The use of gene expression profiling in cancer decision support systems.
By: Vibhu Agarwal (vibhua@stanford.edu)

Introduction: Traditionally, histo-pathological techniques have been used to classify cancer tumors into known types. These techniques made use of chemical staining of tissues followed by visual analysis of several morphological markers by a pathologist. However the limitations of this method in recognizing and classifying primary cancers such as such as diffuse large B-cell lymphoma sub-types, as well as metastatic cancers of unknown primary are well known. Similarly breast cancer tumor classification using such techniques has been unsuccessful. Since the dispensation of effective therapy depends on the correct diagnosis of the cancer type, non targeted treatments often lead to side effects and poor prognosis. According to Foo et al. (1998) 70 – 80% of breast cancer patients on chemotherapy based on the traditional predictors would have survived without it. Identification of the correct tumor sub-type also helps in predicting survival rate and deciding on the correct therapy for treatment (more aggressive tumors with a poor prognosis can be exposed to stronger and more cytotoxic drugs).

Gene expression profiling: The set of genes that are expressed at any given time by a tissue and the relative strengths of their expressions, constitute a molecular signature for that tissue. This was demonstrated by Ross et al. by recording the expression of 9703 cDNA sequences in 60 cancer cell lines from th National Cancer Institute (NCI60). Using the expression values of each of these cDNA sequences, a hierarchical clustering of the cell lines showed that lines with a common tissue of origin, clustered together. This idea has since formed the basis of several expression profiling diagnostic tools for breast cancer.

The ability to query several thousands of genes (and if required the entire set of genes in the human genome) simultaneously is on account of DNA micro-array technology, in which probes for individual

genes are spotted on to a glass slide with great precision. When cDNA molecules that have been tagged with a fluorescent dye such as Cy3 or Cy5 hybridize with their respective probes, a visual indication of the genes represented in a cDNA sample and their relative amounts is obtained. Pioneering work on developing DNA micro-array technology was performed by Patrick Brown's group at Stanford university (Schena et al.).

Gene expression analysis methods, can be broadly classified into two types namely, supervised and unsupervised. Unsupervised analysis leads to a clustering of similar genes and thereby reveals higher order relationships in the data. Clustering may be done in a hierarchical manner or by simply partitioning the data space into flat clusters (with no sub-relationships between the clusters). According to a review of various clustering techniques employed in gene expression analysis, hierarchical agglomerative clustering is probably the most widely used technique (Eisen et al.). In addition to clustering, analysis of expression data also employs ordination techniques such as principal component analysis, that allow conversion of expression data into a smaller set of linearly uncorrelated data, thus reducing dimensionality.

Supervised methods attempt to classify the expression data on the basis of prior class information. The classifier can then be used to predict the class of future data. Ensuring that a classifier does not over-predict requires among other things, careful selection of training and test data. Supervised analysis can be used to find gene groups that have significantly different expression levels. It can also be used to identify a set of genes that correlate well to a particular property of the sample and thus act as a signature with a useful predictive value.

As the usefulness of gene expression analysis and it's potential applications are getting established, there has been a rapid growth in the volume of publicly available gene expression data, for example the Gene Expression Omnibus (GEO) from NCBI and the ArrayExpress from the European Bioinformatics Institute. It is also recognized that the results of expression analysis depend on the method employed. The Critical Assessment of Microarray Data Analysis (CAMDA) is a forum that encourages participants to analyze a standardized data set within a competitive agenda (in recent, the goals of CAMDA have expanded to include other kinds of large data sets and it is now known as the Critical Assessment of Massive Data Analysis). Such comparative analyses reveal that flagging and filtering of outlying data, normalization and other per-processing steps are essential to carrying out a micro-array analysis that is repeatable. As a consequence, it becomes important to have standard protocols for expression analysis so that results across studies can be interpreted and analyzed within a common framework.

Applications in cancer decision support systems: Known applications of gene expression analysis pertain to aiding diagnosis, assessment of prognosis and predicting the response to a therapy (Duffy et al.). In subsequent sections, examples from each of these three areas of application have been discussed.

A well known example of gene expression aided cancer diagnosis is the identification of AML cells from ALL cells based on analysis of about 6800 genes from 38 patients (Golub et al.). This was one of the earliest studies on gene expression based classification of cancer cells and even though differentiating these two kinds of cancers through traditional techniques is not a difficult problem, the

study illustrated the potential of this technique to more difficult cases, such as when the cells present a more atypical morphology and defy correct diagnosis through conventional means. The technique has also shown success in distinguishing sub-types with a class of cells known as small round blue cell tumors (SRBCTs). The tumor, which may be a neuroblastoma, rhabdomyosarcoma, Ewing's tumor, a non-Hodgkin lymphoma or a Burkitt lymphoma, is hard to classify and thus presents a challenge for determining the correct therapy as well as prognosis. By analyzing 6500 genes from micro array data, followed by PCA based dimensionality reduction, Khan et al. were able to arrive at set of 96 genes that could correctly classify the SRBCT samples.

A second example of an assisted diagnosis application is the identification of metastasis of unknown primary. In such a case the the primary site can not be discerned by the use of clinical examinations, standard pathological tests (routine examinations of blood and urine, X-rays) or histological tests. By using gene expression data from 100 primary cancers as a training set, and a another 75 sets from blinded primary and metastatic cancers, Su et al. developed a classifier that could predict primary sites with an accuracy of 95% in test data. Later a similar approach was used by Giordano et al. to correctly classify adenocarcinoma samples that had metastasized from the lung, colon and ovary.

By using survival rate, time to progression or other phenotypic indicators of prognosis, the same technique may be used to estimate prognosis in cancer patients. Applying this to 96 lymphocyte samples from diffuse large B cell lymphomas (DLBCLs), Alizadeh et al. were able to measure the expression of 17856 genes using a custom made micro-array and this revealed a sub set of tumors with a distinct expression signature but with no visible histo-pathological differences from the rest of the

tumors. A hierarchical clustering of the tumors based on germinal B-cell genes revealed two sub types that each had oppositely regulated genes in their expression signatures. These two sub-types (germinal center B-like DLBCL and activated B-like DLBCL) are known to have very different survival rates in patients and thus warrant completely different approaches to treatment.

Gene expression profiling has also been used to assess the prognosis for a given type of cancer. MammaPrint™ is a FDA approved, commercially available recurrence assay (according to Agendia, the agency that is licensed to sell this test, it is the first such test to obtain FDA approval) that predicts the outcome independent of other factors influencing prognosis (such as estrogen receptor status). It is based on the work of researchers at the Netherlands Cancer Institute that examines a set of 70 genes, believed to constitute a prognosis signature for breast cancers. Their approach is to develop a supervised classifier that examines a set of informative gene expressions from a tissue sample and then classify the tumor as either a good prognosis or a bad prognosis tumor. In their study of 98 primary breast cancer patients, who were all less than 55 years o age and lymph node negative, they identified 34 patients who developed distant metastases within 5 years and 18 who remained disease free for at least 5 years. These were patients who had sporadic tumors; in addition to these there were 18 patients in the study who had BRCA1 germ line mutations and 2 who had BRCA2 mutations. By using a micro array capable of identifying 25,000 human genes and provide a quantitative assessment of the gene expression (through the intensity fluorescence), they were able to identify 5000 genes that were significantly regulated in these samples (a two fold difference against the average reference signal, $P<0.01$ in at least 5 tumors). By using a hierarchical clustering algorithm, they clustered each of the 5000 genes across the 98 tumors and again, each of the 98 tumors across the 5000 genes. The results of

the cluster analysis were plotted on a dendrogram (Figure 1). It was observed, that even with unsupervised clustering the 98 tumors separated into two broad groups. The larger group of 62 tumors (plotted on the upper part of the dendrogram) were found to have fewer instances of distant metastases (only 34%) compared to the smaller group of 36 tumors (70% had distant metastases within 5 years). Similarly, examining the clustering of genes (along the horizontal axis), it was observed that ESR1 along with several other co-regulated genes cluster together (left hand side) while genes associated with lymphocytic infiltration are present in another cluster (right side). Examining the estrogen receptor (ER) status of the tumors, of the 39 ER negative tumors, 34 were found in the lower cluster. This was also noticed for the BRCA1 tumors (16 out of the total 18 are in the lower cluster).

From this initial discovery based on a non-supervised clustering, the study sought to identify a prognosis signature, that could reliably identify lymph node negative patients that have a high/low risk of relapse within 5 years. This was done by applying a three step, supervised classification of the 78 sporadic tumors. In the first stage, 5000 genes that were significantly regulated in more than 3 out of these 78 tumors were selected and a correlation coefficient of each of these with the disease outcome was calculated. There were 231 genes that had a significant correlation coefficient x ($x<-0.3$ or $x>0.3$) and in the next step, these were rank ordered. Finally, the prognosis signature set was constructed by adding 5 genes at a time from the top of the rank ordered list and testing the classifier performance using a leave-one-out cross validation approach, till the performance did not improve further. Seventy genes from the starting list of 231 made it to this prognosis signature set. An average good prognosis signature was calculated and the 78 tumors were rank ordered based on their correlation with this average profile. A plot of the 78 tumors versus the genes in the signature (ordered by their level of

contribution to the classifier) is show in in Figure 2. The classifier performance (specificity and sensitivity) can be adjusted by modifying the cut-off suitably; in such a test, one would expect a preference for high sensitivity, even at the expense of some specificity.

An interesting aspect of this study was to examine the functional annotations of the genes in the prognosis signature. It was found that genes involved in ell cycle, invasion, metastasis, angiogenesis and signal transduction were significantly up-regulated in the poor prognosis signatures. Some examples of these genes and their associated GO terms are presented in Annexure A. It is also interesting to observe that while efforts to predict breast caner prognosis with individual genes have often shown contradictory results (for example, as reported by Bieche et al.), the MammaPrint™ prognosis, even though it does not have many of the genes known to be associated with breast cancer prognosis, provides a reliable prediction method. This is because it has greater predictive power, which it draws from simultaneously querying the full set of genes in the prognosis signature, as opposite to individually.

Finally, gene expression analysis can also be used to predict the response to a particular therapy for cancer. Certain types of prostrate and breast cancers are candidates for endocrine therapy. However the markers for endocrine therapy response in these cancer types lack specificity. For example, only 50% of the ER positive breast cancers respond to some form of endocrine therapy. A gene expression analysis of laser microdissected breast tumor samples (Ma et al.) revealed 9 genes that were differentially expressed in the hormone response and hormone resistive patients. Out of these 9, two exhibited differential expression in both the microdissected tissue samples as well as whole tissue

samples. These two genes (HOXB13 and IL17BR) have shown a higher specificity and sensitivity in predicting outcome following adjuvant tamoxifen therapy.

There appear to be relatively few instances of studies focused on discovering markers that reliably predict response to cytotoxic drugs. While the initial research in this direction was based on cell lines and on cancer xenografts in mice, the results of these studies had a limited applicability to human models. One such initiative (Holleman et al.) attempted to discover differentially expensive genes (in a panel of 14,500 genes) in drug sensitive and drug resistant acute lymphoblastic leukemia cells. The study found 172 genes to be differentially expressed in sensitive and resistant B-cell lineage leukemic cells (22 in the case of daunorubicin, 59 for vincristine, 42 for prednisolone and 52 for asparaginase). A hierarchical clustering was able to correctly assign the cases to the correct drug resistant category about 80% of the time. It was also observed that a combined gene expression profile (for all four drugs) was associated with an increased probability of relapse. Examples of other similar studies are available but diagnostic assays based on expression profiles that are regulator approved and predict response to a particular therapy could not be found as of this writing. Recent publications discussing the on on-going work in proteomic profiling indicate that this may hold greater promise for such diagnostic objectives. Possibly this is the direction that therapy response diagnosis will take in future.

Key issues: For gene expression analysis to be successful as a diagnostic tool in the clinical setting, certain important issues must first be addressed. One of the main problems requiring urgent attention is the lack of reproducibility. Gene expression assays show variability within the same assay, between assays on the same platform, and between assays across different platforms. The magnitude of variation

also increases in this order. Duplicate spots on the same micro array may correlate 95% or more, but this correlation goes down to 60-80% if the sample is split and compared across two different arrays (Churchill, 2002). Poor correlations are found when comparing expression results from different platforms such as spotted DNA arrays and Affymetrix. Approaches to contain this variation in results include the use of spot replicates within the same array, using dye reversal technique (Schulze and Downward, 2001), improved laboratory practices and appropriate statistical techniques in analysis. This is essential for clinicians to conduct their own studies and meaningfully compare results across different studies.

Equally important for inter-study and inter-platform comparisons among gene expressions is the availability and adoption of standards. A basic requirement of standardization across methods and platforms is the availability of a universal reference. In 2003, a consortium of industry representatives as well as research institutes working with expression technologies convened and drafted a set of standards for reference RNA material. This consortium (the External RNA Controls Consortium) had defined two standards namely the Assay Process Reference and the Array Hybridization Reference. The first (APR) is a pool of Human cRNA from several samples, the second is meant to be a control reference as it has no sequence similarity with Human RNA. From a survey of literature n the subject it appears that at present there is a multiplicity of similar external controls that have been developed by both for-profit and non-profit entities for instance Agilent technologies' spike-in set, Affymetrix's GeneChip poly-A control kit, National Institute of Aging/Agilent's spike-in set to name a few. Multiple standards negate the very purpose of establishing a standard and this could be a challenge for studies that focus on the same cancer or therapeutic problem but use different platforms for their work.

Another dimensions of variability in expression studies is related to the method of tissue handling which is known to contribute to large variations. This again is an unresolved problem because different laboratories follow different methods of collecting and storing tissue samples. Finally, as demonstrated in the CAMDA meetings, a standardization of analysis protocols that are used to report and compare gene expression study results is also needed in order to effectively leverage the global effort in this direction. The end objective of translational research in the area of gene expression analysis must be develop reproducible, simple and inexpensive methods that can empower clinicians to examine various kinds of cancer tissue and make reliable predictions.

Conclusion: The genesis, progression and prognosis of all cancers is driven by a sets of genes, each unique to a particular type of cancer. As Bert Vogelstein put it "The revolution in cancer research can be summed up in a single sentence; cancer is, in essence, a genetic disease". As a result, the genetic signature that a cancer produces, is a far more dependable marker of the identity and the characteristics of that cancer than histo-pathological markers that have been used traditionally in cancer diagnosis. Development of micro array technologies allows thousands of genes that are expressed in cancer cells to be examined for signatures of expression that may provide clinically relevant information about the cancer. Such signatures, in certain kinds of cancers have been shown to be more effective than histo-pathological examination in the identification of the cancer type, assessment of prognosis, as well as in predicting the response to certain therapeutic agents.

However, before micro array based gene expression analysis can be fully exploited to provide early

diagnosis and to chart out the optimal course of treatment, certain crucial issues will need to be resolved. These relate to the high variability in results, lack of reference standards and methods as well as the high complexity, costs and skills involved in conducting gene expression analysis. Nevertheless, this technique has already been implemented in several regulator approved assays that are commercially available today. This is an encouraging sign for its potential to buttress our efforts against cancer.

REFERENCES:

1. Alizadeh A.A. *et al*. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** (6769): 503–511, (2000).
2. Bieche, I. et al. Genetic alterations in breast cancer. Genes Chromosomes Cancer 14, 227±251, (1995).
3. Churchill, G. A. Fundamentals of experimental design for cDNAmicroarrays. Nat. Genet. 32(Suppl.): 490-495, (2002)
4. 'Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomized trials. The Lancet 352(9132): 930–942, (1998).
5. Eisen, M. B., Spellman, P. T., Brown, P. O. and Botstein, D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 95, 14863-14868, (1998).
6. Foo, X . Tamoxifen for early breast cancer: an overview of the randomised trials. The Lancet 351 (9114): 1451–1467, (1998).
7. Giordano, T. J., Shedden, K. A., Schwartz, D. R., Kuick, R., Taylor, J. M., Lee, N., Misek, D. E., Greenson, J. K., Kardia, S. L., Beer, D. G., Rennert, G., Cho, K. R., Gruber, S. B., Fearon,

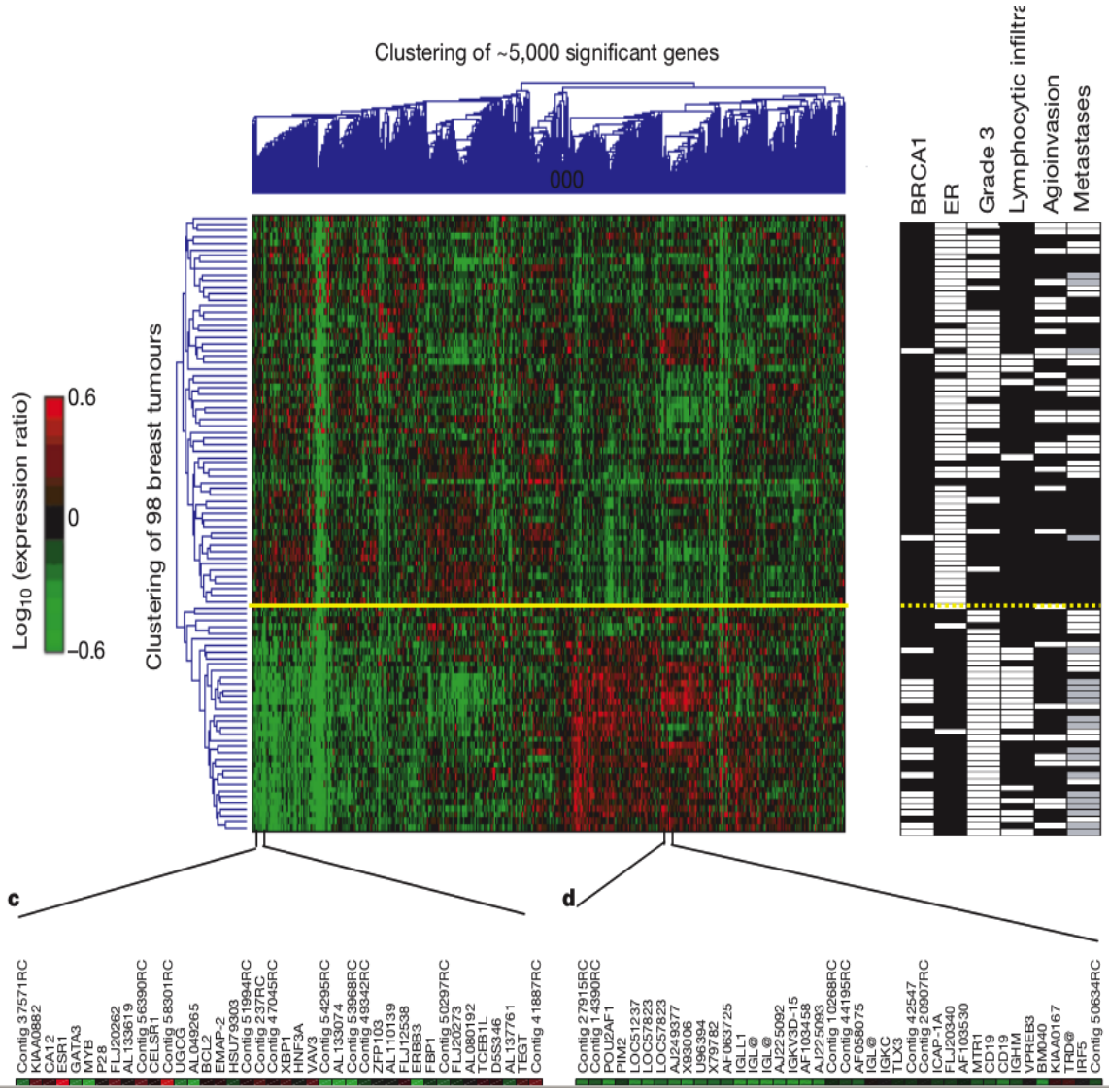
E. R. and Hanash, S. Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles. Am. J. Pathol. 159, 1231-1238, (2001).

8. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531-537, (1999).
9. Holleman, A. et al. Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment. N. Engl. J. Med. 351, 533-542 , (2004).
10. Khan, J. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7, 673-679., (2001)
11. Ma, X. J. et al. A two-gene expression DNA Microarray-Based Gene Expression Profiling in Cancer ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer Cell 5: 607-616, (2004)
12. Ramaswamy S. *et al*. A molecular signature of metastasis in primary solid tumors. Nature Genetics 33 (1): 49–54, (2002).
13. Ross D.T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics 24 (3): 227–235, (2000).
14. Schena, M.; Shalon, D.; Davis, R. W. and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270: 467-470., (1995).
15. Schulze, A. and Downward, J. Navigating gene expression using microarrays--a technology review. Nat. Cell Biol. 3, E190-E195, (2001).

16. Sørlie M. *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences 98 (19): 10869–10874, (2001).
17. Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F. Jr. and Hampton, G. M. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res. 61: 7388-9733, (2001).
18. van't Veer L.J. *et al.* "Gene expression profiling predicts clinical outcome of breast cancer". Nature **415** (6871): 530–536, (2002).
19. van de Vijver M.J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine 347 (25): 1999–2009, (2002).

Figure 1

Cluster analysis of 5000 genes from 98 breast cancer tumor samples



source: van't Veer L.J. *et al.*

Figure 2

A plot of the 78 tumors versus the genes in the signature (ordered by their level of contribution to the classifier)

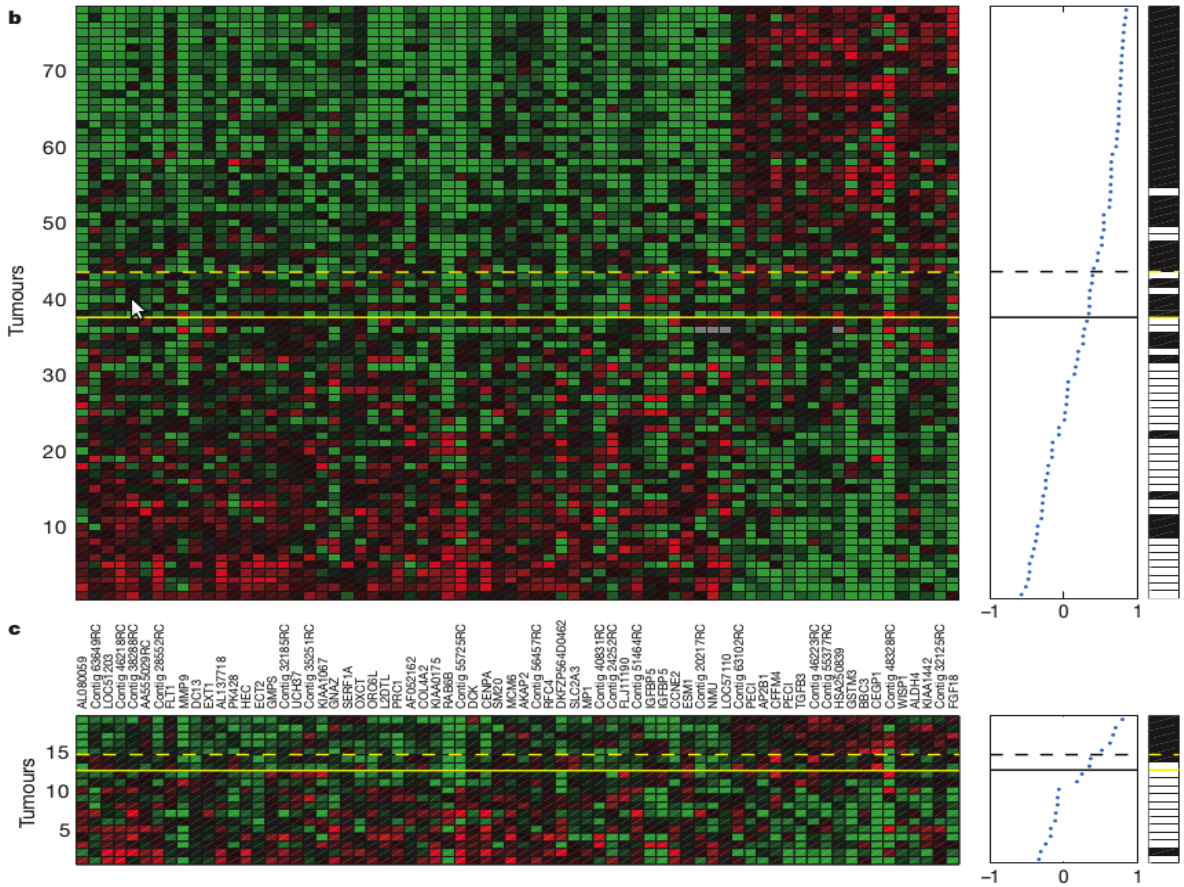
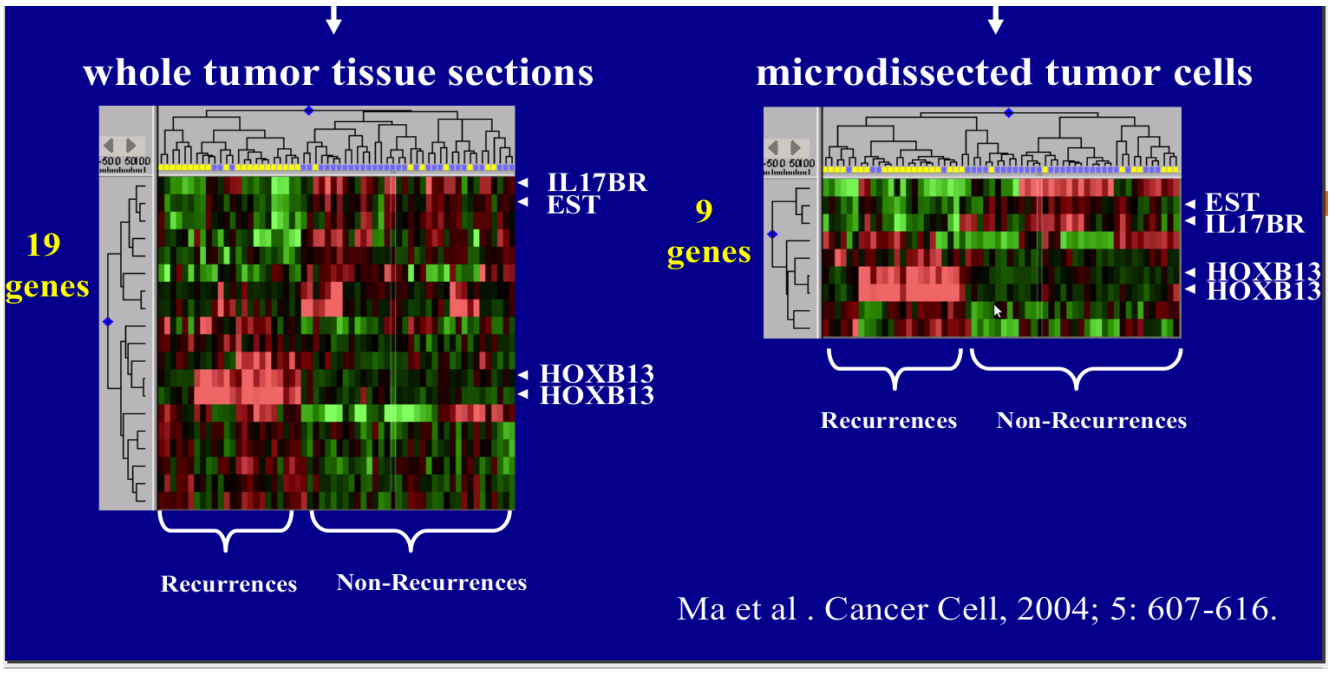


Figure 2 Supervised classification on prognosis signatures. **a**, Use of prognostic reporter genes to identify optimally two types of disease outcome from 78 sporadic breast tumours prognostic classifier with optimal accuracy; dashed line, with optimized sensitivity. Above the dashed line patients have a good prognosis signature, below the dashed line the

source: van't Veer L.J. *et al.*

Figure 3

Gene expression from of laser microdisected breast tumor samples in the microdisected tissue samples as well as whole tissue



Annexure A

A partial list of genes in the 70 gene prognosis signature set of the Netherlands Cancer Institute and their associated GO terms.

Gene	GO terms (SwissProt)
Cyclin E2	DN replication initiation, Cell division, regulation of cyclin dependent protein kinase activity.
MCM6	DNA strand elongation, S phase of the mitotic cycle
MMP9	Macrophage differentiation
MP1	Proteolysis
RAB6B	Small GTPase mediated signal transduction
ESM1	Regulation of cell growth, angiogenesis
FLT1	Positive regulation of cell proliferation. Positive regulation of angiogenesis.