Pacific Symposium on Biocomputing 14:504-515 (2009)

MASTER REGULATORS USED AS BREAST CANCER METASTASIS CLASSIFIER[*]

WEI KEAT LIM, EUGENIA LYASHENKO, ANDREA CALIFANO[†]

Center for Computational Biology and Bioinformatics, Department of Biomedical Informatics, Columbia University, 1130 Saint Nicholas Avenue, New York, NY 10032, USA

Computational identification of prognostic biomarkers capable of withstanding follow-up validation efforts is still an open challenge in cancer research. For instance, several gene expression profiles analysis methods have been developed to identify gene signatures that can classify cancer sub-phenotypes associated with poor prognosis. However, signatures originating from independent studies show only minimal overlap and perform poorly when classifying datasets other than the ones they were generated from. In this paper, we propose a computational systems biology approach that can infer robust prognostic markers by identifying upstream Master Regulators, causally related to the presentation of the phenotype of interest. Such a strategy effectively extends and complements other existing methods and may help further elucidate the molecular mechanisms of the observed pathophysiological phenotype. Results show that inferred regulators substantially outperform canonical gene signatures both on the original dataset and across distinct datasets.

1. Introduction

A key application of genome-wide expression profile analysis is the identification of small gene signatures that effectively classify cancer sub-phenotypes with differential prognosis [1, 2]. For instance, both supervised and unsupervised methods have been extensively used to identify genes whose expression is predictive of patient outcome. In breast cancer, two recent large-scale Gene Expression Profile (GEP) studies identified independent ~70 gene panels that were 60-70% accurate in predicting progression to metastatic cancer in less than five years in their respective patient cohorts [3, 4]. However, paradoxically,

[*] This work is supported by the National Cancer Institute (R01CA109755), the National Institute of Allergy and Infectious Diseases (R01AI066116), and the National Centers for Biomedical Computing NIH Roadmap Initiative (U54CA121852).

[†] To whom correspondence should be addressed. Email: califano@c2b2.columbia.edu

there was virtually no statistical significance in the overlap between the two signatures (only one gene in common between them) and the signatures performed poorly in classifying samples from the other study. Such apparent paradox cannot be explained based only on technical reasons such as sample collection, microarray technology, and differences in data analysis [5]. Furthermore, such cases are quite the norm rather than the exception [6]. Thus the identification of robust prognostic biomarkers is still an open challenge in cancer research.

We argue that the fundamental reason why genes in prognostic signatures are so unstable and study-dependent is related to their role as “passengers” rather than “drivers” of the phenotypic differences (e.g., poor outcome). Since regulatory networks often act as amplification cascades, genes that are most differentially expressed tend to be further downstream from the somatic or inherited determinants of the prognostic differences. Due to the complex combinatorial interplay of regulatory proteins in the cell, these downstream genes are also the most unstable, as many co-factors and potential noise sources are involved in the transcriptional cascade that leads to their differential expression. Indeed, oncogenes and tumor suppressors are not generally the most differentially expressed genes although they may show an outlier behavior in some samples [7]. Thus, rather than looking for differentially expressed genes between two phenotypes of interest, we argue that one should look instead for regulators that are causally responsible for the implementation of the observed differential expression patterns. This was previously suggested [5, 8] but never implemented because accurate, genome-wide maps of regulatory processes in tumor related human cells have been lacking.

Recently, we introduced and extensively validated ARACNe, an algorithm for the dissection of transcriptional networks that can infer the targets of transcription factors (TF) from microarray expression profile data. ARACNe was shown to scale up to the complexity of mammalian transcriptional networks [9], producing accurate networks that are cell-phenotype specific. The algorithm has been successfully biochemical validated in B and T cells [9, 10] and more recently in glial cells (manuscript submitted), showing an accuracy greater than 80% in identifying targets validated by Chromatin Immunoprecipitation (ChIP) assays. Additional enhancements allow the algorithm to produce transcriptional maps that are fully directed by utilizing explicit knowledge both of which genes encode specific TFs and of their binding site when available [11, 12].

Here, we show that ARACNe inferred transcriptional regulation maps can be effectively interrogated for the unbiased inference of TFs that may induce or suppress specific gene signatures associated with poor prognosis in breast cancer. Furthermore, we show that Master Regulator is a better and more robust classifier than the large gene signatures proposed in [4] and [13].

2. Results

To identify and compare transcriptional regulators that are determinants of the gene expression signatures associated with rapid progression to metastasis in breast cancer, we used two datasets of 295 [3] and 286 [4] breast cancer samples each (hereafter denoted as the NKI and Wang datasets). The authors had previously used supervised analysis methods on these datasets to determine a 70-gene and a 76-gene prognostic signatures, respectively, (S_{NKI} and S_{Wang}). While these signatures were able to accurately classify the original datasets in cross-fold validation tests, using a Support Vector Machine (SVM) classifier [14], classification performance of the $\mathrm{S}_{\mathrm{NKI}}$ signature on the Wang data or of the $\mathrm{S}_{\mathrm{Wang}}$ signature on the NKI data was shown to be quite poor.

We performed the following analysis independently on the NKI and on the Wang dataset (For simplicity, we illustrate it only for the NKI data): (a) First, the NKI dataset was processed by ARACNe to infer a genome-wide transcriptional interaction network N_{NKI}. (b) Then, the network was interrogated using the Master Regulator Analysis (MRA), see Methods, to identify a repertoire of candidate TFs, i.e. master regulators MR_i, whose regulons R_i, (transcriptional target set) was highly overlapping with the signature S_{NKI}. (c) Finally, the MR_i genes were tested as classifiers

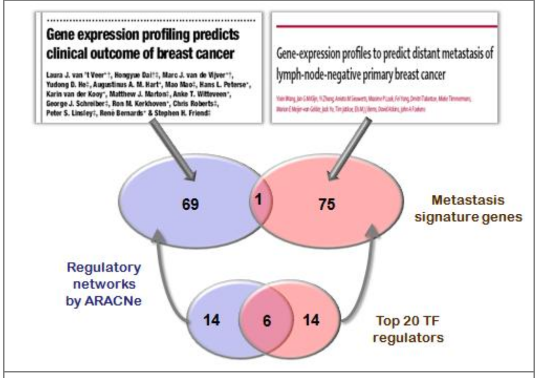
Gene expression profiling predicts clinical outcome of breast cancer
Laura J. van 't Veer*†, Hongyue Dai†‡, Marc J. van de Vijver*†, Yudong D. He‡, Augustinus A. M. Hart*, Mao Mao‡, Hans L. Peterse*, Karin van der Kooy*, Matthew J. Marton‡, Anke T. Witteveen*, George J. Schreiber‡, Ron M. Kerkhoven*, Chris Roberts‡, Peter S. Linsley‡, René Bernards* & Stephen H. Friend‡
Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer
69
1
75
Metastasis signature genes
Regulatory networks by ARACNe
14
6
14
Top 20 TF regulators

Figure 1. Graphical representation of the overlap of metastasis signature genes and the top 20 TF regulators between the two datasets.

Table 1 Top 20 Master Regulators inferred by MRA independently from NKI and Wang datasets.

GID_{NKI}	SYM_{NKI}	FET p-val	GID_{WANG}	SYM_{WANG}	FET p-val
9232	PTTG1	4.5×10^{-16}	7534	YWHAZ	4.9×10^{-5}
2305	FOXM1	6.0×10^{-11}	5933	RBL1	1.1×10^{-4}
7832	BTG2	4.1×10^{-6}	9232	PTTG1	5.3×10^{-4}
7534	YWHAZ	1.1×10^{-5}	7027	TFDP1	7.4×10^{-4}
7704	ZBTB16	1.2×10^{-5}	10736	SIX2	2.7×10^{-3}
4782	NFIC	1.4×10^{-5}	3148	HMGB2	4.1×10^{-3}
1054	CEBPG	1.6×10^{-5}	4208	MEF2C	5.0×10^{-3}
9735	KNTC1	1.9×10^{-5}	4605	MYBL2	5.7×10^{-3}
2737	GLI3	4.5×10^{-5}	4488	MSX2	6.2×10^{-3}
8522	GAS7	4.5×10^{-5}	1869	E2F1	6.7×10^{-3}
7027	TFDP1	5.9×10^{-5}	6095	RORA	1.3×10^{-2}
147912	SIX5	6.9×10^{-5}	2305	FOXM1	1.3×10^{-2}
1869	E2F1	1.2×10^{-4}	3607	FOXK2	1.4×10^{-2}
3608	ILF2	2.4×10^{-4}	6936	C2orf3	1.5×10^{-2}
4904	YBX1	2.4×10^{-4}	4603	MYBL1	1.5×10^{-2}
3223	HOXC6	5.2×10^{-4}	23054	NCOA6	1.7×10^{-2}
1746	DLX2	5.8×10^{-4}	7528	YY1	2.2×10^{-2}
51123	ZNF706	7.0×10^{-4}	677	ZFP36L1	2.2×10^{-2}
3148	HMGB2	8.6×10^{-4}	171017	ZNF384	2.3×10^{-2}
3066	HDAC2	9.3×10^{-4}	10520	ZNF211	2.4×10^{-2}

outcome against the S_{NKI} signature, both in the NKI and Wang set using five-fold cross-validation based on an SVM classifier.

Surprisingly, from an unbiased list of 852 TFs (determined by Gene Ontology annotation) shared by the two datasets, the analysis produced a 30% overlap among the top 20 inferred master regulator TFs ($p = 2\times10^{-6}$, by FET), which is especially significant when compared to the 1 gene overlap between the two signatures ($p = 0.25$, FET) (Figure 1). Note that both ARACNe and the MRA were performed independently on each of the two datasets. The top 20 TFs are listed in Table 1.

To study the stability of MR genes as a function of the training set, we repeated this procedure using different GEP subsets from the NKI dataset to infer prognostic signatures. We generated 100 sample sets by randomly selecting 34 out of 78 NKI samples with poor outcome and 44 out of 118 samples with good outcome in each training set, preserving the ratio of poor to good prognosis samples in the original study. We then computed the Pearson correlation between the prognostic groups

Table 2. Average overlap between S_{NKI} signatures and *top 20 Master Regulators* inferred from 100 GEP sample sets.

Predictor	Overlap
Signature genes	8.6%
Top 20 regulators	47.8%

and the expression level for all informative genes across the 78 samples [13]. The top 70 genes with the highest absolute correlation coefficient were selected as the corresponding S_{NKI}, and the top 20 Master Regulator TFs were identified using the MRA analysis. Finally, we compared the overlap of signature genes and regulators derived from each set to the ones from the original set, and reported the average overlap percentage in Table 2. Results clearly show that while the signature gene selection is highly affected by the training sample choice, with only 8.6% average overlap between signatures, the inferred master regulator TFs were much more consistent, with a 47.8% overlap despite the highly discordant signatures that were used to infer them. Although the causal role of the inferred regulators in the phenotype remains to be validated, this observation suggests that master regulators are less sensitive to training set selection and experimental variability.

In order to study whether the signature regulators possess the same or better prognostic capabilities than the signature genes, we used an SVM classifier to differentiate patients that developed distant metastasis within 5 years from patients that were metastasis-free for at least 5 years. We used both the inferred Master Regulators and the NKI/Wang signatures as the features for SVM-based classification. Gaussian Radial Basis Function kernel was employed in the SVM classifier along with the default parameters specified in MATLAB Bioinformatics Toolbox. Five-fold cross-validation was used to evaluate the performance in predicting the patients' prognostic group. Master regulators were tested independently, using lists of increasing size from the single best MR to a list of the top 70 MRs.

Two methods were used to assess the enrichment of TFs' regulon genes in genes that are differentially expressed between the two prognostic groups. The first one used the Fisher Exact Test (FET) to determine the statistical significance of the overlap between TFs' regulon genes and the prognostic signatures. The second used the Kolmogorov-Smirnov test, as implemented by the Gene Set Enrichment Analysis (GSEA) method, to assess enrichment of the TFs' regulon genes in

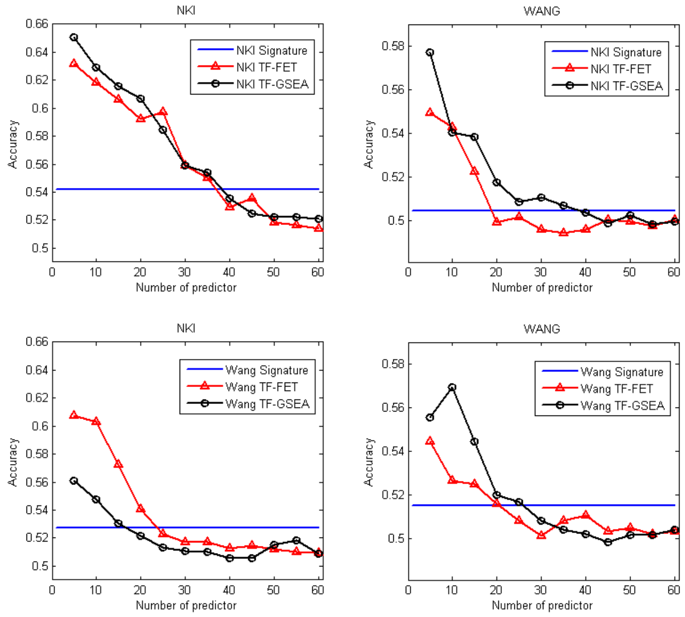

Figure 2: Accuracy of metastasis classification using signature genes and top TFs derived from (a) NKI dataset, and (b) Wang dataset. Predictions were made on NKI (left) and on Wang dataset (right) in both cases.

differentially expressed genes without having to select a specific threshold (see Methods). Both methods produced similar performance.

Results of prognosis classification for the SVM classifier using either (a) the S_{NKI} or S_{Wang} signature genes, (b) Master Regulator TFs ranked by FET p-value, and (c) Master Regulator TFs ranked by GSEA are shown in Figure 2. Since protein activity is not necessarily proportional to the mRNA concentration of the corresponding genes, due to post-transcriptional/translational modifications, we used the regulon of a TF as an indicator of the TF's activity. In particular, gene expressions were first converted into z-score. In each sample, all genes were sorted by the z-score and then used as the reference list in GSEA. The TF's activity level was approximated by the Normalized Enrichment Score (NES) computed for its regulon.

Classification performance was assessed by five-fold cross-validation. The analysis shows that even a handful of Master Regulators can predict progression to metastatic tumors with higher accuracy than the ≥70-gene signatures. Indeed, performance of the Master Regulator classifier invariably decreased as more genes were used. Classification accuracy using the gene-signatures drops when the S_{NKI} is used to classify Wang samples or vice-versa (i.e. in cross-dataset studies), while Master Regulator prediction remained highly predictive. The decrease in classification power as a function of the number of Master Regulators used in the classifier also suggests that the method produces informative ranking of the TFs', with more significant Master Regulators producing more accurate classification. In summary, the most significant master regulators were much better conserved across training sets (48% vs. 9%), consistently outperformed differentially expressed signatures in five-fold cross-validation studies, and were better able to classify samples in independent studies.

3. Methods

3.1. *ARACNe Network Inference*

ARACNe uses an information theoretic approach to dissect physical transcriptional interactions between TFs and their targets [12]. Briefly, the algorithm first uses a large GEP dataset to distinguish candidate interactions between a TF and other genes in the GEP by computing pairwise mutual information (MI). It employs a computationally efficient Gaussian kernel estimator to estimate the MI as:

$$I^*[X;Y] = \frac{1}{M}\sum_i \log\frac{f(x_i, y_i)}{f(x_i)f(y_i)} \tag{1}$$

where the 1-dimensional and 2-dimensional kernel estimators with kernel *K* and the position-dependent kernel width, *h*, are defined by Equations (2) and (3) respectively.

$$f(x) = \frac{1}{M}\sum_j K\left\{\frac{(x-x_j)^2}{h^2}\right\} \tag{2}$$

$$f(x,y) = \frac{1}{M}\sum_j K\left\{\frac{(x-x_j)^2 + (y-y_j)^2}{h^2}\right\} \tag{3}$$

ARACNe first eliminates interactions that are below a minimum MI threshold, defined based on statistical significance threshold. Then, the Data Processing Inequality (DPI) theorem from information theory is used to eliminate the vast majority of interactions that are likely mediated by another TF.

Bootstrap sampling, a nonparametric technique for statistical inference, is employed in the networks reconstruction process in order to accommodate the noise in microarray data and the error in MI estimation. The sampling procedure generates bootstrap datasets that are obtained by randomly selecting samples with replacement from the original dataset. One hundred bootstrap datasets are used to create the *bootstrap networks*. Due to the sampling procedure with replacement, the generated bootstrap datasets will consist of some replicated samples, and thus increase MI values for all edges computed from these datasets. To avoid this artifact, an infinitesimal amount of uniformly distributed noise is added on the bootstrap samples so that all the repeats with identical values would produce a randomized order. A consensus network is then constructed by retaining edges supported across a significant number of the bootstrap networks.

3.2. *Gene Sets Enrichment Analysis*

GSEA uses the Kolmogorov-Smirnov statistical test to assess whether a predefined gene set (in this case the Master Regulator set) is statistically enriched in genes that are the two extremes of a list ranked by differential expression between two biological states [15]. The algorithm is very useful to detect differential expression of a set of genes as a whole, even though the fold-change may be small for each individual gene.

Since gene regulons include both TF-activated and TF-repressed genes, GSEA was extended to assess enrichment of two complementary gene sets against N ranked genes. For instance, suppose that is expected to be a Master Regulator capable of activating a signature. Then one would expect the TF-activated gene set to be enriched in genes that are upregulated, while the TF-repressed gene set should be enriched in genes that are downregulated. Standard GSEA would dilute the evidence supporting TFs that can function both as activators and repressors.

The extended GSEA, for two complementary gene sets, proceeds as follows: (a) Compute signal-to-noise ratio (S2N), the difference of means scaled by the standard deviation $(\mu_A\text{-}\mu_B)/(\sigma_A\text{-}\sigma_B)$, between phenotype A and B, for each of the N genes in microarray. Order the N genes by S2N from the most positive to the most negative values, denoted by R; (b) identify hits independently for the positive gene set $S+$ in R, and the negative gene set S- in $\bar{R}$, in which $\bar{R}$ is the inversed ranking of R with the inverted S2N values; (c) Combine R and $\bar{R}$ and reorder the S2N values by keeping the hits for both *S*+ and *S*-, denoted as R_C (see Figure 3); (d) Compute a running score by walking down the combined ranking R_C. The score will increase by $|r_i|^p/\Sigma_{i\epsilon S}|r|^p$ if the i^{th} gene is a hit, or otherwise decrease by *1/(2N-S)*, where *S* is the combined total number of genes in *S*+ and *S*-. Finally, (e) Enrichment Score (ES) is determined as sum of the maximum and the minimum deviation from zero along the running score.

We randomly permuted the phenotype labels and repeated steps (a – d) for 10,000 times to compute the ES null distribution. Statistical significance of the ES can be computed by comparing the observed ES to the null distribution. A MATLAB function implementing the extended GSEA described above is available from http://www.dbmi.columbia.edu/~wkl7001/gsea2.m.

3.3. *Master Regulator Analysis (MRA)*

Given two phenotypes A and B and a transcriptional interaction network model, the Master Regulator analysis attempts to identify TFs that may induce the transition from A to B.

For each TF in the transcriptional interaction model, we test whether activation or inhibition of the TF may affect genes to produce a change

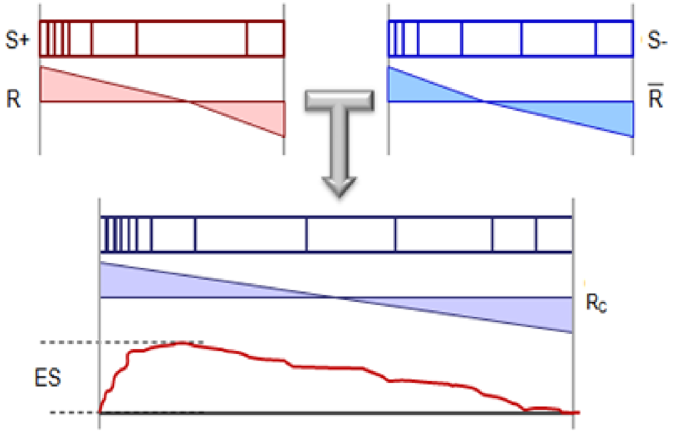

Figure 3. Extended GSEA assesses two complementary gene sets at once. Hits are identified independently for the positive gene set *S*+ in *R*, and the negative gene set *S*- in $\bar{R}$, in which $\bar{R}$ is the inversed ranking of *R* with the inverted S2N values. The rankings *R* and $\bar{R}$ are then combined and reordered according to the S2N values, denoted as R_C. All the identified hits, for both *S*+ and *S*-, are remained throughout the reordering process. ES is defined as sum of the maximum and the minimum deviation from zero along the running score.

as similar as possible to that observed between B and A. Specifically, we first assume that TF activation may lead to the transition. In this case, we test whether TF-activated targets (the R+ regulon of the TF) are enriched in the genes that are overexpressed in B and whether the TF-repressed targets (the R- regulon of the TF) are enriched in genes that are underexpressed genes in B. This is accomplished with one of two tests:

FET Analysis: In this analysis, genes that are differentially expressed in B vs. A are first filtered by a statistical significance threshold (e.g., 5% FDR by t-test or u-test) and then divided into two groups, S+ and S-, containing respectively genes that are overexpressed or underexpressed in B vs. A. The statistical significance p^+ of the intersection between R+ and S+ as well as the statistical significance p^- of the intersection between R- and S- is computed, given the total number of considered genes, using the FET test. Since the two tests are statistically independent, the final p-value is assessed as $p = p^+ \times p^-$.

Gene Set Enrichment Analysis: In this test, rather than selecting gene signatures S+ and S-, the GEP genes are ranked in a list $L_{A\rightarrow B}$, from the most underexpressed to the most overexpressed in B vs. A. Then, the extended GSEA analysis is used to simultaneously assess whether R+ and R- are respectively enriched in genes that are overexpressed and underexpressed in $L_{A\rightarrow B}$. The procedure is then repeated to test whether TF silencing may lead to the transition. In this case, the role of R+ and R- is inverted but the procedure is not affected. Finally, Master Regulators are sorted by statistical significance (p-value).

4. Discussion

In this manuscript, we have presented an approach for identifying robust prognostic markers using transcriptional regulatory networks in the breast cancer context. Rather than establishing signatures of genes that are differentially expressed in poor prognosis vs. good prognosis samples, the method attempts to identify the upstream transcriptional regulators of the signature that are consistent with the network topology.

We have shown that top Master Regulator genes, independently inferred from different datasets or from different subsets of the same dataset, are far more stable and robust than top genes correlated to disease outcome. For instance, overlap of the top 70 correlated genes after subsampling was only 8.6% while overlap of the top 20 Master Regulators inferred from the same datasets was 46.8% on average.

Finally, we have shown that in independent testing using two large datasets for which prognostic signatures were previously published, the top inferred Master Regulators consistently and substantially outperformed the signatures in five-fold cross-validation tests using an SVM-based classifier. This was especially visible when the signature or Master Regulators inferred from one dataset were used to classify samples from the other.

Since they are inferred from a network model that explicitly encodes the regulation logic (i.e. edges are directed), these findings also generate testable hypotheses and rational models for understanding the oncogenic processes leading to the phenotypic difference between poor and good prognosis samples.

References

1. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O. & Staudt, L. M. (2000) *Nature* **403,** 503-11.
2. 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. & Lander, E. S. (1999) *Science* **286,** 531-7.
3. 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., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. & Bernards, R. (2002) *N Engl J Med* **347,** 1999-2009.
4. Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M. E., Yu, J., Jatkoe, T., Berns, E. M., Atkins, D. & Foekens, J. A. (2005) *Lancet* **365,** 671-9.
5. Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. (2005) *Bioinformatics* **21,** 171-8.
6. Michiels, S., Koscielny, S. & Hill, C. (2005) *Lancet* **365,** 488-92.
7. Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A. & Chinnaiyan, A. M. (2005) *Science* **310,** 644-8.

8. Rhodes, D. R. & Chinnaiyan, A. M. (2005) *Nat Genet* **37 Suppl,** S31-7.
9. Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R. & Califano, A. (2005) *Nat Genet* **37,** 382-90.
10. Palomero, T., Lim, W. K., Odom, D. T., Sulis, M. L., Real, P. J., Margolin, A., Barnes, K. C., O'Neil, J., Neuberg, D., Weng, A. P., Aster, J. C., Sigaux, F., Soulier, J., Look, A. T., Young, R. A., Califano, A. & Ferrando, A. A. (2006) *Proc Natl Acad Sci U S A* **103,** 18261-6.
11. Mani, K. M., Lefebvre, C., Wang, K., Lim, W. K., Basso, K., Dalla-Favera, R. & Califano, A. (2008) *Mol Syst Biol* **4,** 169.
12. Margolin, A. A., Wang, K., Lim, W. K., Kustagi, M., Nemenman, I. & Califano, A. (2006) *Nat. Protocols* **1,** 662-671.
13. 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., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. & Friend, S. H. (2002) *Nature* **415,** 530-6.
14. Vapnik, V. (1995) *The Nature of Statistical Learning Theory* (Springer.
15. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. (2005) *Proc Natl Acad Sci U S A* **102,** 15545-50.