been found in e.g. gastrointestinal stromal tumors (GISTs), synovial sarcomas, malignant peripheral nerve sheath tumors (MPNSTs), and in subsets of liposarcomas. The more pleomorphic tumor types, such as high-grade variants of leiomyosarcomas, malignant fibrous histiocytomas (MFHs), fibrosarcomas, and subtypes of liposarcomas, show a greater variability among the expression profiles, but interestingly subsets with distinctive expression profiles can be identified also among these tumors. The data available place many of the genes hypothesized to be involved in the development of a certain type of STS, such as the KIT gene in GIST development, among the top discriminating genes. Thereby expression profiling provides novel insights into the pathogenesis of STS. Although much work remains to be done to validate the data and to define optimal discriminating gene lists, the current lessons from gene expression studies in STS are encouraging and imply that genetic signatures may serve as diagnostic and prognostic markers and may help identify novel therapeutic strategies.**

The DNA chip technologies utilize the advances in high-throughput oligonucleotide synthesis or PCR amplification in order to generate arrays of immobilized DNA probes that through hybridization to RNA or cDNA sequences can detect quantitative differences. Array-based gene expression analysis has become an important tool in projects aiming to refine diagnosis and prognosis. These techniques will probably provide important data for the development of targeted therapies, and studies that have applied these technologies to bone and soft tissue sarcomas (STS) have provided exciting clues to the biology of mesenchymal tumors.

Technological issues in expression profiling

A number of different technologies may be used in microarray analysis, many of them available as commercial "ready-to-use" packages. Commercial systems are easy to set up, are generally more reproducible, but may be inflexible and expensive. Arrays produced by academic facilities are cheap and flexible, but quality controls and reproducibility may be suboptimal. The main commercial alternative has been synthetic oligonucleotide arrays from Affymetrix, but several commercial alternatives are now appearing that avoid the wide patents owned by Affymetrix. Whereas cDNA-based expression arrays have been the main academic platform, such facilities are now increasingly moving to oligonucleotide spotted arrays. Which technology to use will depend on the application, e.g. whether the aim is a global genomic screen of large tumor panels or a future routine screening with less expensive arrays containing a moderate number of preselected probes in a pathology lab. The various options when it comes to probe design, sample preparation, labelling and

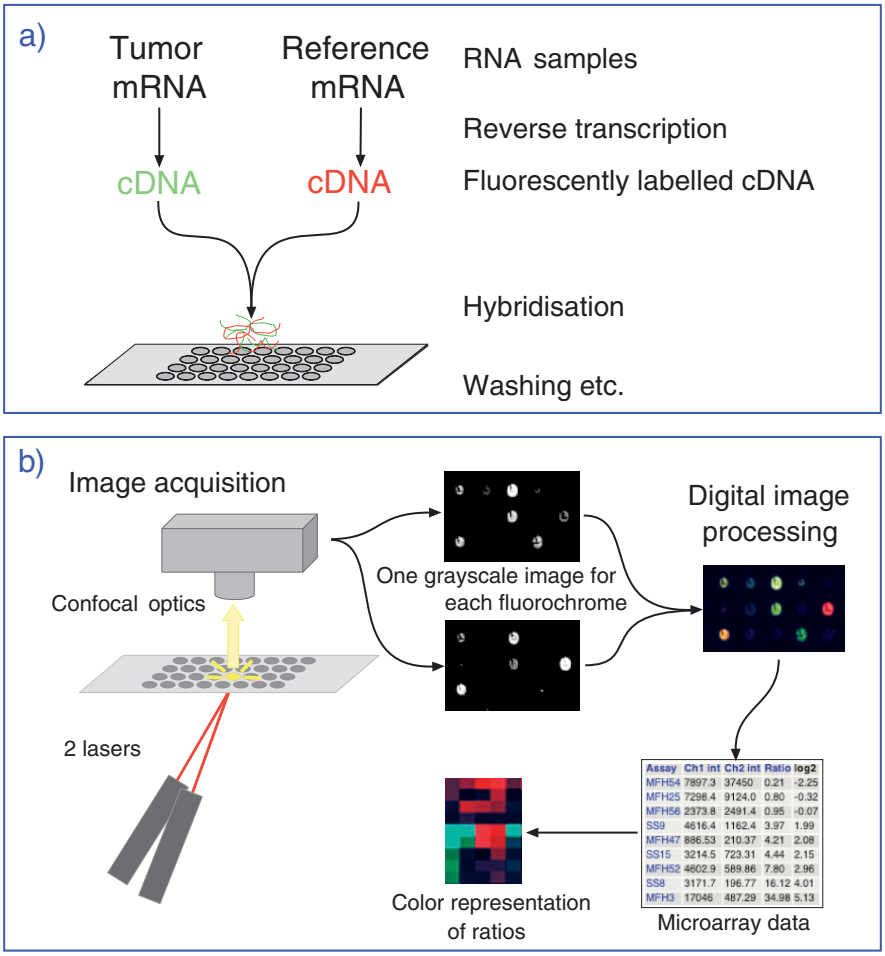


Figure 1. Basics of microarray analysis. a) RNA purified from tumor samples is used to make cDNA that is labelled with a fluorescent dye. This labelled cDNA is hybridized to the microarrays together with a reference cDNA labelled with a different dye, which is used as an internal standard. b) After hybridization and washing, the slide is scanned by a special laser scanner, and the amount of cDNA from the sample and the reference that has hybridised to each spot is measured as the intensity values for the two fluorescent dyes. The ratio values, showing the expression levels in samples relative to reference, are frequently displayed as coloured squares, where red designates higher and green lower expression.

hybridization procedures introduce artefacts and biases, and at present, the jury is still out as to which method most correctly represent the "truth". However, any of the technology platforms, when correctly applied, will produce a wealth of important and reproducible data, and the main take home message is that within each study great care must be taken so that every step is done in exactly the same way for all samples. For review of the various techniques see Duggan et al. (1999), Lipschutz et al. (1999), and Ramaswamy et al. (2002). An overview of the procedures for the analysis of spotted arrays is given in Figure 1.

Sample preparation

Analyses of gene activity measure the relative abundances of all gene transcripts (mRNAs) in the samples. A major complication is the fragility of mRNAs, which are easily degraded by potent and robust enzymes (RNAses) present in the tissues or cells. Thus, when tissues are collected, it is critical that either the RNA is quickly extracted (or the tissue is dissolved in a solution that inactivates the RNAses), or they are cooled on ice, to keep cells intact until freezing at –80º C or extraction is possible. When the tissues become anoxic and die, the RNA quickly deteriorates. Although formalin

fixation will inactivate RNAses, this is too slow to prevent degradation, and thus paraffin blocks are not suited for this kind of analysis, although more robust qualitative assays for specific mRNAs may still be possible. However, variable RNA quality is unavoidable for tumor samples, and one should be aware that uneven degradation of individual samples will result in artefacts, because degradation will affect the signal obtained from each spot depending on the probe length or distance from the 3' end of the gene.

Any purification procedure that gives intact and pure RNA may be used, but requirements are more stringent than for many other assays. Any RNAse present may partially degrade the samples during the labelling reaction, which also may be more sensitive to inhibition by other impurities, such as polysaccharide-protein complexes, that may copurify with RNA.

Some labelling protocols use purified (polyA) mRNA, whereas others use total RNA, and with few exceptions, the labelling is done by synthesis of cDNA copies that incorporate fluorescently labelled nucleotides. The cDNA is made by a reverse transcriptase (RT), initiated from small synthetic primers, that either prime from the 3' polyA tail (and other A-rich sequences), or randomly along the whole sequence. The simplest procedure does this in the presence of labelled nucleotides (direct labelling). Because the fluorescent groups are large and are not easily accommodated by especially some types of RT, and especially if the sequence contains stretches of the corresponding nucleotide, this may affect labelling efficiencies, leading to unreliable results (called dye effects). One way to avoid this is to add the fluorochromes after cDNA synthesis using biotinylated nucleotides ("post-labelling").

Direct labelling is rather simple, but requires considerable amounts of RNA, from 10 to 50 micrograms per microarray. This translates to 10^6–10^8 cells, or 20 mg to 1 g of tissue, depending on type. Thus, various methods to amplify either the target RNA or the signals from the arrays are frequently used. In particular, if studies are to be done on microdissected material, this is mandatory. Details about these are beyond the scope here, but the procedures employ various manipulations to achieve representative multiplication of target mRNAs, either by PCR-based techniques, or by making cDNA copies that include RNA promoters so that multiple cRNA copies can be made from each cDNA. Using such procedures, nanogram amounts of RNA may be used, but care is needed to avoid artefacts due to the many manipulations, giving various biases (Nygaard et al. 2003). Whatever method is used, the results will usually not compare well with those from experiments done with another method.

Reference samples

A reference sample, labelled with a different fluorochrome, is generally included as an internal control in hybridizations to spotted arrays, allowing calibration of the signals between experiments. For each spot the signal from the target to be investigated and reference can be measured separately, and the ratio between the two net signals is calculated (Figure 1b), and this is the measure of relative expression level that is used in further analysis (see bioinformatics section). It is important for this purpose that the reference gives a signal in most of the spots, because otherwise one cannot calculate a ratio. Additionally, it is important that there is a reproducible supply of the reference, so that new samples or panels can be compared with older results. At the biological level, it is important to understand the relevance of this reference, as it determines the interpretation of increases or decreases in expression. For profiling of tumor panels, a commercially provided mixture of RNA from several cell lines is advantageous since these are prepared in large batches at an industrial scale with limited batch variation. We have produced a tumor-focused reference that includes sarcoma samples. This reference gave a better representation of the genes expressed in sarcomas than the commercial cell line-based reference (Berner, Namløs and Myklebost, unpublished). A commercial reference sample from human tissues, which would be expected to be more suited to tumor profiling than cell lines is also available. Importantly, for tumor profiling, it is the pattern of signals between the tumor samples that is important, whereas the apparent up-regulation or down-regulation relative to a cell line reference has no particular biological significance.

cDNA microarray

As the name implies, the probes of these arrays consist of cDNAs, partial copies of human mRNAs, usually from the large collection of the IMAGE Consortium. These cDNAs are generally short, i.e. 500-2000 bases, and made from the 3' end of the mRNA, primed from the polyA tail. Because most protocols for labelling of samples for microarray analysis are primed from the 3' end, it is convenient that the probes also are from this end, as the 5' parts of the mRNA is frequently lost due to partial degradation, and it is difficult to produce long labelled cDNAs. Most labs use the 40 k Unigene set of clones, consisting of 40 000 bacterial clones, each containing (in principle) one single IMAGE cDNA, each supposed to represent one unique gene (see bioinformatics section). The bacteria are grown, plasmid purified, and the cDNA inserts amplified in 96-well format. Because of contamination problems, both at the plate replication steps at the repository, and during probe preparation locally, these clone collections may result in some probes containing more than one PCR fragment, some containing the wrong probe, and some being absent. Up to 30 % erroneous probes have been reported for IMAGE-based arrays (Knight 2001). Thus, for important results, verification by other techniques such as real-time PCR or Northern blotting is necessary, and also sequence verification of subsets of the array probes is recommended.

An advantage of the cDNA arrays is that differential mRNA splicing is less of a problem due to the length of the probes. On the other hand, crosshybridization of transcripts from genes having some sequence similarity to the probe is a significant problem.

The cDNA arrays are invariably produced by robotic printing of the PCR fragments onto solid surfaces, most commonly specially prepared glass microscope slides. This gives spot sizes of around 100 μm, thus production of arrays of 40 000 spots is possible on standard microscope slides.

Oligonucleotide arrays

Like cDNA slides, oligonucleotide arrays may be produced by robotic printing, but some commercial suppliers, including Affymetrix and Rosetta, synthesize the oligonucleotides directly on the solid surface. The probes may be short, e.g. 22 bases for Affymetrix, or long, e.g. 60-70-mers is most common for spotted arrays. Because they are short, the selection of which part of the gene to represent is critical. When done carefully, cross-hybridization with other genes can be avoided, and all probes can have very similar hybridization properties, thus giving strong and reliable results. If mRNA splicing information is available, one or more probes can usually be designed to represent all known splice variants. An important advantage for in-house production of oligonucleotide arrays is the reduced handling and crosscontamination risk. Thus, on these arrays it is unlikely that a spot will contain a wrong probe.

Technological issues in profiling by genomic microarrays

One may also use microarrays to analyse DNA copy numbers in tumors. In this way, amplifications and deletions can be assayed genome-wide. Comparative genomic hybridization (CGH), since its initial description by Kallioniemi et al. (1992), has been widely used to detect and map changes in copy number in tumors. In CGH, DNA from a test sample (tumor) and reference (blood) are differentially labelled and hybridized competitively to normal metaphase chromosomes. The ratios of fluorescence between the test and reference DNA is calculated along several copies of each chromosome, providing information on the relative copy number in each chromosome segment. In recent years the genome representation has been replaced by an array of large genomic clones, cDNAs or oligonucleotides (Pinkel et al. 1998, Pollack et al. 1999). The array format CGH provides a number of advantages over the use of chromosomes, high resolution and dynamic range, improved reproducibility, direct mapping of changes to the genome sequence and increase throughput by allowing automation.

In principle, every part of the genome is covered in classical CGH. However, it is estimated that amplicons need to cover many million basepairs to be detected, although it is likely that shorter amplicons may be detected if highly amplified. Whereas both cDNA and oligonucleotide arrays may be used, the short probes suffer from weak signals and thus reduced reproducibility, so many groups

use genomic arrays for this purpose. The probes on these arrays consist of cloned chromosomal fragments, usually bacterial artificial chromosomes (BACs) or P1 artificial chromosomes (PACs) of 100–300 kilobases. Large inserts provide sufficient intense signals so that accurate measurements can be obtained over a broad dynamic range. BACs and PACs are more demanding to produce than cDNA probes, and thus complete coverage genomic arrays have not yet been published. Initial reports of genomic arrays used whole BACs and PACs isolated from large scale cultures, since then, different methods have been described to create representations of genomic clones by ligation-mediated PCR, or different types of degenerated oligonucleotide primed (DOP-)PCR. A common design is to include one BAC every million base-pairs, thus giving arrays of 3–4000 BACs (Figure 2a) (Snijders et al. 2001, Fiegler et al. 2003). Some BAC sets are focused, i.e. include specific regions known to be involved in cancer, or may completely cover a specific region (Figure 2d), some are defined through cytogenetic mapping, and some are sequenced so that the gene content is precisely known (Figure 2c).

A major advantage of genomic profiling is the increased stability of DNA compared to RNA. The requirement for intact targets are less for the DNA-labelling protocols, sufficient quality can be obtained even from paraffin blocks, allowing access to vast collections of archive material (although not if demineralized with acid, which destroys DNA and therefore EDTA may be used instead, Paris et al. 2003).

Classical DNA purification procedures involving organic extractions give relatively pure and intact genomic DNA for copy number analysis. Typically less than 1 microgram of total genomic DNA is labelled directly with fluorescent nucleotides by random priming. Recently various reports have shown promising methods for whole genome amplification (Lage et al. 2003, Hosono et al. 2003). These methods can generate enough material from minute clinical samples, so that gene dosage alterations down to threefold can be reproducibly detected with as few as 1000 cells of starting material (Lage et al 2003).

For DNA copy number profiling a sex-matched normal diploid genome is used as reference, usually normal leukocytes. Recent work has also shown polymorphism in DNA copy number between individuals (Lucito et al. 2003). This polymorphism can involve segments as large as several megabases that may or may not be present in different individuals. It is therefore recommended to utilize a pool of normal DNA or normal DNA from the same patient in question as reference.

Bioinformatics

The post-genomics era we currently experience, with thousands of candidate genes and proteins identified, requires well-developed bioinformatics systems in order to handle the vast amounts of data generated (Simon et al. 2003). A basic understanding of the bioinformatics involved at several stages is important when planning or interpreting microarray experiments. Below we describe procedures for analysis of spotted expression arrays. Some of the commercial arrays, like those of Affymetrix, are single channel assays, using different procedures for filtering and calibration, but down-stream analysis is more or less the same. For genomic arrays, the procedures are also different, in that normalization is done with normal DNA, and the target gives ratios above or below a fixed multiplum (reflecting non-diploid karyotypes) of the normal value, indicating deletion or gain. A complication there is to determine the correct normalization factor in samples with aneuploidy or numerous aberrations, where it might be difficult to identify reliably regions with normal copy number ratios. Another difference in this case is that the probes must be analyzed in their chromosomal positional context, and each spot represents a large genomic segment, in most cases containing multiple genes. Thus dedicated software is required for this purpose.

Gene definition

A basic problem in post genomics is annotation, which is the definition and unequivocal naming of each gene and its splice variants. This is getting even more complicated as it turns out that the sequences of many human genes actually overlap. The Unigene system defined each tentatively unique gene as a collection (contig) of overlapping cDNA sequences, and gave each a unique Unigene cluster ID (Wheeler et al. 2003). For each of these

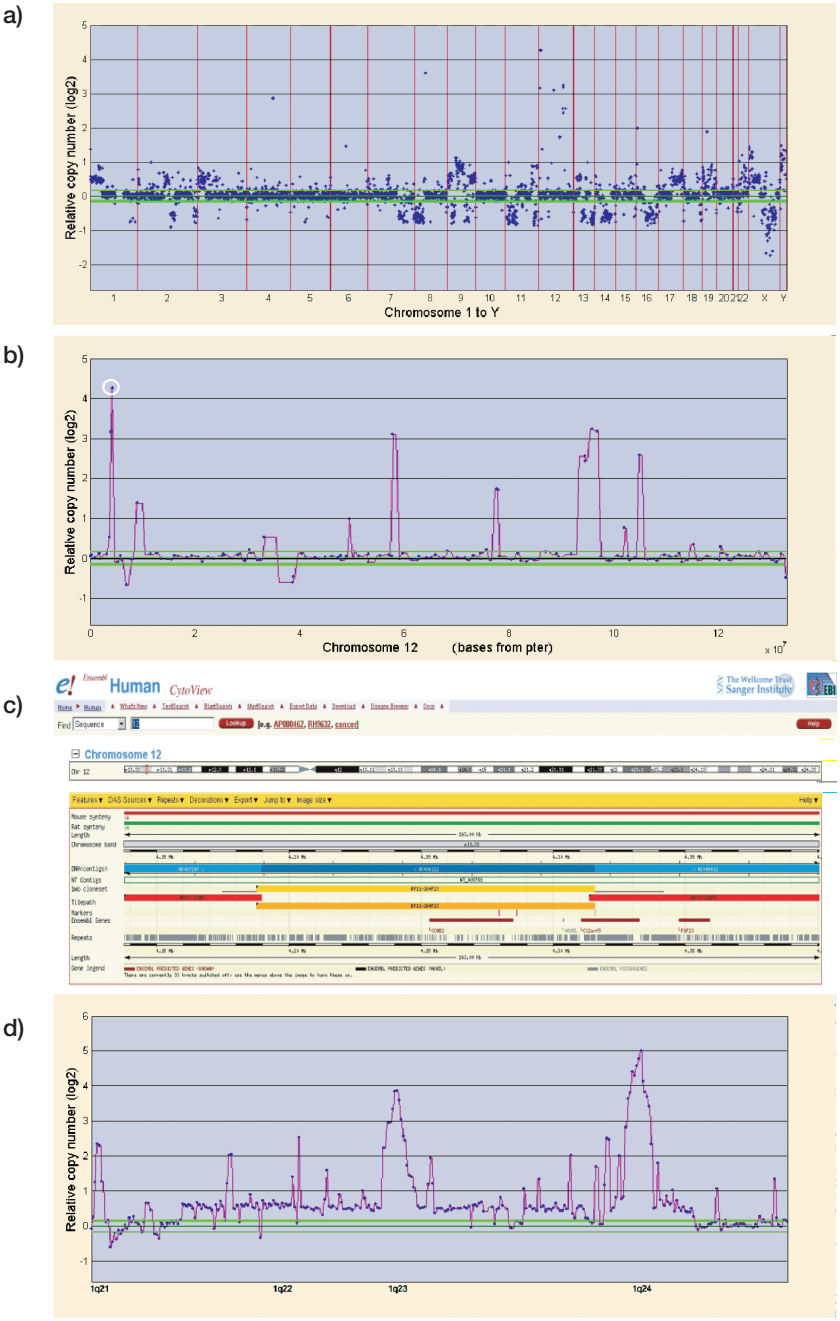


Figure 2. Examples from genomic profiling of a MPNST using arrayCGH. a) Genome-wide profile using an array with clones spaced at approximately 1Mb intervals. The CGH ratios for each BAC or PAC array element is plotted as a function of its genome location, with chromosome 1 to the left and Y to the right. Each ratio is the log mean of the quadruplicate array measurements. Vertical lines indicate chromosome boundaries. Plots were generated using M-CGH (Wang, Meza-Zepeda, Kresse and Myklebost, unpublished). b) Expanded view showing normalized copy number variation throughout chromosome 12. A white circle identifies a highly amplified sequence in 12p13, also indicated in a). c) Information linked to the corresponding genomic clone (white circle) found in Ensembl. The arrowhead indicates the BAC at the peak of the amplicon in a and b, and the CCND2 gene (coding for cyclin D2) is highlighted as the most likely gene driving the amplification. d) A region-specific array covering the tiling path between 1q21 and 1q25. This region shows well-defined amplicon (Meza-Zepeda

clusters, an IMAGE cDNA clone was selected as representative for that putative gene, and included in the 40 k Unigene clone set widely used to make arrays. However, as new sequences are obtained, in many cases, two or more Unigene clusters were found to represent the same gene, particularly as a result of alternative splicing, but also because some cDNA clones were primed inside the gene rather than from the 3' end. Other clusters were split, because some cDNAs may have been artificially joined during cloning or erroneously aligned. Thus, the Unigene IDs may change over time, and the annotation of gene names for each spot, stably represented by its IMAGE ID, must be regularly updated. More recently, as the genome sequence became available, the gene IDs from the Ensembl database (ensembl.org) are being used to annotate new probe sets. This has the advantage that each probe is directly connected to the genome sequence, and information about splice variants is to a large extent available. Furthermore, sequences shared between genes may be identified. However, it is complicated to compare results obtained with sets defined by these different means, and also those defined by commercial procedures (e.g. Affymetrix, Compugen).

Image analysis and data extraction

When performing microarray experiments, computers are used at many steps. The array needs to be scanned, and the settings and properties of the scanner may have a major impact on the results. In particular, the background level, the linearity and to what extent some signals are above the saturation level is important. Each spot and the relevant background level needs to be reliably identified, and weak spots may be filtered away. Data from spotted arrays are represented as the ratio between signals from target and reference as described above. The reliability of the signals correlates with intensity, so that weak spots are less reliable, as one would expect, partly because the background, as measured on the surrounding surface not blocked by DNA, may poorly represent that within the spot. However, when the ratios have been calculated, information about signal intensities is lost, and in general weak signals will contribute equally to downstream analysis. Thus, procedures for determination of the level of intensity that is regarded as reliable, and for handling such spots, are critical. With weak signals, experimental variation or modest differences in balance between the two channels (reference and target), may give large variation or even inversion of the ratios. If representative duplicate probes are present on the arrays, adaptive filtering may be implemented, that sets the filter at a level giving a defined level of reproducibility (Jenssen et al. 2001). Other procedures use the variation in background intensity or relate to its average level or some fixed value, and spots below this threshold in one or both of the channels are removed. However, one then looses information about transcripts that are absent in a target (but present in the reference), including deleted genes, as well as those being expressed in the target but absent in the reference. More complex procedures may take care of this issue, e.g. by replacing negative or zero values of the reference with a low, fixed value, to stabilize the ratios.

Whereas aggressive filtering will improve the reproducibility of each experiment, it may also drastically reduce the information available when sets of probes are to be compared.

Data preparation for comparison

To compare experiments, the ratios need to be normalized. This is usually done by setting the total intensity in each channel, or the sum of intensities of a set of "house hold" genes, equal, assuming the total level of expression is the same for both samples. This can be done by multiplying all intensities of one channel with the required factor, or, if the intensities are not expected to be linearly distributed, by intensity-based "Lowess" normalization (Yang et al. 2002). In most cases, $\log_2$ ratios are used, so that 1 corresponds to double, and -1 to half the expression level. Depending on what one wants to emphasize, one may remove genes which do not change much across a panel, which one is not interested in. Furthermore, to give equal emphasis to the remaining genes, disregarding the amplitude of changes, one may calibrate the variation across the panel for each gene, so that the mean is 0 (mean centering) and maximum in all cases is 1. Whereas duplicate probes for some genes may be a good internal control, they will affect e.g. clustering, and one may either only keep one, or calculate an average. It is difficult to ana-

lyze genes that are absent (filtered away) in some of the samples, and one usually omits those absent in a certain fraction of the samples. As mentioned above, if a large fraction of the genes may be weak in one or more of the samples, perhaps because the sample labelled poorly, aggressive filtering may quickly remove a majority of the results in one or more samples. Procedures are available for imputation of missing values, i.e. calculation of likely values based on those in the most similar samples to improve downstream analysis (Troyanska et al. 2001).

The contribution of mRNA from stromal cells will affect the results, and thus the surrounding tissue will contribute to the tumor pattern (see example in the clinical applications section). To some extent these gene clusters can be removed in advance (Perou et al. 2001), focussing the analysis on the profile of the cancer cells. However, one may then lose information about stromal cells that may be important prognostically, e.g the degree of infiltration of immune cells (Wang et al. 2001).

Data mining in microarray studies

It is beyond the scope of this article to review the many ways data sets may be analyzed, and we refer to Quackenbush (2001) for further introductory reading. In tumor profiling, one generally aims to do two main types of analyses; class discovery, where one looks for entirely new class structures among the samples, or class prediction, where one tries to identify expression fingerprints that can determine to which predetermined class a sample belongs. This could be used e.g. for classification of problem cases into the classical histological classes, or for prediction of whether a sample will respond to chemotherapy. In many cases interesting data are represented by a spread sheet-like display, in which each line represents the values for a gene (probe) in all the samples, and each column contains the values for all genes in one sample. To facilitate interpretation, the ratios are represented as colors, with increasing red intensity representing log ratios larger than 0, and green the negative, whereas ratios around zero are approaching black.

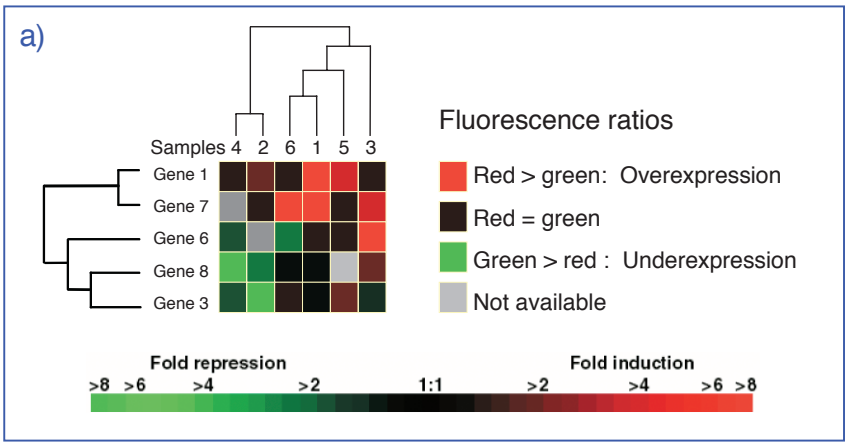
For class discovery, unsupervised methods are used, where the microarray data only are used to look for subclasses in the set. For this purpose, the way in which the filtering, centering and normalization is done will have important impact. The most common procedure is hierarchical clustering (Figure 3, Eisen et al. 1998), by which, to put it simply, the sample columns and the gene lines are resorted, so that those that are most similar lie side by side. So nothing is changed in the data, just the samples and genes are grouped according to similarity. This will reveal some of the inherent structure in the data, but the result is dependent among other things on how similarity is defined, and the order by which the values are compared. A problem with hierarchical clustering is that each sample or gene can only be placed in one relation, whereas clearly there may be relations in several directions. Using other algorithms, such as self-organizing maps and K means clustering, one may predefine the number of groups.

In supervised analysis, one uses other information about the samples to guide the microarray analysis. The majority of the samples are generally included in a learning set to define a set of genes that correlates with the properties of predetermined groups. Subsequently, this classification is used to evaluate the rest of the samples, which make up the test set. It is important to understand that with thousands of measurements and perhaps hundreds of samples (most sarcoma studies are done with tens of samples or less), there will always be many genes that just by chance correlate completely with any way to divide the panel. Therefore, validation in separate and sufficiently large sample sets is mandatory.

One may also correlate gene expression patterns with survival, e.g. through a modified log rank test (Jensen et al. 2002). This will identify genes with expression patterns that correlate with survival.

Functional interpretation

Besides classification and prediction purposes, microarray analyses yield detailed information on genes that are associated with each type of cancer or with its behaviour. These genes may reveal unknown processes involved in cancer development or progression, many of which may be diagnostically valuable or be developed as new candidate targets for therapy. However, the lists of genes are frequently quite hard to interpret, since many genes have multiple names (and indeed multiple functions), many of which do not give hints



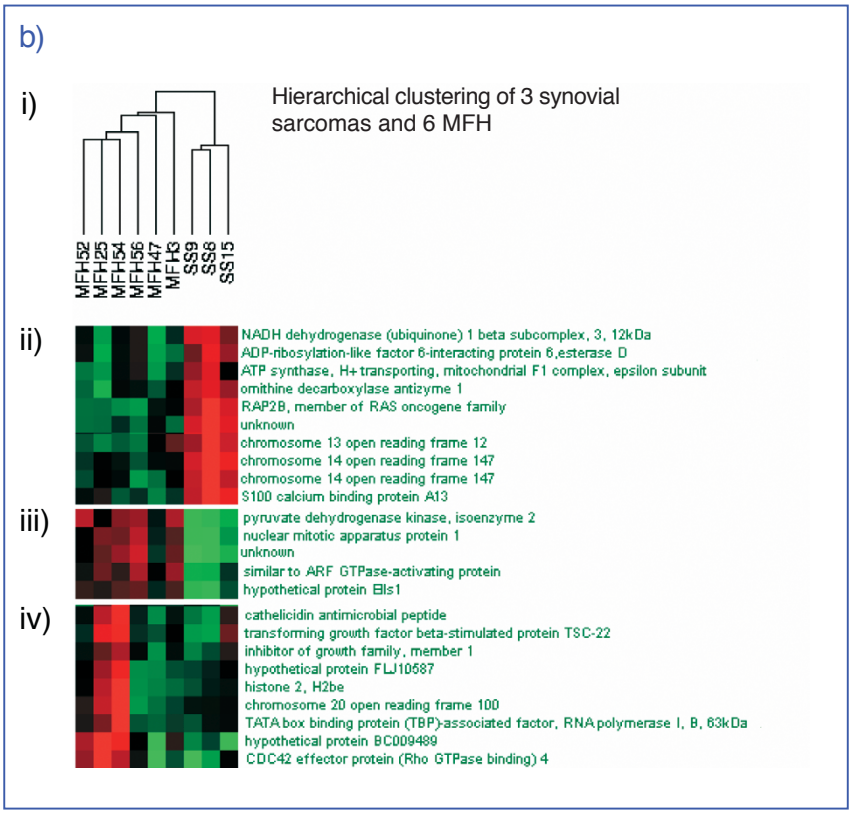


Figure 3. Hierarchical cluster analysis of expression data. a) Example showing how the rows of patient samples, and the lines of gene values, are sorted so that those with the most similar expression profile are beside each other. The length of the tree-like "dendrograms", showing how the various samples and clusters are related to each other, indicate the relative similarity. Red indicates signal from target (tumor) and green from reference. b) Real example of expression data, showing comparison of expression profiles of a group of 3 synovial sarcomas and 6 MFH. i) shows the unsupervised clustering of the samples based on all available expression data (5300 values remaining after removing weak spots etc.). The dendrogram shows separate clusters for each subtype, and indicates some subgroups within the MFH. ii) and iii) are clusters of genes with similar expression profile among the samples, that may be used to distinguish the two groups. ii) genes that are highly expressed in synovial and low in MFH samples, and iii) vice versa. iv) shows genes that are expressed in a subgroup within the MFHs, as responsible for a subcluster within this group, as indicated by the separate branches in i). The color

of their function, and still a majority of genes have no name at all. To interpret the results, clustering analysis may give hints on function because the genes behave similarly (cluster with) other genes of known function, thus some regulatory or functional relationship may be inferred. Furthermore, data mining tools that scour the literature or other databases for information on each gene can be used (e.g. PubGene.org, Jenssen et al. 2001). Several efforts are ongoing to annotate all genes with functional information, such as the Gene Ontology Consortium (GeneOntology.org), and these can be used to interpret the functions of a set of candidate genes.

Clinical applications of expression profiling in cancer

Gene expression profiling has revealed novel classification patterns in several tumor types, including STS, breast cancer, lymphoma, malignant melanoma, and acute myeloid leukemia (Golub et al. 1999, Alizadeh et al. 2000, Bittner et al. 2000, Perou et al. 2000 Sorlie et al. 2001), but application of molecular diagnosis has been tested in only a few large sample sets with most tumors being of epithelial origin. Ramaswamy et al. (2001) applied 16K oligonucleotide arrays to 218 malignant solid tumors of 14 types to study whether gene expression patterns could be used to predict tumor origin. An expression-based classifier was developed and had an overall diagnostic accuracy of 78%. The lowest accuracy was observed for poorly differentiated cancers, which represent anaplastic lesions that due to lack of morphological hallmarks may be hard to diagnose also based on morphology. Many of the anaplastic tumors probably display expression profiles quite different from those of their original tumor type. Furthermore, multi-class distinction between related tumor types may provide a greater challenge than pair-wise comparison between tumor types. Giordano et al. (2001) used 7K oligonucleotide arrays to compare the gene expression patterns in adenocarcinomas from the lung, colon, and the ovary. A correct classification with regard to organ specificity was achieved in 91% of the tumors. Interestingly, the two discordant tumors, one colonic tumor and one ovarian tumor, proved to represent a metastatic colonic adenocarcinoma in the ovary (which clustered closer to the ovarian group) and a pleomorphic mesenchymal tumor. The differentially expressed genes identified in such studies have a great potential to prove diagnostically useful, and these data constitute an important step towards the identification of organ-specific expression profiles.

The studies by Ramaswamy et al. (2002) and Giordano et al. (2001) demonstrated that gene expression profiling can be applied to tumors of unknown origin in order to denote the primary tumor type. Despite the advances in differential diagnosis using gene expression data, such applications do not constitute the major diagnostic aim, which will probably rather be the possibility to distinguish yet not identified tumor subtypes. Thereby, these technologies will serve as complementary analyses for diagnosis, and will also provide data on prognosis and form the basis for the development of novel therapeutic strategies. Molecular profiling may also shed light on the cellular origin of tumors and could thereby provide a better understanding of the precursor cells and on tumor pathogenesis.

Expression profiling in STS

The morphologic classification of STS is complex, has repeatedly been modified with novel clinicopathological entities being recognized, and now includes more than 50 histotypes. However, the morphologic classification is rather robust and when novel technologies, such as immunohistochemistry, cytogenetics, and RT-PCR have been introduced and compared with the morphological classification, these generally support and thereby validate the morphologic classification. However, even within the currently identified histopathologic subtypes, an extensive variability in morphology and clinical behaviour exists. Since 1/3 of the patients develop metastases and STS are poorly chemosensitive, novel prognostic factors and therapeutic possibilities are needed. The introduction of imatinib (known as Gleevec or STI571), an inhibitor of the KIT receptor tyrosine kinase, in the treatment of GIST is one of the best examples of the efficiency of targeted therapies. The story behind imatinib and the clinical responses obtained suggests that detailed molecular characterization of other types of STS may be of importance for

the development of targeted therapies. The introduction of expression arrays provides an exciting opportunity to identify such molecular targets for targeted therapies also in other types of STS.

Tumor heterogeneity

Since STS are often large with heterogeneous morphology, concern has been raised as to whether a random sample from a STS can provide accurate gene expression data that are representative of the entire tumor. These issues have been addressed by studies of multiple tumor biopsies from leiomyosarcomas and MFHs, which have been compared with the gene expression patterns determined from single tumor samples from the same histopathologic type of tumor. These studies have shown that the intra-tumor variability of the gene expression profiles in STS are within the variability of replicate experiments from the same tumor piece and that the variations are minimal compared to the inter-tumor variability (Shmulevich et al. 2002, Francis, unpublished observations). These findings indicate that, unless clear tumor heterogeneity is observed, an accurate molecular profiling can be obtained from single samples from STS, and that tumor size will not have a major effect on the data obtained. However, the expression levels may vary between the center and the periphery of the tumor for certain genes; overexpression of the PDGF receptor has been found to be more pronounced in the tumor periphery, whereas lysozyme and cathepsin E showed reduced expression in the tumor periphery (Shmulevich et al. 2002).

Expression-based subclassification of STS

The first studies applying gene expression arrays to STS have now been reported and these data do in general support the traditional, morphologic classification. Nielsen et al. (2002) used 22K and 42K cDNA microarrays to analyze a heterogeneous group of 41 STS, including GIST, synovial sarcomas, liposarcomas, leiomyosarcomas, MFHs and malignant peripheral nerve sheath tumors (MPNST, schwannomas). Based on unsupervised clustering of differential expression of some 5,500 genes, 5 major clusters were identified and included synovial sarcomas, GISTs, MPNSTs, a cluster containing a subset of the leiomyosarcomas, and a cluster containing heterogeneous tumor types such as liposarcomas, MFHs, and some of the leiomyosarcomas. When histopathological data was introduced to perform a supervised clustering analysis, differentially expressed transcripts that correlated with these groups could be identified. Class prediction for GIST tumors involved 125 genes, including *KIT*, synovial sarcomas were recognized by a 104-gene cluster containing e.g. *SSX*, *EGFR*, and genes involved in the retinoic acid pathway, leiomyosarcomas were associated with muscle-related genes and the MPNSTs showed expression of nerve sheath-related genes (Nielsen et al. 2002). Segal et al. (2003) also used 12K Affymetrix arrays to test whether gene expression profiling could identify novel classification schemes. In a sample set containing 51 STS, synovial sarcomas, myxoid/round cell liposarcomas, clear-cell sarcomas, and GISTs showed distinct expression profiles. The discriminating genes within this study included *SCF* and *KIT* for GIST, *WNT5A* and *FRIZZLED-1* for synovial sarcoma, genes associated with the melanocytic lineage in clear-cell sarcomas, and *CDK4* and *MDM2* in dedifferentiated liposarcomas (Segal et al. 2002). In this study some of the fibrosarcomas clustered closely to the synovial sarcoma, which may reflect a common origin for some of these tumors. The homogenous expression profiles of GISTs have also been demonstrated by Allander et al. (2001) with 13K cDNA microarrays. The *KIT* gene was indeed the most highly expressed gene and also had a top-ranking in the discriminator gene list, which suggests that the GISTs develop as a clonal expansion of cells that have acquired a *KIT* mutation and that these tumors are not affected by the extensive genetic instability that characterize many carcinomas as well as a subset of STS.

Synovial sarcomas show, among several histological features, signs of epithelial differentiation, and on the basis of such components the tumors are classified into the two major subtypes; biphasic (composed of epithelial cells arranged in glandular-like structures and spindle cells) and monophasic (composed of fibrosarcoma-like spindle cells, possibly with scattered epithelioid areas). The synovial sarcomas are characterized by the X;18-translocation, which results in the *SYT-SSX* fusion gene (Clark et al. 1994). Indeed, two major, alternative fusions exist with the *SSX1* or the *SSX2*

genes, and that *SSX1* fusions correlate with the biphasic histotype and *SSX2* with the monophasic type. The fusion proteins have transcriptional activity, and although their normal downstream targets are unknown, it is likely that their levels, and perhaps those of additional genes, will be affected by the aberrant transcription factors. This observation strongly indicates that different precursor cells for the two major types of synovial sarcomas exist, and these tumors are thought to arise from a mesenchymal stem cell with a capacity for epithelial differentiation. Nagayama et al. (2002) used 23K cDNA micorarrays and demonstrated that synovial sarcomas cluster separately compared to other types of STS, and in the hierarchical cluster analysis the synovial sarcomas clustered close to malignant peripheral nerve sheath tumors. Genes related to migration or differentiation of neural crest cells, e.g. coding for ephrin-B3, endothelin-3, retinoic acid receptor and collagen IX, showed frequent up-regulation, which raises the possibility of a connection between synovial sarcoma and neuroectodermal differentiation. Allander et al. (2002) used 6K cDNA microarrays and examined the expression profiles in 14 synovial sarcomas compared to MFHs and fibrosarcomas. Up-regulation was identified for e.g. IGF2, which acts through the IGF1 receptor, and of the negative regulators IGF1, IGFBP2, and ERBB2. Regarding morphology, biphasic tumors tend to cluster together, whereas monophasic tumors seem to be divided into separate clusters and keratin-encoding genes are among the genes that distinguish these subsets of synovial sarcomas (Allander et al. 2002, Nagayama et al. 2002, Fernebro et al., unpublished data).

Fritz et al. (2002) performed expression profiling as well as microarray-based CGH in pleomorphic, dedifferentiated liposarcomas. The genomic profiling detected the highest amplification levels in dedifferentiated liposarcomas for the genes *MDM2*, *GLI* and *CDK4*, and served as class predictors for dedifferentiated liposarcoma. All these genes are localized to 12q13-15, which indicates a close relationship between the dedifferentiated liposarcomas and well-differentiated liposarcomas, in which amplification of this segment is closely associated with the typical marker chromosomes. Clustering based on the expression levels of 1,600 genes allowed most of the tumors to be separated into pleomorphic or dedifferentiated liposarcomas, with the heat shock protein HSP90 and the adaptor protein gene *SCAP* showing higher expression in the pleomorphic liposarcomas. Also these data are in agreement with the notion that 1q22, to which band these genes are localized, is amplified in pleomorphic liposarcomas (Rieker et al. 2002).

Expression profiling may have its greatest potential for clinical use to distinguish between the pleomorphic and undifferentiated types of several common types of STS, e.g. high-grade malignant leiomyosarcomas, pleomorphic liposarcomas, and MFHs, where a novel expression-based tumor classification may offer a more objective and reproducible classification. Consequently, identification and further investigation of subgroups among the currently identified histotypes is the focus of most of the currently performed studies that apply DNA arrays to STS. Ren et al. (2003) applied Affymetrix arrays with 22K probes to analyze 35 STS. Two clusters were identified based on the expression of 92 genes and ESTs, and the gene expression profiles in these two groups were found to correlate with tumor differentiation and clinical aggressiveness. Three genes that showed frequent up-regulation, *p16*, a7-integrin, and neurotrophin, were further evaluated using RT-PCR, which confirmed frequent up-regulation of *p16*. This report thus indicates that expression profiling can subclassify leiomyosarcomas. The results also suggest that expression profiling can predict tumor behaviour and may thus be of clinical use in decisions about adjuvant therapies. The more pleomorphic tumor types, including MFH, are characterized by complex genetic alterations and less consistent expression profiles. Whereas MFHs have in some studies failed to form a separate cluster, a subset of these tumors have in other studies clustered separately (Nagayama et al. 2002, Nielsen et al. 2002, Segal et al. 2002, Ren et al. 2003). The identification of a subset of MFHs with a particular expression profile could be of diagnostic value and might also have prognostic applications.

Subclassification of pediatric solid tumors may be challenging. Kahn et al. (2001) applied 6K cDNA microarrays to train artificial neural networks (ANNs) in the classification of small round blue-cell tumors, including Ewing sarcomas,

rhabdomyosarcomas, neuroblastomas, and Burkitt lymphomas. Using the top-96 genes, the authors could train the artificial network to correctly cluster the tumors within their diagnostic categories. Wai et al. (2002) applied 2K Affymetrix arrays to distinguish cell lines derived from Ewing sarcomas, neuroblastomas, and malignant melanomas of soft parts, and did among other findings identify overexpression of WNT-signaling pathway in neuroblastoma. Kahn et al. (1998) also reported consistent gene expression patterns in alveolar rhabdomyosarcomas and demonstrated that genes related to the PAX3-FKHR fusion were among the most consistently expressed.

Genomic profiling of sarcomas

ArrayCGH can detect deletions, duplications, non-reciprocal translocations and gene amplification, phenomena frequently seen in STS. Analysis of tumor genomes by arrayCGH has focused mainly on genome-wide arrays (Figure 2a) or on particular regions of the genome where aberrations frequently occur (Figure 2d).

Unlike most epithelial tumors, STS can be defined by their molecular pathology. From a genetic perspective sarcomas can be classified into two main groups, tumors with relatively simple near-diploid karyotypes and few, rather specific, chromosomal rearrangements, and a second group of tumors with complex karyotypes, characterised by severe genomic and chromosomal instability. The first group of tumors generates distinct and homogeneous expression profiles. In the second group, specific profiles of copy number changes seem to appear as a result of a selection for particular changes affecting gene expression and genetic instability. In several small studies, copy number fingerprints have been used for classification (Fritz et al. 2002, Weiss et al. 2003, Wilhelm et al. 2003).

Relative small number of sarcoma samples have been analysed by array-based CGH. 16 dedifferentiated and pleomorphic sarcomas have been profiled using a 300 element genomic array, half of them localised to 12q (Fritz et al. 2002). Several class predictors were identified based on particular genes located in the chromosomal subregion 12q13-q15. This region, frequently amplified in well-differentiated liposarcomas, was over-represented in all the dedifferentiated but not in the pleomorphic liposarcomas. Within the 12q amplicon, *MDM2*, *GLI* and *CDK4* were coamplified. Other candidate genes from this region were also over-expressed, indicating a possible role in liposarcoma development. Two other chromosomal regions, 6q25 and 20q13, were also frequently increased in copy number. More recently, pediatric osteosarcomas were analysed using cDNA microarrays (Squire 2003). DNA from 9 samples were hybridized to a 19200-clone cDNA microarray, and copy number gains or amplification were found in 6p, 8q and 17p, all consistent with the pattern observed by metaphase CGH. The higher resolution of arrayCGH, on the other hand, allowed definition of amplicon boundaries for the 17p amplicon. Taking advantage of the increased resolution of arrayCGH, a more precise map of amplicon boundaries and maxima can be identified. Precise mapping of these regions to the genome sequence aids the identification of candidate oncogenes. Expression levels of the candidate genes in tumor panels and cell lines can be used to identify the most likely gene that drives the amplification and contributes to the oncogenic phenotype. The genes identified may give light into tumorigenesis, as well as possible targets for new therapies.

Summary and future perspectives

In summary, the current although limited data from expression profiling in STS indicate that distinct expression profiles exist for several subtypes of STS. These especially include tumors that are characterized by specific chromosomal translocations (e.g. synovial sarcoma and Ewing sarcoma) or oncogene mutation (e.g. GISTs containing mutations in the *KIT* gene). The more pleomorphic tumors that are also generally characterized by complex genetic alterations show less consistent expression profiles, although subsets of these tumors with distinct expression profiles can be identified (Nielsen et al. 2002). Furthermore, several genes and pathways suggested to be involved in the different subtypes of STSs, such as *KIT* in GIST and the wnt-pathway in synovial sarcoma have been found to be over-expressed and to contribute to the subclassification obtained.

However, the first studies available have been performed during the earliest phase of expression arrays and therefore differ considerably in

basic layout; different technological platforms (cDNA microarrays, oligonucleotide arrays, and CGH arrays) have been used, a variable number of genes and/or ESTs have been analyzed, various references RNA has been utilized, and different statistical methods have been used to handle the data obtained. Therefore, it is expected rather than surprising that various classifying gene lists have been suggested, and many of the genes that contribute to the classifications have not yet been characterized. The minimal set of genes that can correctly classify samples into their diagnostic categories depends on the number of genes included in the array, the size of the tumor material studied and probably also on the tumor type.

The current lessons from gene expression studies in STS are encouraging and suggest that these methods may be used as adjunct techniques in future differential diagnosis of STS and that classification systems based on gene expression levels may be feasible in STS. However, much work remains to be done in order to confirm and validate the very first results from expression profiling studies presented herein, and it will be critical for further development that larger collaborative, but consistently diagnosed and treated tumor panels are studied. Also, functional evidence and characterization of the pathways involved will be required before these new data can be introduced in clinical decision-making. However, novel technologies are available also for data evaluation; integrated information about a specific target based on studies at DNA level, the RNA level and the protein level can be obtained, the results can be validated in model systems, over-expressed clones can be tested in representative cohorts using the tissue microarray technology (TMA), and finally the molecular data needs to be correlated with clinical, histopathological and prognostic data.

Now that the first data on up-regulated and down-regulated genes in STS are available, this knowledge is likely to influence future studies in this field since different investigators can mine data published on the web and test candidates in independent tumor materials. Thereby, the gene arrays have a great potential to reveal the biology and the pathogenesis behind STS with important consequences for diagnosis as well as for prognosis and therapeutics and it is with great interest that genetic signatures for local recurrences and/or metastases as well as expression data related to therapeutic response in STS are foreseen.

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JL, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibishirani R, Sherlock G, Chand WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000; 403: 503-511.

Allander SV, Nupponen NN, Rigner M, Hostetter G, Maher GW, Goldberger N, Chen Y, Carpten j, Elkahloun AG, Meltzer PS. Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogenous gene expression profile. Cancer Res 2001; 61: 8624-8628.

Allander S, Ilei PB, Chen Y, Antonescu CR, Bittner M, Ladanyi M, Meltzer PS. Expression profiling of synovial sarcoma by cDNA microarrays. Am J Pathol 2002; 161: 1587-1596.

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marinkola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000; 406: 536-540.

Clark J, Rocques PJ, Crew AJ, Gill S, Shipley J, Chan AM, Gusterson BA, Cooper CS. Identification of novel genes, SYT and SSX, involved in the t(X; 18)(p11.2;q11.2) translocation found in human synovial sarcoma. Nat Genet 1994; 7: 502-508.

Duggan D, Bittner M, Chen Y, Meltzer P, Trent J. Expression profiling using cDNA microarrays. Nat Genet 1999; 21: 10-14.

Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998 95: 14863-8.

Fiegler H, Carr P, Douglas EJ, Burford DC, Hunt S, Scott CE, Smith J, Vetrie D, Gorman P, Tomlinson IP, Carter NP. DNA microarrays for comparative genomic hybridisation based on DOP-PCR amplification of BAC and PAC clones. Genes Chromosomes Cancer 2003; 36: 361-74.

Fritz B, Schubert F, Wrobel G, Schwaenen C, Wessendorf S, Nessling M, Korz C, Rieker RJ, Montgomery K, Kucherlapati R, Mechtersheimer G, Eils R, Joos S, Lichter P. Microarray-based copy number and expression profiling in dedifferentiated and pleomorphic liposarcoma. Cancer Res 2002; 62: 2993-2998.

Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, Du J, Kingsmore SF, Egholm M, Lasken RS. Unbiased whole-genome amplification directly from clinical samples. Genome Res 2003; 13: 954-64.

Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 2001 28: 21-8.

Jenssen TK, Kuo WP, Stokke T, Hovig E. Associations between gene expressions in breast cancer and patient survival. Human Genetics 2002; 111: 411-420.

Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, Smith PD, Jiang Y, Gooden GC, Trent JM, Meltzer PS. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998; 58: 5009-5013.

Khan J, Wei JS, Rigner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS. Classification and diagnostic prediciton of cancers using gene expression profiling and artificial neural networks. Nature Med 2001; 7: 673-679.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridisation for molecular cytogenetic analysis of solid tumors. Science 1992; 258: 818-21.

Knight J. When the chips are down. Nature 2001; 410: 860-861.

Lage JM, Leamon JH, Pejovic T, Hamann S, Lacey M, Dillon D, Segraves R, Vossbrinck B, Gonzalez A, Pinkel D, Albertson DG, Costa J, Lizardi PM. Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. Genome Res 2003; 13: 294-307.

Lipshutz R, Fodor S, Gingeras T, Lockhart T. High density synthetic oligonucleotide arrays. Nat Genet 1999; 21: 20-24.

Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, Rodgers L, Brady A, Sebat J, Troge J, West JA, Rostan S, Nguyen KC, Powers S, Ye KQ, Olshen A, Venkatraman E, Norton L, Wigler M. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res 2003; 13: 2291-305.

Nagayama S, Katagiri T, Tsunoda T, Hosaka T, Nakashima Y, Araki N, Kusuzaki K, Nakayama T, Tsuboyama T, Nakamura T, Imamura M, Nakamura Y, Togushida J. Genome-wide analysis of expression in synovial sarcomas using a cDNA microarray. Cancer Res 2002; 62: 5859-5866.

Nielsen TO, West RB, Linns SC, Alter O, Knowling MA, O´Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, van der Rijn M. Molecular characterization of soft tissue tumors: a gene expression study. Lancet 2002; 359; 1301-1307.

Nygaard V, Loland A, Holden M, Langaas M, Rue H, Liu F, Myklebost O, Fodstad O, Hovig E, Smith-Sorensen B. Effects of mRNA amplification on gene expression ratios in cDNA experiments estimated by analysis of variance - art. no. 11. BMC GENOMICS 2003; 4: 11.

Paris PL, Albertson DG, Alers JC, Andaya A, Carroll P, Fridlyand J, Jain AN, Kamkar S, Kowbel D, Krijtenburg PJ, Pinkel D, Schrèoder FH, Vissers KJ, Watson VJ, Wildhagen MF, Collins C, Van Dekken H. High-resolution analysis of paraffin-embedded and formalin-fixed prostate tumors using comparative genomic hybridisation to genomic microarrays. Am J Pathol 2003; 162: 763-70.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature 2000; 406: 747-52.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridisation to microarrays. Nature Genetics 1998; 20: 207-11.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nature Genetics 1999; 23: 41-6.

Ramaswamy S, Golub T. Molecular classification of cancer: class discovery and class prediction in clinical oncology. J Clin Oncol 2002; 20:1932-1941.

Ramaswamy S, Tamayo P, Rifkin R, Mukjerjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T, Gerald W, Loda M, Lander ES, Golub TR. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 2001; 98: 15149-15154.

Ren B, Yu YP, Jing L, Liu L, Michalopoulos GK, Luo JH, Rao UN. Gene expression analysis of human soft tissue leiomyosarcoma. Hum Pathol 2002; 34: 549-558.

Rieker R, Joos S, Bartsch C, Willeke F, Schwarzbach M, Otano-Joos M, Ohl S, Hogel J, Lehnert T, Lichter P, Otto HF, Mechtersheimer G. Distinct chromosomal imbalances in pleomorphic and in high grade dedifferentiated liposarcomas. Int J Cancer 2002; 99: 68-73.

Segal NH, Pavlidis P, Antonescu CR, Maki RG, Noble WS, DeSantis D, Woodruff JM, lewis JJ, Brennan MF, Houghton AN, Cordon-Cardo C. Classification and subtype prediction of adult soft tissue sarcoma by functional genomics. Am J Pathol 2003; 163: 691-700.

Shmulevich I, Hunt K, El-Naggar A, Taylor E, Ramdas L, Laborde P, Hess KR, Pollock R, Zhang W. Tumor specific gene expression profiles in human leiomyosarcoma. Cancer 2002; 94: 2069-2075.

Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003; 95: 14-18.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG. Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet 2001 29: 263-4.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen ME, van der Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 2001; 98: 10869-10874.

Squire JA, Pei J, Marrano P, Beheshti B, Bayani J, Lim G, Moldovan L, Zielenska M. High-resolution mapping of amplifications and deletions in pediatric osteosarcoma by use of CGH analysis of cDNA microarrays. Genes Chromosomes Cancer 2003; 38: 215-25.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB. Missing value estimation methods for DNA microarrays. Bioinformatics 2001; 17: 520-5.

Wai DH, Schaefer KL, Schramm A, Korsching E, Van Valen F, Ozaki T, Boecker W, Schweigerer L, Dockhorn-Dworniczak B, Poremba C. Expression analysis of pediatric solid tumor cell lines using oligonucleotide microarrays. Int J Oncol 2002; 20: 441-451.

Weiss MM, Kuipers EJ, Postma C, Snijders AM, Siccama I, Pinkel D, Westerga J, Meuwissen SG, Albertson DG, Meijer GA. Genomic profiling of gastric cancer predicts lymph node status and survival. Oncogene 2003 22: 1872-9.

Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L. Database resources of the National Center for Biotechnology Information: update. Nucleic Acids Res 2004; 32: D35-40.

Wilhelm M, Veltman JA, Olshen AB, Jain AN, Moore DH, Presti JC, Jr., Kovacs G, Waldman FM. Array-based comparative genomic hybridisation for the differential diagnosis of renal cell cancer. Cancer Res 2002; 62: 957-60.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002 30: e15.