BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 2 2005, pages 171–178
doi:10.1093/bioinformatics/bth469

Outcome signature genes in breast cancer: is there a unique set?

Liat Ein-Dor[1,†], Itai Kela[1,3,†], Gad Getz[1,†], David Givol[2] and Eytan Domany[1,*]

[1]Department of Physics of Complex Systems, [2]Department of Molecular Cell Biology and [3]Department of Immunology, Weizmann Institute of Science, Rehovot 76100, Israel

Received on June 4, 2004; revised on August 2, 2004; accepted on August 3, 2004
Advance Access publication August 12, 2004

ABSTRACT

Motivation: Predicting the metastatic potential of primary malignant tissues has direct bearing on the choice of therapy. Several microarray studies yielded gene sets whose expression profiles successfully predicted survival. Nevertheless, the overlap between these gene sets is almost zero. Such small overlaps were observed also in other complex diseases, and the variables that could account for the differences had evoked a wide interest. One of the main open questions in this context is whether the disparity can be attributed only to trivial reasons such as different technologies, different patients and different types of analyses.

Results: To answer this question, we concentrated on a single breast cancer dataset, and analyzed it by a single method, the one which was used by van't Veer *et al*. to produce a set of outcome-predictive genes. We showed that, in fact, the resulting set of genes is not unique; it is strongly influenced by the subset of patients used for gene selection. Many equally predictive lists could have been produced from the same analysis. Three main properties of the data explain this sensitivity: (1) many genes are correlated with survival; (2) the differences between these correlations are small; (3) the correlations fluctuate strongly when measured over different subsets of patients. A possible biological explanation for these properties is discussed.

Contact: eytan.domany@weizmann.ac.il

Supplementary information: http://www.weizmann.ac.il/physics/complex/compphys/downloads/liate/

INTRODUCTION

Several attempts were made to predict survival of cancer patients in general (Bair and Tibshirani, 2004; Beer *et al*., 2002; Khan *et al*., 2001; Nguyen and Rocke, 2002; Rosenwald *et al*., 2002), and of breast cancer patients in particular (Ramaswamy *et al*., 2003; Sorlie *et al*., 2001; van't Veer *et al*., 2002) on the basis of gene expression profiling. Sorlie *et al*. (2001) used an unsupervised approach, hierarchical clustering, to assign breast carcinoma tissues to one of five different subtypes, each with a distinctive expression profile. Robustness of these survival-related subclasses was demonstrated (Sorlie *et al*., 2003) by applying the same analysis procedure to two independent breast carcinoma datasets (van't Veer *et al*., 2002; West *et al*., 2001). van't Veer *et al*. (2002) applied a supervised approach to identify a gene expression signature, based on 70 genes, capable of predicting a short interval to the development of distant metastases. First, they randomly selected a set of 78 patients, a training set, which was used to measure the correlation between each gene's expression and disease outcome. The genes were ranked according to this correlation, and the 70 most-correlated genes were used to construct a classifier discriminating between patients with good- and poor prognosis. The remaining 19 patients served as the test set to validate their prognosis classifier. A follow-up study (van de Vijver *et al*., 2002) proved the efficiency of this classifier as a survival predictor on a large set of 295 tumor specimens. In a third study, Ramaswamy *et al*. (2003) identified a set of 128 genes separating metastases from primary tumors. A refined set of 17 metastases-associated genes were tested on a large diverse set of primary solid tumors, and were found to successfully distinguish patients with good versus poor prognosis.

The predictive success of these studies was frustrated by the fact that the sets of survival-related genes identified by these three studies had only a few genes in common. Only 17 genes appeared in both the list of 456 genes of Sorlie *et al*. (2001) and the 231 genes of van't Veer *et al*. (2002); merely 2 genes were shared between the sets of Sorlie *et al*. (2001) and Ramaswamy *et al*. (2003) Such disparity is not limited to breast cancer but characterizes other human disease datasets (Lossos *et al*., 2004) such as schizophrenia (Miklos and Maleszka, 2004).

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

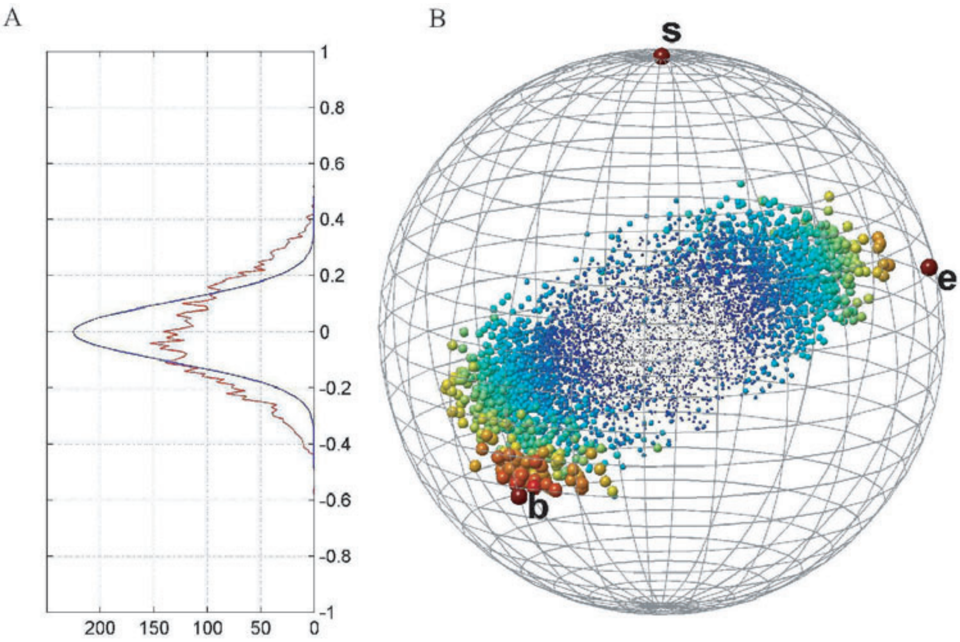

Fig. 1. (**A**) The histogram of the genes' correlation with the real survival vector (projection onto the vertical **s** axis—red curve), and with a random permutation of the survival vector (blue curve). (**B**) Globe of genes in the 'world' spanned by the normalized survival (**s**), BUB1 (**b**) and ESR1 (**e**). The survival is located at the north pole, while BUB1 (chosen from a large cluster of genes characterized by negative correlation with survival) and ESR1 (chosen from a large cluster of genes characterized by positive correlation with survival) are on the sphere's surface and their relative locations are determined by their angles with survival and with each other. All other (normalized) genes are represented by spots whose size and color illustrate how close the gene is to the surface (large red spots are close and small blue are far). The genes create an elongated structure at an angle $<\pi/2$ with **s**, implying that a large number of genes exhibit non-vanishing correlations with survival.

In this work, we explore this surprising phenomenon, and suggest new explanations for the lack of agreement between the sets of genes.

MATERIALS AND METHODS

Public dataset

The data van't Veer *et al*. (2002) contain gene expression profiles of primary breast tumors, from 96 sporadic young patients with grade T1/T2 tumors <5 cm in size, and N0 (no lymph node metastases). Of the 96 sporadic patients, 34 were treated by modified radical mastectomy and 62 underwent breast-conserving treatment, including axillary lymph node dissection followed by radiotherapy. Hybridization ratios were measured with respect to a reference made by pooling equal amounts of cRNA from all the sporadic carcinomas, on microarrays containing 25 000 human genes (Hughes *et al*., 2001).

Preprocessing of data

The full expression matrix of van't Veer *et al*. (2002) had 24 481 rows (genes) and 117 columns (samples). We applied filtering criteria, based on the entire set of 117 samples, yielding 5852 genes that exhibited a 2-fold change of expression with a P-value < 0.01 in five or more samples [van't Veer *et al*. (2002) applied the same filtering criteria on 98 samples, while discarding the test set of 19 samples, yielding 5000 genes]. We discarded from the set a single sample (sample 54) that contained >20% missing values [van't Veer *et al*. (2002) decided to include this patient in their analysis]. Like van't Veer *et al*. (2002) we also based our analysis on 96 'sporadic' patients free of BRCA1/2 germ line mutations.

Correlation analysis

For each gene, we test the null hypothesis that its gene expression profile is uncorrelated with the survival vector (over all 96 samples). We randomly permuted the survival vector (10^5 times) and calculated the correlation of the expression of each gene with the randomized survival vector. The P-value is the fraction of times one gets an absolute correlation larger or equal to the absolute correlation of the unpermuted data. Correction for multiple comparisons was performed using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995). Bounding the expected FDR by 10% yielded a list of 1234 genes for which the null hypothesis can be rejected. Histograms of the correlation (measured for 5852 genes) with the true survival and with a randomly permuted survival vector, are shown in Figure 1A.

Dividing the data into ten different divisions of 77/19

To examine how different experiments of 77 samples influence the composition of the 70 most-correlated genes with survival, we used the bootstrapping method (Tibshirani, 1993). Bootstrapping is a computer simulation enabling the overcoming of finite size effects. It assumes that the sample is a good approximation of the population. By generating a large number of new samples from the original sample sets, we can estimate the statistical parameters of the population. To keep the good/poor prognosis ratio of the original training set (33/44) we divided the 96 samples into a poor prognosis set of 45 samples, and a good prognosis set of 51. We chose with repetitions a random set of 33 samples from the poor prognosis set, and 44 from the good prognosis. We repeated this procedure ten times and found the top 70 genes for each 'training set' composition.

Measuring the STD of a gene based on a sample size of 77

We assumed that the degree of the polynomial fit for the average STD curve (Fig. 5) is the degree of the polynomial fit to the STD curve of each individual gene. Using this assumption, we found the polynomial fit to the STD curve of each gene in the data, and used it to estimate their STD values in a sample size of 77.

RESULTS

Many genes are related to survival

As was mentioned before, several microarray studies yielded gene sets whose expression profiles successfully predicted survival in breast cancer. However, the overlap between these gene sets was almost zero. This lack of agreement can be attributed to different chips, different methods of sample preparation, mRNA extraction and analysis of the data and, most importantly, to genuine differences between the patients (tumor grade, stage, etc.). To eliminate these sources of variation, we focused on data from a single experiment (van't Veer *et al*., 2002). The data consist of 96 samples and 5852 genes (see Materials and Methods). Disease outcome is represented by a survival vector **s**, of 96 binary components, with 1 representing good prognosis (metastasis-free time interval >5 years), and 0 representing poor prognosis (<5 years). The projection of the 96-dimensional expression vector of each gene onto a three-dimensional space [spanned by the survival vector (**s**) and the expression vectors of ESR1 (**e**) and BUB1 (**b**)] is shown in Figure 1B.

We chose to use ESR1 and BUB1 as representative genes of two large clusters characterized by positive- and negative correlation with survival, respectively.

The 5852 genes comprise an oblate spheroid shaped cloud, tilted with respect to the equator. If survival is replaced by a random binary vector, the oblate spheroid cloud lies on the plane of the equator. Since the vertical component of each gene is the correlation of its expression with survival (Fig. 1A), this difference is a striking geometrical manifestation of the fact that the expression vectors of very many genes (1234—at an FDR of 10%, see Materials and Methods) are related to survival.

According to our model, if the experiment is repeated on a different group of patients (with the same clinical characteristics), the overall appearance of the new 'globe' will be quite similar, but the positions of individual genes will swarm around. This swarming will suffice to change drastically the relative ranking of the genes on the basis of their correlation with survival.

Many sets of 70 genes can be used to predict survival

This dataset is characterized by three main properties: first, many genes are correlated with survival; second, the differences between these correlations are small; and third, the correlation-based rankings of the genes depend strongly on the training set (shown later). These properties may indicate that the top 70 genes are not superior to others in predicting disease outcome. To test this hypothesis, we selected the same 77 patients (out of 78; see Materials and Methods, and van't Veer *et al*., 2002) and ranked all genes according to their correlation with survival. We used the 5852 genes to build a series of classifiers (following the method used by van't Veer *et al*., 2002), based on consecutive groups of 70 genes. For each classifier, we measured the training and the test error, and found seven other sets of 70 genes, producing classifiers with the same prognostic capabilities as those based on the top 70. The genes of some of these seven classifiers appeared way down in the correlation-ranked list; the 70 genes of the first classifier are ranked between 71 and 140; classifier 2, 141–210; classifier 3, 211–280; classifier 4, 281–350; classifier 4, 351–420; classifier 5, 421–490; classifier 6, 561–630; classifier 7, 701–770. The location of these seven sets on the globe and their predicting performance is shown in Figures 9 and 11, respectively (see Supplementary information), and the corresponding Kaplan–Meier plots are shown below (Fig. 2).

To ensure that the aforementioned phenomenon is not unique to the specific training and test sets selected by van't Veer *et al*. (2002), we repeated the procedure described above for 1000 different compositions of training sets (of 77 samples) and test sets (19 samples). Each training set was used to rank the genes, and for each case the sequence of classifiers described above was constructed, and the training and test errors were measured for each classifier. Note, that when repeating this procedure for a randomized survival vector, the training error curve fluctuates around 37.5 mistakes (50% rate of errors) while the test error fluctuates around 9.5 mistakes, independent on the genes' rank. The results shown in Figure 3 imply that indeed, for each of the training sets, classifiers based on very-low-ranked genes are capable of predicting survival with quality similar to the high-ranking ones.

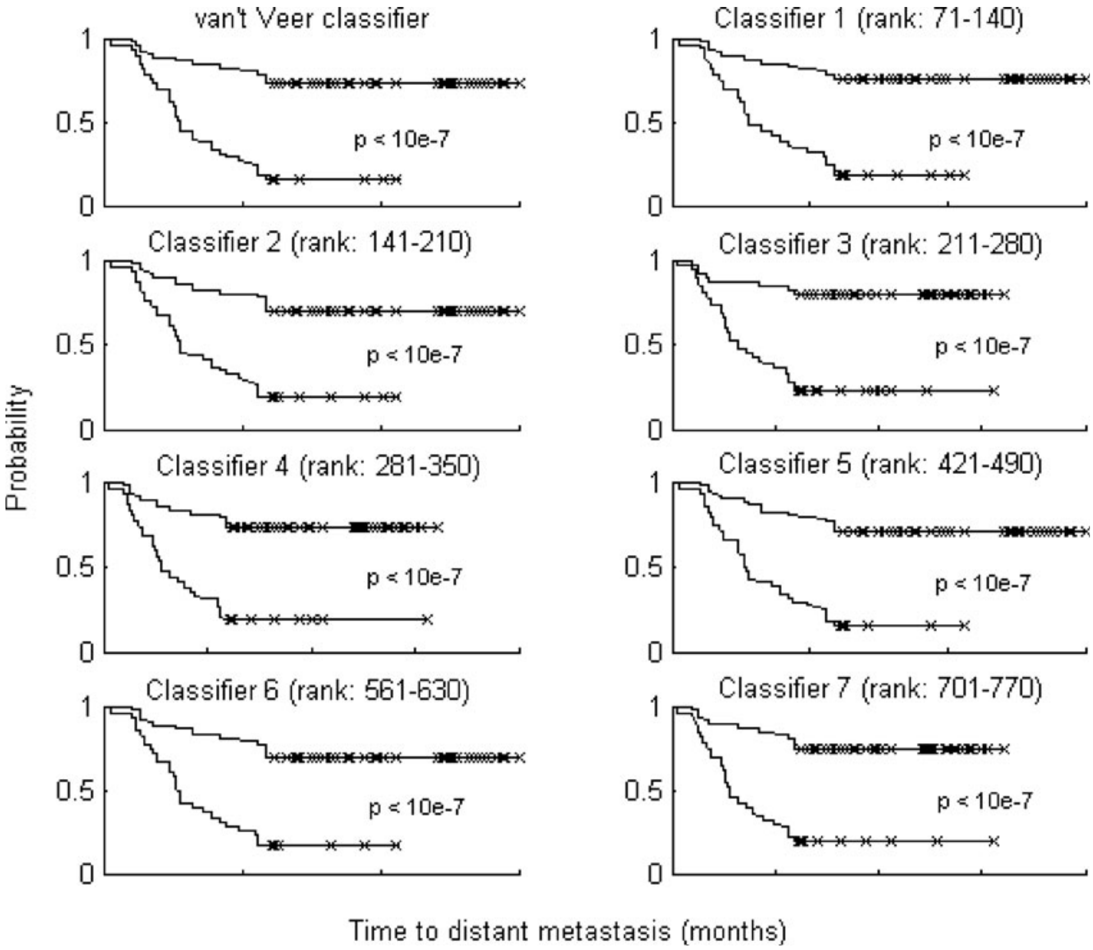

Fig. 2. Kaplan–Meier analysis of van't Veer *et al.*'s classifier and of the seven alternative classifiers as obtained from classifying all 96 samples. Upper curves describe the probability of remaining free of metastasis in the group of samples classified as having a good prognosis signature, while the lower curves describe the poor prognosis group.

To give a quantitative meaning to this claim, we generated the histogram presented in the inset of Figure 3, which shows that $>$70% of the 1000 training sets produced at least one classifier with the same (or better) performance as the one based on its own top 70 genes. The average number of such classifiers is four. The surprising summary of these observations is that (1) the list of the 'top 70 genes' of highest correlation with survival depends strongly on the training set of (77) patients on which the correlation was measured and (2) even with a fixed training set, one could have easily singled out a different group of 70 much lower ranked genes with as good a prognostic performance as that of the top-ranked genes.

Our results imply that although the top 70 genes may provide good prediction, other groups of 70 genes may do the same. Hence, these 70 genes cannot be considered as the main candidates for targeting anti-cancer treatment. Such candidates should be selected from the much longer list of genes related to survival, as demonstrated by the following list of cancer-related genes, present in the seven classifiers mentioned above. We list several of these genes, and indicate next to each one its correlation rank (in parentheses) measured on the training set selected by van't Veer *et al.* (2002).

Negative correlation with survival IL-6 (rank $=$ 502) is anti-apoptotic, and therefore supports tumor survival (Lotem *et al.*, 2003); CDC25B (402) (Nilsson and Hoffmann, 2000), CKS2 (297) (Urbanowicz-Kachnowicz *et al.*, 1999), CDC2 (229) (Winters *et al.*, 2001) and CDC20 (341) (Singhal *et al.*, 2003) are known to function in cell cycle regulation or DNA replication; oncogenes NRAS (260) (Boon *et al.*, 2003) and EZH2 (92) (Varambally *et al.*, 2002) enhance cancer aggressiveness.

Positive correlation with survival It may be caused by some indirect relation to tumor growth, affecting survival through indirect mechanisms like immunity, apoptosis or inhibition of oncogenes. Examples: BIN1/AMPH2 (477) by binding to MYC functions as a tumor suppressor (Sakamuro *et al.*, 1996); BIK (342) is pro-apoptotic (Li *et al.*, 2003) via binding to BCL2 (1106) (Li *et al.*, 2003). The positive

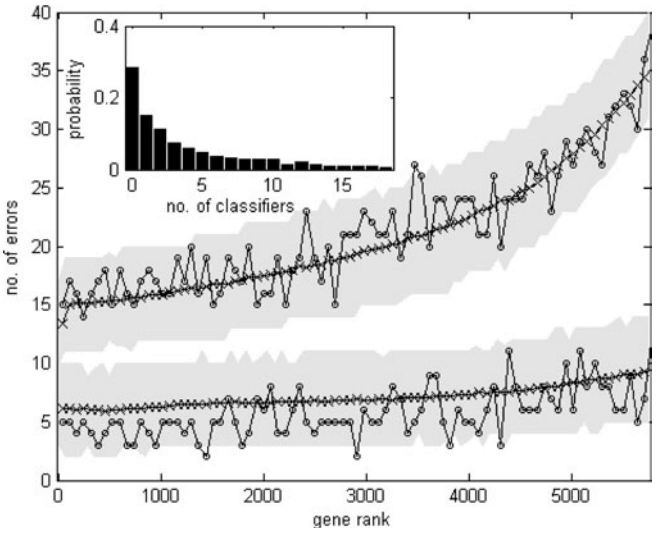

Fig. 3. The average performance of a series of classifiers generated by consecutive sets of 70 genes. The fluctuating curves present the number of errors produced by the classifiers resulting from one particular selection of training and test sets (upper, training errors out of 77 samples; lower, test errors out of 19). The x-axis represents the rank of the genes in the classifiers. The average over 1000 partitions is plotted as black x's; the two gray areas are the 95% confidence intervals of the training and test errors. Inset: histogram of the number of classifiers whose training and test errors are at least as low as those of the first classifier (based on the 70 genes with highest correlation to survival). Of the 1000 partitions, for ~28% no such classifier was found, whereas for ~6% five were found. Note that >70% of the training sets produce at least one classifier with the same performance as the top 70 genes; the expected number of such classifiers is around 4.

correlation of FLT3 (220) is due to its strong effect on dendritic cells and T-cells to enhance anti-tumor immunity (Ciavarra *et al.*, 2003). BRAK (237) is highly expressed in all normal tissues but low in malignant cells (Hromas *et al.*, 1999); IGFBP4 (225) induces apoptosis (Byron and Yee, 2003; Zhou *et al.*, 2003). Expression of GATA3 (255) is highly correlated with ER status (Bertucci *et al.*, 2000). Similarly, MYB (285) is also positively correlated with breast cancer outcome since it is a target of ER (Bertucci *et al.*, 2000; Guerin *et al.*, 1990) which is positively correlated with outcome. None of the genes listed above is ranked among the top 70.

Note that as opposed to claims made in (Gruvberger *et al.*, 2003), the success of the classifier is not due to the correlation of outcome to ER status. Creating a dataset which lacks this correlation, our seven classifiers, as well as van't Veer *et al.*'s (2002), kept their prognostic capabilities (see Supplementary information).

A gene's rank may fluctuate

Say we measure the correlation r of a gene's expression with survival on the basis of a sample of N patients drawn at random from a larger group with similar clinical characteristics. If a different set of N is drawn, the correlation will be different. If these statistical fluctuations of r are sizeable, they may change the ranking of a gene from high in one sample to a much lower rank in another; the smaller the N, the larger the fluctuations of r. In order to estimate the effect of these fluctuations on the composition of gene lists such as those of van't Veer *et al.* (2002), we repeatedly selected different subgroups of 77 samples out of the 96 (in each group we maintained the overall good/poor prognosis ratio) and for each subgroup identified the 70 genes that have the highest correlation with survival. The significant variation of the membership of the top 70 genes is clearly shown in Figure 10 of the Supplementary information. Note that every pair of these training sets has at least 58 samples in common, which significantly reduces the fluctuations of r and variation of the genes' ranks. In spite of this, the average overlap between two such gene groups is only 33.7/70. To better estimate the 'true' fluctuations of r for independent subgroups of 77 we used bootstrapping (Tibshirani, 1993), drawing subgroups from the 96 samples with repeats (see Materials and Methods). This reduces the expected overlap of two top 70 gene lists to 12.2/70. Figure 4 shows how large the variation of gene rank is, measured for 10 subgroups. Genes whose correlation with survival ranked high over one subgroup are likely to become low ranked in another. Hence, different sets of 77 patients, drawn from a clinically similar pool, will yield different lists of 'top 70 genes' with respect to correlation with survival.

Measuring the correlation fluctuations

In order to study how the fluctuations of the correlation with survival vary with the sample size K, we created n_K non-overlapping subgroups of size K from the 96 available samples. We calculated the correlation of each gene g with survival, measured over each subgroup, and from these n_K values we estimated the standard deviation (STD) of the correlation. We repeated this procedure five times (each time creating a different set of subgroups), to obtain $\sigma_g(K)$, the average STD, for each of the 5852 genes, for K ranging from 2 to 48 (the maximal K allowing for non-overlapping subgroups). Finally, we extrapolated the correlation noise (estimated by $\langle\sigma\rangle$, the STD averaged over the genes), from $K = 0$–96 (Fig. 5). As shown in Figure 5 the correlation noise decreases as the samples size increases. For sample size of 77 (the size of the training set), the expected average noise is ~0.1, whereas the significant genes found by van't Veer *et al.* (2002) and by our study show correlation between 0.3 and 0.5. In light of this small signal to noise ratio, the phenomenon shown in Figure 4 is not surprising.

Focusing on sample size $K = 77$ (Fig. 6), one can see that even relatively-low-ranked genes (around 1000), may have a non-negligible probability to be included among the 70 top ranked genes. Conversely, genes ranked among the top 70 can easily fluctuate to much lower ranks. The relatively low

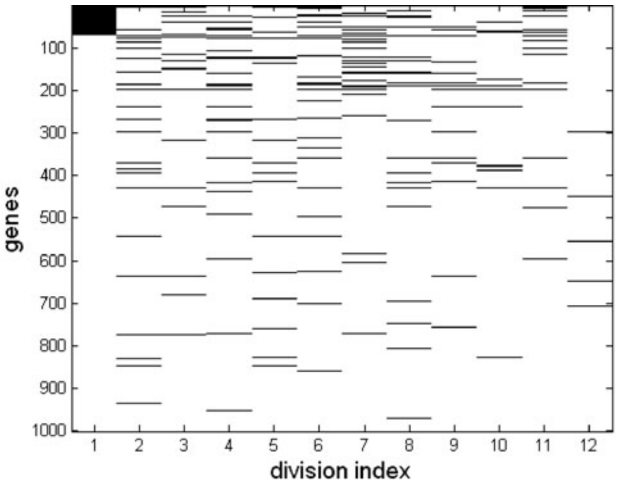

Fig. 4. Ten sets of top 70 genes, identified in 10 randomly chosen training sets of $N = 77$ patients (using bootstrapping—see Materials and Methods). Each row represents a gene and each column a training set. The genes were ordered according to their correlation rank in the first training set (leftmost column). For each training set, the 70 top-ranked genes are colored black. The genes that were top ranked in one training set can have a much lower rank when another training set is used. The two rightmost columns (columns 11 and 12) mark those of the 70 genes published by van't Veer *et al*. (2002) and the 128 genes appearing in (Ramaswamy *et al*., 2003) that are among the top 1000 of our first training set.

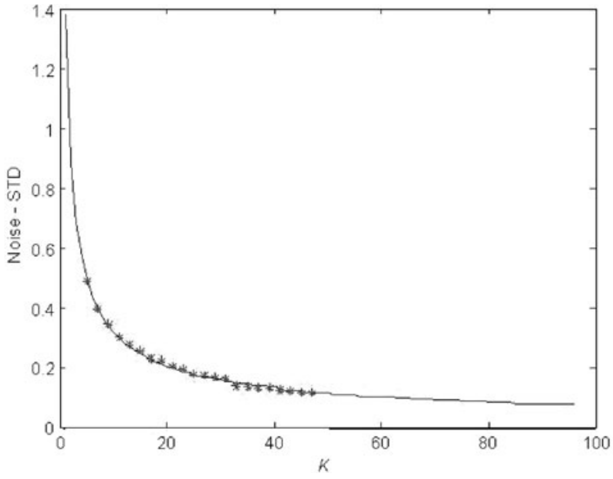

Fig. 5. Standard deviation (STD) of a gene's correlation with disease outcome, averaged over 5852 genes (y-axis) as a function of sample size K (x-axis). The curve is the polynomial fit to the results obtained for K between 2 to 48. This curve was used to extrapolate the STD to larger values of K. (The values extrapolated to $K = 77$ were used to calculate the error-bars presented in Fig. 6.)

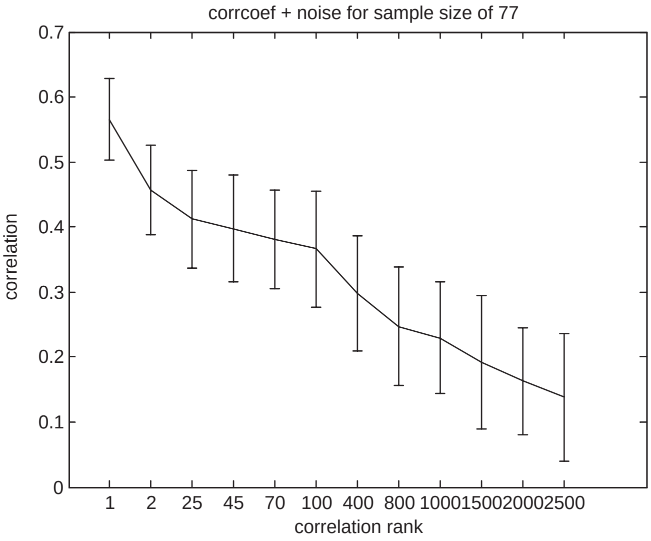

Fig. 6. Correlation of genes with survival versus their ranks. The correlation of each gene (y-axis) was measured based on the 96 samples, and the genes were ordered according to their correlation magnitude (x-axis). The error bars represent the noise (STD) of a gene based on sample size 77 (see Materials and Methods).

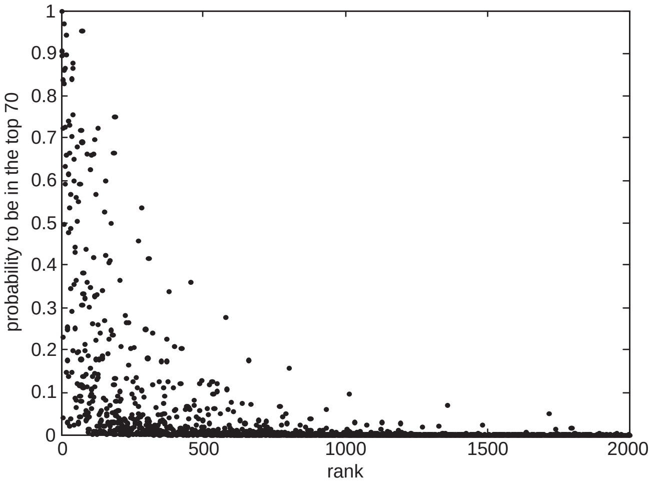

Fig. 7. The probability of genes to be included in a list of top 70. The genes were ranked on the basis of their correlation with outcome, as measured over the 77 samples of one particular (randomly chosen) training set.

signal-to-noise ratio explains the phenomenon demonstrated in Figure 4. In order to estimate the actual probability of each gene to be included in a list of top 70, we generated, at random, 10 000 training sets, each of 77 samples. For each such training set we identified the top 70 genes. The fraction of times (among 10 000) that each gene appeared in the top 70 is shown in Figure 7.

Taking into account the correlation noise, we defined an alternative gene score (instead of correlation coefficient), by calculating its probability to have a correlation above a given threshold for a given sample size (see Supplementary information). Figure 8 presents the probability of genes to

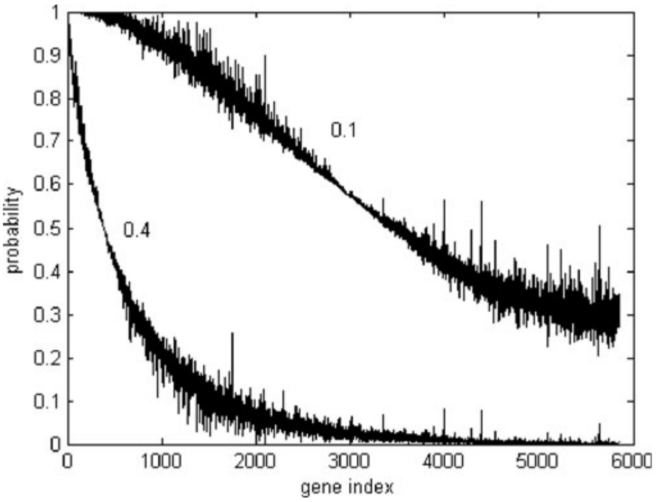

Fig. 8. The probability (y-axis) that genes have a correlation higher than a given threshold, calculated on the basis of noise derived for a training set of 77 samples. The x-axis represents the gene ranks according to their correlation with all 96 samples. The left curve corresponds to threshold of 0.4 and the right curve to threshold 0.1.

have a correlation higher than a given threshold (y-axis) calculated on the basis of the noise derived for a samples size of 77. The x-axis represents the genes' ranks according to their correlation coefficient with all 96 samples.

DISCUSSION

In this work, we investigated a single breast cancer dataset (van't Veer *et al*., 2002) in an attempt to explain the inconsistency between lists of survival-related genes derived from different experiments. While no single gene has a very high correlation with outcome, for many the correlation has intermediate values (Fig. 1). The differences between these correlation values are small, and the relative ranking of genes on the basis of correlation with survival changes drastically when a different training set is used. These large fluctuations in gene rank indicate that the identities of the top 70 ranked genes are not robust, and hence will not be reproduced in a different experiment. In spite of this sensitivity, the predictive power of several sets of genes is quite good. The main lesson is that whenever any arbitrary decision (e.g. choice of training and test set) is taken throughout analysis of the data, one has to generate a large ensemble of the different ways in which this arbitrary decision could be taken, and perform a statistical analysis of the results obtained over this ensemble. A high sensitivity of the results to the arbitrary decisions may indicate that the conclusions, e.g. the list of survival-related genes, are not unequivocal. In light of the inconsistency between lists of survival-related genes generated from the same dataset, the disagreement between lists obtained from different datasets is not surprising. A possible biological explanation for this may be the individual variations and heterogeneities associated with markers for outcome, even within a clinically homogenous group of patients.

Perhaps one has to divide the patients into smaller subgroups (Sorlie *et al*., 2003) on the basis of some yet unknown attribute and for each subgroup of tumors look for it's much sought 'primary, master genes' that control the metastatic potential. The correlations with survival of such a master gene may be very high in its own subgroup and low in others. The large fluctuations in the correlation of such a gene's expression with survival, measured over different training sets, are due to the fluctuating fraction of how many members of the gene's subgroup are in the training set. It is important to note that such a master gene will not necessarily be top-ranked with respect to correlation measured in a very large sampling of patients, composed of a mixture of subgroups.

Since one may need much larger numbers of patients to identify such survival-wise-homogenous subgroups and their associated, potential master genes, one should separate two issues: the quest for survival-related master genes and the construction of prognostic tools on the basis of a short gene list. One can produce fairly reliable prognostic tools; many genes are related to survival, and using a large enough subset of them will compensate for the fluctuations in the predictive power of individual genes for individual patients. Membership in a prognostic list, however, is not necessarily indicative of the gene's importance in cancer pathology. Rather, in order to study the potential targets for treatment, one must scan the entire, wide list of survival-related genes. By focusing only on those genes that were singled out from one dataset as its preferred prognostic tool, one may miss important key players, in breast and also in other types of cancer.

ACKNOWLEDGEMENTS

This work was supported by the Ridgefield Foundation, a Weizmann-Tel Aviv Sourasky Medical Center research grant, and by the NIH grant #5 P01 CA 65930-06.

REFERENCES

Bair,E. and Tibshirani,R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.*, **2**, E108.

Beer,D.G., Kardia,S.L., Huang,C.C., Giordano,T.J., Levin,A.M., Misek,D.E., Lin,L., Chen,G., Gharib,T.G., Thomas,D.G. *et al.* (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, **8**, 816–824.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, **57**, 289–300.

Bertucci,F., Houlgatte,R., Benziane,A., Granjeaud,S., Adelaide,J., Tagett,R., Loriod,B., Jacquemier,J., Viens,P., Jordan,B., Birnbaum,D. and Nguyen,C. (2000) Gene expression profiling of primary breast carcinomas using arrays of candidate genes. *Hum. Mol. Genet.*, **9**, 2981–2991.

Boon,K., Edwards,J.B., Siu,I.M., Olschner,D., Eberhart,C.G., Marra,M.A., Strausberg,R.L. and Riggins,G.J. (2003) Comparison of medulloblastoma and normal neural transcriptomes identifies a restricted set of activated genes. *Oncogene*, **22**, 7687–7694.

Byron,S.A. and Yee,D. (2003) Potential therapeutic strategies to interrupt insulin-like growth factor signaling in breast cancer. *Semin. Oncol.*, **30**, 125–132.

Ciavarra,R.P., Brown,R.R., Holterman,D.A., Garrett,M., Glass,W.F.,II, Wright,G.L.,Jr, Schellhammer,P.F. and Somers,K.D. (2003) Impact of the tumor microenvironment on host infiltrating cells and the efficacy of flt3-ligand combination immunotherapy evaluated in a treatment model of mouse prostate cancer. *Cancer Immunol. Immunother.*, **52**, 535–545.

Gruvberger,S.K., Ringner,M., Eden,P., Borg,A., Ferno,M., Peterson,C. and Meltzer,P.S. (2003) Expression profiling to predict outcome in breast cancer: the influence of sample selection. *Breast Cancer Res.*, **5**, 23–26.

Guerin,M., Sheng,Z.M., Andrieu,N. and Riou,G. (1990) Strong association between c-myb and oestrogen-receptor expression in human breast cancer. *Oncogene*, **5**, 131–135.

Hromas,R., Broxmeyer,H.E., Kim,C., Nakshatri,H., Christopherson,K.,II, Azam,M. and Hou,Y.H. (1999) Cloning of BRAK, a novel divergent CXC chemokine preferentially expressed in normal versus malignant cells. *Biochem. Biophys. Res. Commun.*, **255**, 703–706.

Hughes,T.R., Mao,M., Jones,A.R., Burchard,J., Marton,M.J., Shannon,K.W., Lefkowitz,S.M., Ziman,M., Schelter,J.M., Meyer,M.R. *et al.* (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.*, **19**, 342–347.

Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Li,Y.M., Wen,Y., Zhou,B.P., Kuo,H.P., Ding,Q. and Hung,M.C. (2003) Enhancement of Bik antitumor effect by Bik mutants. *Cancer Res.*, **63**, 7630–7633.

Lossos,I.S., Czerwinski,D.K., Alizadeh,A.A., Wechser,M.A., Tibshirani,R., Botstein,D. and Levy,R. (2004) Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *N. Engl. J. Med.*, **350**, 1828–1837.

Lotem,J., Gal,H., Kama,R., Amariglio,N., Rechavi,G., Domany,E., Sachs,L. and Givol,D. (2003) Inhibition of p53-induced apoptosis without affecting expression of p53-regulated genes. *Proc. Natl Acad. Sci. USA*, **100**, 6718–6723.

Miklos,G.L. and Maleszka,R. (2004) Microarray reality checks in the context of a complex disease. *Nat. Biotechnol.*, **22**, 615–621.

Nguyen,D.V. and Rocke,D.M. (2002) Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18**, 1625–1632.

Nilsson,I. and Hoffmann,I. (2000) Cell cycle regulation by the Cdc25 phosphatase family. *Prog. Cell Cycle Res.*, **4**, 107–114.

Ramaswamy,S., Ross,K.N., Lander,E.S. and Golub,T.R. (2003) A molecular signature of metastasis in primary solid tumors. *Nat. Genet.*, **33**, 49–54.

Rosenwald,A., Wright,G., Chan,W.C., Connors,J.M., Campo,E., Fisher,R.I., Gascoyne,R.D., Muller-Hermelink,H.K., Smeland,E.B., Giltnane,J.M. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

Sakamuro,D., Elliott,K.J., Wechsler-Reya,R. and Prendergast,G.C. (1996) BIN1 is a novel MYC-interacting protein with features of a tumour suppressor. *Nat. Genet.*, **14**, 69–77.

Singhal,S., Amin,K.M., Kruklitis,R., DeLong,P., Friscia,M.E., Litzky,L.A., Putt,M.E., Kaiser,L.R. and Albelda,S.M. (2003) Alterations in cell cycle genes in early stage lung adenocarcinoma identified by expression profiling. *Cancer Biol. Ther.*, **2**, 291–298.

Sorlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Sorlie,T., Tibshirani,R., Parker,J., Hastie,T., Marron,J.S., Nobel,A., Deng,S., Johnsen,H., Pesich,R., Geisler,S. *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.

Tibshirani,B.E.a.R.J. (ed.) (1993) *An Introduction to the Bootstrap*. Chapman and Hall, NY.

Urbanowicz-Kachnowicz,I., Baghdassarian,N., Nakache,C., Gracia,D., Mekki,Y., Bryon,P.A. and Ffrench,M. (1999) ckshs expression is linked to cell proliferation in normal and malignant human lymphoid cells. *Int. J. Cancer*, **82**, 98–104.

van de Vijver,M.J., He,Y.D., van't Veer,L.J., Dai,H., Hart,A.A., Voskuil,D.W., Schreiber,G.J., Peterse,J.L., Roberts,C., Marton,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Varambally,S., Dhanasekaran,S.M., Zhou,M., Barrette,T.R., Kumar-Sinha,C., Sanda,M.G., Ghosh,D., Pienta,K.J., Sewalt,R.G., Otte,A.P., Rubin,M.A. and Chinnaiyan,A.M. (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, **419**, 624–629.

West,M., Blanchette,C., Dressman,H., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,J.A.,Jr, Marks,J.R. and Nevins,J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

Winters,Z.E., Hunt,N.C., Bradburn,M.J., Royds,J.A., Turley,H., Harris,A.L. and Norbury,C.J. (2001) Subcellular localisation of cyclin B, Cdc2 and p21(WAF1/CIP1) in breast cancer. association with prognosis. *Eur. J. Cancer*, **37**, 2405–2412.

Zhou,R., Diehl,D., Hoeflich,A., Lahm,H. and Wolf,E. (2003) IGF-binding protein-4: biochemical characteristics and functional consequences. *J. Endocrinol.*, **178**, 177–193.