Reviewer's report

Title: Low expression of a few genes indicates good prognosis in estrogen receptor positive breast cancer

Version: 2 Date: 29 April 2009

Reviewer: Carsten Denkert

Reviewer's report:

In the present work the author presents a new method for feature selection and cancer classification based on microarray data: Genes are identified that have a bimodal distribution of expression values. Using the theory of mixture models, a cutoff is determined and samples are classified as positive or negative for each of the genes. In an training-test environment, an optimal classifier is derived by combining these binary variables.

As an application of this method, a survival model is constructed for estrogen receptor positive breast cancer.

The main value of the manuscript is in the description of the bioinformatical approach, the value of the biomarkers for clinical purposes needs clarification.

Some major issues need to be clarified, as follows:

- Major Compulsory Revisions - related to the bioinformatics

1. The gcrma method for preprocessing of the Affymetrix data includes a quantile normalization that gives each chip the same empirical distribution. Were all the 813 samples of this study preprocessed in one batch together? If not, please comment on the comparability of the gene expression values between different preprocessing batches.

2. The study includes data from different Affymetrix microarray generations (hgu133a and hgu133plus2). How does the author cope with the shifts of the gene expression scales between different Affymetrix array generations that has been described elsewhere (cf. Kilpinen et al., Genome Biol. 2008, 9:R139)?

3. The method for feature selection should be declared and discussed in more detail. In this context it is unclear, wether feature selection targets bimodal distributed gene expression (“multistate genes”) or genes “consisting of a large normal component of low expression values and a long right tail of high expression values”. Please discuss the usage of the ratio for the low component as parameter for feature selection. Another natural parameter to assess bimodality would be the likelihood ratio of the two component mixture model compared to a one component model.

- Major Compulsory Revisions - related to the clinical background
4. The manuscript should follow the REMARK guidelines for description of a biomarker more closely. Most importantly, the relationship of AP4 to established clinical parameters should be described in more detail. Which clinical parameters were associated with relapse in univariate and multivariate analysis? How does the multivariate analysis change if AP4 is included in addition to established parameters?

5. Is there any value for AP4 that goes beyond the information already included in the standard parameter tumor grade? The author states that "tumor grade is not a significant predictor of metastasis in AP4+ and AP4-" (page 5). This issue is not really relevant, the main question would be if AP4 would be of any interest in the separated subsets of G1, G2 and G3 tumors. This should be investigated by a stratified analysis as well as a multivariate analysis, the tables showing the results should be included.

6. The same analysis should be included for nodal status as well as tumor size, the data should be included in the additional tables, as well.

7. The statement "chemotherapy is likely ineffective an AP4-" (page 5) seems rather speculative. Basic tumor biology suggests that highly proliferating tumors respond better to chemotherapy. However, there are no specific data from the present investigation that really links AP4 and response. Top2alpha is also a marker of tumor proliferation, which would explain its association with AP4. The author should move the paragraph on chemotherapy response to the discussion part of the manuscript, which is a more appropriate place.

- Minor Essential Revisions

8. The sentence “If necessary, the left tail of v is trimmed by 10% so that the resulting two components consist of the values below and above a cutoff c.” (page 3) is unclear. Why can this be necessary and how often (for how many genes) is it necessary?

9. Please specify the threshold for treating probabilities as zero that is needed for calculation of the ratio. The standard eps=0.0001 used in the package flexmix seems to be very stringent for gene expression data.

10. The statement “for probes negatively correlated with relapse, the roles of high and low components are reversed" is unclear (page 3). Do you look for genes consisting of a large normal component of high expression values and a long left tail of low expression values in this case?

11. Please add density plots (like Fig. 1) or histograms of the gene expression of ESR1 and all genes in the AP4 classifier to the paper. Please report the r values that lead to the identification of those genes.

12. The author should add a table showing the clinical variables in the different cohorts that were investigated.
13. Discussion: While the bioinformatics approach is very interesting, the AP4 classifier seems to be just another way to describe tumor proliferation, which can be more easily done by markers, like Ki67, by assessing histological grade or by counting the mitoses. The author should discuss the clinical value of AP4 that goes beyond these standard markers.

14. The methods, results and discussion parts are mixed, for example Figure 4 (Ki67) is mentioned in the discussion for the first time. The paragraph “An optimal survival model derived from P…” (page 4) describes an important result and should be moved from the Methods to the Results section. The manuscript should be structured more clearly.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

We declare that we have no competing interests.

Carsten Denkert
Jan Budczies