Microarray Based Diagnosis Profits from Better Documentation of Gene Expression Signatures

Dennis Kostka[1*¤], **Rainer Spang**[2]

1 Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany, **2** Computational Diagnostics Group, Institute for Functional Genomics, University of Regensburg, Regensburg, Germany

Microarray gene expression signatures hold great promise to improve diagnosis and prognosis of disease. However, current documentation standards of such signatures do not allow for an unambiguous application to study-external patients. This hinders independent evaluation, effectively delaying the use of signatures in clinical practice. Data from eight publicly available clinical microarray studies were analyzed and the consistency of study-internal with study-external diagnoses was evaluated. Study-external classifications were based on documented information only. Documenting a signature is conceptually different from reporting a list of genes. We show that even the exact quantitative specification of a classification rule alone does not define a signature unambiguously. We found that discrepancy between study-internal and study-external diagnoses can be as frequent as 30% (worst case) and 18% (median). By using the proposed documentation by value strategy, which documents quantitative preprocessing information, the median discrepancy was reduced to 1%. The process of evaluating microarray gene expression diagnostic signatures and bringing them to clinical practice can be substantially improved and made more reliable by better documentation of the signatures.

Citation: Kostka D, Spang R (2008) Microarray based diagnosis profits from better documentation of gene expression signatures. PLoS Comput Biol 4(2): e22. doi:10.1371/journal.pcbi.0040022

Introduction

It has been shown that microarray based gene expression signatures have the potential to be powerful tools for patient stratification, diagnosis of disease, prognosis of survival, assessment of risk group and selection of treatment [1–3]. These signatures are computational rules for deriving a diagnosis from a patient's expression profile. Micorarray based molecular diagnosis constitutes a promising approach, but so far it has been difficult to exploit its potential. Before gene expression signatures can impact clinical practice, they need to be communicated to other health care centers with data for external evaluation [2], and ultimately to practitioners for use in clinical routine. This requires unambiguous documentation of the signature.

We believe documentation does not receive sufficient attention. For instance, it has been observed that the reconstruction of published signatures can require an expert re-analysis of the study [4] or may not be possible at all [5]. Documenting a signature is conceptually different from reporting a list of genes contributing to the classification rule, since the latter does not determine how an external patient should be diagnosed. Moreover, we show that even the exact quantitative specification of the classification rule does not constitute a ready-to-use signature. In addition, a procedure transforming raw external expression data to the signature-specific scale has to be provided.

Based on eight clinical microarray studies [6–13] we show that common documentation standards have the following shortcoming: When using the documented information only, a study-external patient might receive a diagnosis different from the one he would have received as part of the original study. To address the problem, we propose guidelines and methodology for building and documenting signatures unambiguously. We demonstrate that our approach reduces ambiguity of diagnoses to a minimum.

Results

Ambiguity of Diagnosis

Before we start with a systematic analysis of the signature documentation problem, we illustrate it in the context of the breast cancer study of Huang et al. [9]. This study includes expression profiles from 89 breast cancer patients. The majority (34) had no recurrence of disease, 18 suffered from a relapse and for the rest no data about recurrence is available. We preprocessed the entire study data using the rma protocol (see Methods), and derived a prognostic signature using a subset of the data (training set). Patient 00291471 was not part of the training set. When applying the signature, this patient was predicted to be relapse free. Next, we classified patient 00291471 a second time, this time as a study-external patient from another health care center. Consequently, the expression profile was preprocessed independently from the study data but using the same preprocessing protocol (rma). Again, we applied the prognostic signature, and this time the patient was predicted to relapse. The reason for this inconsistency is that any quantitative specification of the classification rule critically depends on the scale of the preprocessed data. The scale, in

Editor: Geoff McLachlan, University of Queensland, Australia

Received May 31, 2006; **Accepted** December 20, 2007; **Published** February 15, 2008

A previous version of this article appeared as an Early Online Release on January 2, 2008 (doi:10.1371/journal.pcbi.0040022.eor).

* To whom correspondence should be addressed. E-mail: dkostka@ucdavis.edu

¤ Current address: UC Davis Genome Center and Department of Statistics, University of California Davis, Davis, California, United States of America

Author Summary

It has been shown that microarray based gene expression signatures have the potential to be powerful tools for patient stratification, diagnosis of disease, prognosis of survival, assessment of risk group, and selection of treatment. However, documentation standards in current publications do not allow for a signature's unambiguous application to study-external patients. This hinders independent evaluation, effectively delaying the use of signatures in clinical practice. Based on eight clinical microarray studies, we show that common documentation standards have the following shortcoming: when using the documented information only, the same patient might receive a diagnosis different from the one he would have received in the original study. To address the problem, we derive a documentation protocol that reduces the ambiguity of diagnoses to a minimum. The resulting gain in consistency of study-internal versus study-external diagnosis is validated by statistical resampling analysis: using the proposed *documentation by value strategy*, the median inconsistency dropped from 18% to 1%. Software implementing the proposed method, as well as practical guidelines for using it, are provided. We conclude that the process of evaluating microarray gene expression diagnostic signatures and bringing them to clinical practice can be substantially improved and made more reliable by better documentation.

turn, depends on the data included in the preprocessing process, which is different in an internal and an external analysis.

Documentation Strategies

We studied the signature documentation problem more systematically on eight clinical microarray studies [6–13] involving different cancer types and representing diagnostic as well as prognostic classification problems. Table 1 gives an overview. We compared two different documentation strategies:

Documentation by reference. The signature is documented by a quantitative classification rule and by referencing the underlying preprocessing protocol. This includes the bookkeeping of user-adjustable parameters, such as the target scale for the standard Affymetrix protocol (mas). External cases are preprocessed by running the specified procedure exclusively on the new data.

Documentation by value. The signature is documented by a quantitative classification rule and by supplying a set of parameters (values) determined by the fitted vsn and rma preprocessing procedures (see Methods). These parameters depend on the data that was preprocessed and determine its scale. External cases are scaled by using the corresponding add-on preprocessing routines (see Methods).

Documentation by reference does not include properties of the study data. A comparable scale is assumed to be generated automatically, given the exact details of the preprocessing routine. In contrast, documentation by value communicates the scale implied by the study data explicitly. Data normalized by rma was documented using both strategies. In case of vsn, normalized data was documented by value only. The reason is that documentation by reference requires applying the preprocessing protocol to a single array. In the case of vsn, the normalization step requires more than one array, so this is not possible. On the other hand, documentation by value works well in this setting. Finally, for the Affymetrix Microarray Analysis Suite the two strategies coincide, since arrays are normalized independently of each other.

Consistency of Diagnosis

We calculated consistency and kappa index as described in the Methods Section. To that end, we repeatedly split each dataset into an internal set (of size 20, 30 and 40) to derive the signature and an external set for validation (using documented information only). This mimics the process of communicating signatures between health care centers. The validation results were compared to a reference diagnosis and the fraction of agreement is reported as consistency. This procedure requires the encoding (documentation) of signature information and its subsequent decoding for use on external data.

The main results are summarized in Table 2, the complete details can be found in supplemental Figures S1 to S8. First, we observed in the study by Beer et al. [6] that documentation by reference can lead to discrepancies between external and internal diagnosis being as frequent as 30%. The median consistency across all studies using documentation by reference was 82%, corresponding to a median discrepancy of diagnoses as high as 18%. This demonstrates the existence and importance of the documentation problem. External researchers will generally not obtain the same diagnoses as the investigators of the original study. More importantly, we observed that the documentation strategy matters: Documenting signatures by value leads to substantially more consistent results than documentation by reference. We observed the biggest effect for the prognostic study by Beer et al., where consistency improved from 70% to 99%. This corresponds to a consistency gain between 27% and 31% (95% CI).

The smallest consistency gain (between 3% and 4%, 95% CI) was observed for the diagnostic study of Willenbrock et al. [13], which poses the easiest classification problem. The median minimal gain in consistency (97.5 % CI) obtained from documenting signatures by value was 15%. On most datasets consistency of rma and vsn preprocessing were comparable; the differences were small. Exceptions are the datasets of Huang et al. [9] and Ross et al. [11], where consistencies obtained with rma were larger. Overall, signatures documented by value displayed high consistency, most of them larger than 98%. Documentation by reference was found to be significantly less consistent (about 15% median consistency loss). Surprisingly, we see no big effect of the size of the internal set on the observed consistency (Supplemental Figures S1-S8). Also, the two learning algorithms (SVM and shrunken centroids, see Methods) seem to perform comparable with respect to consistency (Figures S1-S8). Note that for data normalized with the Affymetrix Microarray Analysis Suite the consistency problem does not exist. Since arrays are normalized independently of each other, one is guaranteed to receive 100% consistency.

Discussion

To the best of our knowledge, this is the first study investigating the problem of documenting diagnostic expression signatures. We were able to demonstrate that common documentation standards are insufficient for unambiguously determining diagnosis. Moreover, we have

Table 1. Summary of Microarray Studies

Study	Disease	Problem	Number of Cases	Accuracy	Documentation Gain
Beer [6]	Adenocarcinoma	Prognostic	84	72% (71%)	29%
Bhattacharjee [7]	Adenocarcinoma	Prognostic	125	63% (62%)	18%
Bild [8]	Ovarian cancer	Prognostic	133	69% (54%)	16%
Huang [9]	Breast cancer	Prognostic	52	81% (65%)	12%
Pomeroy [10]	Medullablastoma	Prognostic	60	67% (65%)	19%
Ross [11]	Childhood ALL	Risk Group	87	81% (81%)	19%
Shipp [12]	DLBCL	Prognostic	58	64% (62%)	11%
Willenbrock [13]	Childhood ALL	Diagnostic	45	100% (58%)	4%

Overview of the eight studies used to investigate the signature documentation problem. As accuracy we report a 10-fold cross validation estimate, averaged over 100 different random partitionings. In parentheses we report the prevalence. Documentation gain denotes the increase in consistency when using documentation by value instead of documentation by reference (see Results).

doi:10.1371/journal.pcbi.0040022.t001

shown that the consistency of diagnostic signatures can be substantially improved by documenting data-dependent preprocessing information. To do so, we have proposed the documentation by value strategy.

We observed a trade-off between the reported performance of preprocessing protocols and the effort required for documenting them [14,15]. While it is known that preprocessing schemes that share information across arrays enhance precision and accuracy of estimated expression differences, improved normalization performance comes at a price: The already normalized expression values for a fixed microarray change when additional arrays are added to the study. This is a problem for applying a signature to external data; the original data needs to be included in the normalization of external arrays. The re-normalization of the complete dataset changes the original expression values, affecting the signature and the molecular diagnosis of patients in the original study. To circumvent this problem, we have altered the widely used preprocessing methods rma [16] and vsn [17] to provide an add-on mode. This allows one to first process a core dataset, and then to add data from additional arrays without changing the normalized core data.

As a summary of our findings we propose the following guidelines for deriving and documenting a diagnostic gene expression signature:

Preprocessing

Preprocess the data using a protocol that allows for later inclusion of arrays without changing the original expression values. For example, this can be done preprocessing arrays independently of each other, or by making use of the software we provide.

Building the Classification Rule

Derive a classification rule using software that provides a complete quantitative specification of the signature for documentation purposes. For example, this can be the nearest shrunken centroid procedure [18] we employed.

Documentation by Value

Document the full quantitative specification of the classification rule. In addition, document preprocessing by including aggregate data-dependent information. For example, this can be done with the software we provide. Ideally, make both parts available as an integrated open source computer program that can readily be used to diagnose new patients.

Diagnosing an External Patient

Bring the raw data to a signature consistent scale. For example by using our software. Apply the documented classification rule to diagnose the new patient.

These guidelines suggest methods we have found to work well in practice, but we do not claim them to be optimal in any sense. Given the heterogeneity of clinical data as well as the diversity of array platforms, it can safely be assumed that there is data where other methods are more appropriate. However, we believe that in these situations the documentation problem still exists, and a similar documentation by value strategy should be developed for the methodology in use.

In our simulation setup, the data we call external are actually arrays from the same study. With real external data, additional problems occur. It has been shown that even when using the same technology and experimental protocols, the resulting data for the same tissue sample varies between different health care centers [15,19]. Also, different centers might not employ the same definitions of outcomes and/or diagnoses. While this effect is not directly linked to documentation, we believe that the benefits of documenting signatures by value are enhanced in situations where external and internal data are more heterogeneous.

Documentation of signatures is significantly easier for preprocessing methods treating arrays independently of each other, as is the case for the Affymetrix Microarray Suite. However, preprocessing protocols sharing information across different arrays are commonly used as they can provide increased accuracy [15]. Overall, methods like mas present a viable alternative to more advanced procedures whenever consistency is of prime importance, as it is the case in clinical microarray studies. In our study we did not find strong evidence favoring one approach over the other.

While microarray based diagnostic signatures hold great promise to improve diagnosis and prognosis of disease, evaluation of a signature's predictive performance is difficult and subject to current research and argument [4,19–23]. It is important to prove that a signature holds independent complementary information to existing prognostic markers. While sharing candidate signatures within the research community can accelerate the process of evaluation, this

Table 2. Documentation by Value Increases Consistency

Dataset		rma (ref)	rma (val)	vsn (val)	Gain
Beer [6]	**Consistency**	**70** (68, 73)	**99** (98, 99)	**99** (98, 99)	**29** (27, 31)
	Kappa index	**40** (34, 46)	**98** (96, 99)	**97** (96, 99)	
Bhattaharjee [7]	**Consistency**	**81** (79, 83)	**99** (98, 99)	**98** (98, 99)	**18** (16, 20)
	Kappa index	**62** (56, 67)	**98** (96, 99)	**97** (95, 98)	
Bild [8]	**Consistency**	**83** (81, 85)	**99** (99, 100)	**98** (98, 99)	**16** (14, 16)
	Kappa index	**67** (62, 71)	**98** (97, 100)	**97** (95, 98)	
Huang [9]	**Consistency**	**87** (86, 89)	**99** (99, 100)	**89** (87, 90)	**12** (11, 13)
	Kappa index	**74** (70, 78)	**99** (98, 100)	**77** (73, 81)	
Pomeroy [10]	**Consistency**	**80** (78, 82)	**99** (98, 99)	**96** (95, 97)	**19** (17, 20)
	Kappa index	**60** (55, 65)	**97** (96, 99)	**93** (90, 95)	
Ross [11]	**Consistency**	**79** (77, 82)	**98** (98, 99)	**94** (93, 96)	**19** (17, 21)
	Kappa index	**59** (53, 64)	**97** (95, 98)	**88** (85, 92)	
Shipp [12]	**Consistency**	**88** (86, 90)	**99** (99, 100)	**98** (98, 99)	**11** (10, 13)
	Kappa index	**75** (71, 80)	**99** (97, 100)	**98** (96, 99)	
Willenbrock [13]	**Consistency**	**96** (95, 97)	**99** (99, 100)	**100** (99, 100)	**4** (3, 4)
	Kappa index	**92** (89, 95)	**99** (98, 100)	**99** (99, 100)	

Each row contains consistency results for one of the eight clinical microarray studies. For the preprocessing schemes rma and vsn, we report consistency and kappa indices (see Methods) for signatures documented by reference (rma only [ref]) and signatures documented by value (val). The last column displays the gain in consistency when using documentation by value (rma only). Confidence intervals are given in parentheses (95%). The size of the internal dataset was 40 arrays, except for the studies of Huang et al. [9] and Ross et al. [11], where the study size restricted us to 30 arrays. Documentation by value significantly increases consistency in all studies. rma, robust multi-chip average; vsn, variance stabilizing method.
doi:10.1371/journal.pcbi.0040022.t002

does not allow for any ambiguity of signatures. We believe that our documentation by value strategy removes this obstacle and greatly facilitates signature evaluation. The additional effort required for documentation by value is small. There are certainly several ways to implement documentation by value, and we provide suggestions for two preprocessing schemes. Overall, we found that the consistency of diagnoses based on gene expression diagnostic signatures critically depends on the documentation of the signature. Optimal consistency was obtained by using the proposed documentation by value strategy.

Methods

Data sources. We studied the signature documentation problem on eight clinical microarray studies [6–13], involving different disease types and representing diagnostic as well as prognostic classification problems. See Table 1 for an overview.

Preprocessing. Raw microarray data is subject to noise. There is variation in the data that is not due to biological signal, but rather to measurement error or experimental artifacts. Prior to data analysis it is therefore common practice to perform data preprocessing. Preprocessing procedures depend on the microarray technology in use. We focused on Affymetrix GeneChip technology, where data preprocessing commonly includes three steps: Background correction, normalization and probeset summary. We focus on three of the most widely used protocols: Standard preprocessing as provided by the Affymetrix Microarray Analysis Suite Version 5 (mas) [24], a procedure called ”robust multi-chip average” (rma) [16], and a variance stabilizing method (vsn) [17]. For vsn and rma we provide an ”add-on” mode. This mode enables one to first process a core dataset, and then to add data from additional arrays without changing the normalized core data.

Learning algorithms. In order to achieve comparable results across studies, we re-learned classification rules using the same two learning algorithms for all datasets. We chose a nearest shrunken centroid classifier (pam) [18] that was shown to perform well in recent comparisons of classification algorithms [20,25]. Additionally we used support vector machines (SVMs) [26] with a linear kernel. In the case of pam, classification rules were documented by gene-specific weights and class-specific shrunken centroids. For the SVMs gene specific scaling factors, the support vectors and the offset were documented. All these parameters depend on the scale of the preprocessed data.

Consistency. Consistency of a signature was assessed through re-learning, using 1000 different random subsets of the study data (sub-sampling), and then analyzing the diagnoses of patients across all sub-sampling runs. We randomly split each of the eight datasets into two parts, which we call the internal and the external set. The size of the internal sets was fixed to 20, 30 and 40 arrays for all datasets. For a given preprocessing and documentation strategy (see Results), we first derived and documented a signature using the internal set. Then we applied this signature to a random sample of the external set, exclusively using the information documented beforehand. This mimics the process of communicating signatures between health care centers.

For evaluation of performance we determined the diagnosis a patient would have received if analyzed in the context of the original study (reference diagnosis). To this end, we concatenated each external case with the (say N) internal arrays, renormalized this complete dataset of $N + 1$ cases and applied the signature to the external case. The result constitutes the reference diagnosis for this patient. Then we compared all prior diagnoses from the sub-sampling runs to the reference. The fraction of matching diagnoses was reported as a consistency index. A consistency of one corresponds to the situation where all diagnoses were identical to the reference. A consistency of zero implies that all diagnoses were different from the reference. In addition we report the kappa index [27], a statistical measure to assess inter-rater reliability.

Statistical analysis. All confidence intervals were calculated assuming Bernoulli models for class predictions. In the case of confidence intervals for consistency gains, an additional convolution of estimated Binomial densities was carried out. More details can be found in supplemental material.

Signature documentation. In the core of this paper lies the documentation by value strategy for documenting diagnostic signatures: In addition to the parameters describing a classification rule, documentation by value also keeps track of the normalization dependent scale that is underlying the signature. This scale does not only depend on the preprocessing strategy, but also on the original data. In the following, we demonstrate how to document the scale for two preprocessing methods, rma and vsn.

Documenting quantile normalization. For rma, background correction is performed on an array-by-array basis. The subsequent normalization step can be documented as follows. Assume we have p probes and n arrays. Let X be the $p \times n$ background corrected probe-level expression matrix on log scale. Let π be the permutation sorting the columns of X and π^{-1} its inverse. Then the quantile normalized [28] version of $\tilde{X}$ of X is defined as:

$$\tilde{X} = \pi^{-1}(\pi(X)1),$$

where 1 is a $n \times p$ matrix with all elements equal to $1/n$. Let $\hat{\mu}$ be equal to the first (and therefore any) column of $\pi(X)1$. Then $\hat{\mu}$ is the vector of (identical) quantiles of each normalized array. To bring an external array to this scale, let $x \in \mathbb{R}^p$ be its raw probe level expression values. If π_x is the permutation sorting the entries of x, a scale-consistent version of x is given by

$$\tilde{x} = \pi_x^{-1}(\hat{\mu}).$$

The normalized array $\tilde{x}$ is consistent with the scale of the other arrays as it has the same quantiles. Since π_x depends on the sorting of the entries of x only, we do not need to worry about a global background correction. That is, up to probeset summary, rma can be documented by keeping track of $\hat{\mu}$.

Documenting the variance stabilizing transformation. In the case of vsn, the raw probe-level expression matrix X is background corrected and normalized simultaneously. Huber et al. [17] relate random variables X_{ki} ($k = 1 \ldots p$ and $i = 1 \ldots n$) to the true abundance μ_k of probe k on array i, given probe k is not differentially expressed:

$$\operatorname{arsinh}(a_i + X_{ki} b_i) = h(X_{ki}; a_i, b_i) = \mu_k + \varepsilon_{ki}, \varepsilon_{ki} \sim N(0, c^2). \qquad (1)$$

The parameters $a_i \in \mathbb{R}\mathbb{R}$, $b_i \in \mathbb{R}^+$and $c \in \mathbb{R}^+$are estimated from the data.

Assume vsn normalized core data is at hand. For n arrays we have normalized expression values $\{\hat{h}(x_{ki}) = h(x_{ki}; \hat{a}_i, \hat{b}_i)\}$ with $i = 1 \ldots n$ and $k = 1 \ldots p$, and corresponding parameter estimates $\{\hat{a}_i, \hat{b}_i\}$. Also, a set K of not differentially expressed genes has been identified. This implies estimates $\hat{\mu} = \{\hat{\mu}_k\}$ for each gene $k \varepsilon K$ and an estimate of the variance of the residuals in (1):

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} \hat{h}_i(x_{ki}) \text{ and } \hat{c}^2 = \frac{1}{n|K|} \sum_{k \in K} \sum_{i=1}^{n} (\hat{h}(x_{ki}) - \hat{\mu}_k)^2.$$

We want to transform an external sample x^e to the scale determined by the n core arrays. By employing the same model as for the core data (Equation (1)) and plugging in the estimates, we get maximum likelihood estimators for the parameters a^e and b^e:

$$(a_e, b_e) = \underset{(a,b)}{\operatorname{argmin}} \left\{ \sum_{k \in K} \frac{(h(x_{ki}; a, b) - \mu_{ki})^2}{2c^2} - \sum_{k \in K} \log(\partial_{x_k} h(x_k^e; a, b)) \right\}$$

which can be calculated numerically. These estimators are completely determined by $\hat{\mu}$,$\hat{c}^2$ and measurement values from the external array. They define the variance stabilizing transformation $h(x_k^e) = h(a_e + b_e x_k^e)$ bringing x^e to the same scale as the core data. Therefore, up to probeset summary, vsn preprocessing is fully documented by $\hat{\mu}$ and $\hat{c}^2$.

Documenting an additive model for probeset summary. Let $X^{(k)}$ be the submatrix of normalized expression values indexed by the probes belonging to probe set k (across all arrays). Let there be l probes in probeset k. Then an additive model assumes

$$X_{ij}^{(k)} = p_i^{(k)} + g_i^{(k)} + \varepsilon_{ij},$$

where $p_i^{(k)}$, $i = 1 \ldots l$, is a probe specific effect and $g_j^{(k)}$, $j = 1 \ldots n$, represents the abundance of mRNA of gene k on array j. The parameters $\hat{p}^{(k)} = \{\hat{p}_i^{(k)}\}$ and $\hat{g}^{(k)} = \{\hat{g}_j^{(k)}\}$ can e.g. be estimated by the median polish procedure [29]. Denote by x^e a (suitably normalized) external array. Let $x^{e,(k)}$ denote the values for probeset k. Then median $(x^{e,(k)} - \hat{p}^{(k)})$ denotes a consistent estimate of the expression of gene k. That is, the additive model is fully documented keeping track of the probe effects $\hat{p}^{(k)}$ for all probesets on the array.

Supporting Information

Figure S1. Beer et al.

Beer et al. [6].

Found at doi:10.1371/journal.pcbi.0040022.sg001 (16 KB PDF).

Figure S2. Bhattacharjee et al.

Bhattacharjee et al. [7].

Found at doi:10.1371/journal.pcbi.0040022.sg002 (16 KB PDF).

Figure S3. Bild et al.

Bild et al. [8].

Found at doi:10.1371/journal.pcbi.0040022.sg003 (16 KB PDF).

Figure S4. Huang et al.

Huang et al. [9].

Found at doi:10.1371/journal.pcbi.0040022.sg004 (15 KB PDF).

Figure S5. Pomeroy et al.

Pomeroy et al. [10].

Found at doi:10.1371/journal.pcbi.0040022.sg005 (16 KB PDF).

Figure S6. Ross et al.

Ross et al. [11].

Found at doi:10.1371/journal.pcbi.0040022.sg006 (15 KB PDF).

Figure S7. Shipp et al.

Shipp et al. [12].

Found at doi:10.1371/journal.pcbi.0040022.sg007 (16 KB PDF).

Figure S8. Willenbrock et al.

Willenbrock et al. [13].

Found at doi:10.1371/journal.pcbi.0040022.sg008 (16 KB PDF).

Text S1. Supplementary Methods and Software

Found at doi:10.1371/journal.pcbi.0040022.sd001 (225 KB PDF).

Acknowledgments

We thank Stefan Bentink, Anja von Heydebreck, Andreas Rosenwald, and Michael Hummel for discussions and helpful advice.

Author contributions. DK and RS jointly designed the experiment and the statistical analysis and participated in the writing of the manuscript. DK wrote the software and the supplemental material.

Funding. The research described in this article was funded by the Deutsche Krebshilfe program, grant 70 3173-Tr3 Molecular Mechanisms of Malignant Lymphoma. The Deutsche Krebshilfe did not have any role in design, analysis, interpretation or report of this article.

Competing interests. The authors have declared that no competing interests exist.

References

1. Ramaswamy S, Perou CM (2003) DNA microarrays in breast cancer: the promise of personalised medicine. Lancet 361: 1576–1577.
2. Simon R (2005) Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol 23: 7332–7341.
3. Whitfield ML, George LK, Grant GD, Perou CM (2006) Common markers of proliferation. Nat Rev Cancer 6: 99–106.
4. Tibshirani R (2005) Immune signatures in follicular lymphoma. N Engl J Med 352: 1496–1497 (author reply).
5. Tibshirani R, Efron B (2002) Pre-validation and inference in microarrays. Stat Appl Genet Mol Biol 1: 1–18.
6. Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, et al. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 8: 816–824.
7. Bhattacharjee A, Richards W, Staunton J, Li C, Monti S, et al. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A 98: 13790–13795.
8. Bild AH, Yao G, Chang JT, Wang Q, Potti A, et al. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature 439: 353–357.
9. Huang E, Cheng S, Dressman H, Pittman J, Tsou M, et al. (2003) Gene expression predictors of breast cancer outcomes. Lancet 361: 1590–1596.
10. Pomeroy S, Tamayo P, Gaasenbeek M, Sturla L, Angelo M, et al. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415: 436–442.
11. Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, et al. (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood 102: 2951–2959.
12. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, et al. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8: 68–74.
13. Willenbrock H, Juncker A, Schmiegelow K, Knudsen S, Ryder LP (2004) Prediction of immunophenotype, treatment response, and relapse in childhood acute lymphoblastic leukemia using DNA microarrays. Leukemia 18: 1270–1277.

14. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP (2004) A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 20: 323–331.
15. Irizarry RA, Wu Z, Jaffee HA (2006) Comparison of Affymetrix GeneChip expressionvmeasures. Bioinformatics 22: 789–794
16. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264.
17. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18 Suppl 1: S96–S104.
18. Tibshirani R, Hastie T, Narasimhan B, Chu G (2003) Class prediction by nearest shrunken centroids, with applications to dna microarrays. Statist Sci 18: 104–117.
19. Biganzoli E, Lama N, Ambrogi F, Antolini L, Boracchi P (2005) Prediction of cancer outcome with microarrays. Lancet 365: 1683 (author reply).
20. Wessels LFA, Reinders MJT, Hart AAM, Veenman CJ, Dai H, et al. (2005) A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics 21: 3755–3762.
21. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365: 488–492.
22. Ransohoff D (2004) Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 4: 309–314.
23. Ntzani E, Ioannidis J (2003) Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. Lancet 362: 1439–1444.
24. Affymetrix. Statistical algorithms description document. Whitepaper. Available at: http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf. Accessed on: 2 October 2007.
25. Lee J, Lee J, Park M, Song S (2005) An extensive comparison of recent classification tools applied to microarray data. Comput Stat Data Anal 48: 869–885.
26. Hastie T, Rosset S, Tibshirani R, Zhu J (2004) The entire regularization path for the support vector machine. J Mach Learn Res 5: 1391–1415.
27. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20: 37–46.
28. Bolstad B, Irizarry R, Astrand M, Speed T (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.
29. Tukey JW (1977) Exploratory data analysis. Reading (Massachusetts): Addison-Wesley.