Author's response to reviews

Title: Translating microarray data for diagnostic testing in childhood leukaemia

Authors:

Katrin Hoffmann (katrinh@ichr.uwa.edu.au)
Martin J Firth (martynf@ichr.uwa.edu.au)
Alex H Beesley (alex@ichr.uwa.edu.au)
Nicholas H de Klerk (nickdk@ichr.uwa.edu.au)
Ursula R Kees (ursula@ichr.uwa.edu.au)

Version: 2 Date: 21 May 2006

Author's response to reviews: see over
Dear Sir/Madam,

Please find enclosed our manuscript 1031022372966385 entitled “Translating microarray data for diagnostic testing in childhood leukemia” by Katrin Hoffmann, Martin J. Firth, Alex H. Beesley, Nicholas H. de Klerk and Ursula R. Kees. It has been revised according to the suggestions, and individual issues raised by the reviewers are listed below together with our actions and comments.

**Reviewer ONE**

**Major compulsory revisions**

1. The statistical methods need to be described in more detail. E.g. they apply a variance filter to eliminate probe sets with a p-value >0.1 without giving details. What exactly refers this p-value to?

   **Action/Comment:** The application of the variance filter has been clarified by stating “First, we applied a variance filter to eliminate non-informative probe sets. We excluded all probe sets from the analysis with a fold-change <1.15 between patient subgroups and a p-value associated with this fold-change of >0.1, calculated using a permutation test (999 permutations).” (Materials and Methods, page 7)

2. After filtering, the top-ranked 1000 probe sets were extracted and combined from the two chips: how exactly? The top 20 genes for subgroup-discrimination are considered: they are highest ranked in what?

   **Action/Comment:** A more detailed description of the RF algorithm has been included in the manuscript to clarify the selection of top1000/ top20 probe sets and their ranking to read “Supervised analysis was then performed separately for HG-U133A and HG-U133B data and each of the subgroups with the remaining, informative probe sets using the decision-tree based algorithm Random Forest (RF, randomForest 3.4 standard settings) [23]. In brief, each RF analysis consisted of 100,000 trees and for each tree, the intrinsic RF reiterative process randomly chooses a subset of samples and probe sets for initial analysis and subsequently uses the remaining samples for testing back. Finally, all probe sets used for RF analysis are ranked according to their ability to discriminate between the groups of interest and for each sample a classification accuracy is obtained, along with a measure of confidence [24].” (Materials and Methods, page 7)

   In addition two references have been added, one of which shows a detailed flowchart of the analysis of microarray data using the RF algorithm (Beesley et al. 2005, British Journal of Haematology).

3. Finally it does not make sense to use PCA as a visualizing tool since it tells nothing about the quality of the algorithm.
Action/Comment: We agree with Reviewer One that PCA is not an appropriate tool to qualitatively assess classification/prediction accuracy and thus the quality of the algorithm. For this reason Table 1 and Table S3 are included in the manuscript, which show the prediction accuracies obtained after 100 cross validations. However, PCA is an established technique for the unsupervised analysis of microarray data and has been extensively used as the algorithm of choice to visualize microarray data (e.g. Yeoh et al. 2002, Cancer Cell; Ross et al. 2003, Blood). We therefore believe that it is valid to show PCA-based classification in this manuscript, since it is supported by the correct cross validation procedure.

Reviewer TWO
Minor essential revisions

1. The authors indicated that they used multiple training and test sets for cross validation and in each of these exercises, different discriminating probe sets would be generated. It is unclear how the authors eventually select the probe sets for their analysis.

Action/Comment: The cross validation procedure is used in this study to obtain an unbiased estimate of the true error rate of the classification procedure (Simon et al. 2003, Journal of the National Cancer Institute), while the decision-tree based algorithm Random Forest is used to select the probe sets that are included in the classifier; see action/comments to concern 2 of Reviewer One. (Materials and Methods, page 7)

2. It is unclear how a case was classified according to the selected probe sets and expression levels.

Action/Comment: The Random Forest (RF), a decision-tree based algorithm, was used for classification; see action/comments to concern 2 of Reviewer One. (Materials and Methods, page 7)

3. Some transcripts may be represented by multiple probe sets. If several probe sets had concordant expression and were selected by computational analysis to be a classifier, it does not seem useful to retain all the probe sets representing the same transcripts in the same classifier.

Action/Comment: We agree with Reviewer Two that probe sets representing the same transcript may not be required in the ultimate classifier, however, we believe that removing these probe sets is not justified at this stage. As indicated in the manuscript the 30 probe sets that are required for accurate discrimination of the six ALL subgroups represent a total of 26 genes, with four genes being represented by two probe sets. As shown in Table 2, the fold-changes and direction in expression for these probe sets are very similar and hence it can be anticipated that their removal will not affect the classifier.

4. The two analysis methods produced overlapping transcript classifiers for the different ALL subgroups. The authors compared the different classifiers originating from the two different analysis programs but did not discuss how well the consensus transcripts will work as classifiers. It is worthwhile doing that analysis.

Action/Comment: We appreciate the comment from Reviewer Two that it would be of interest to perform the analysis using a consensus of the two classifiers. However, we believe that such an analysis is beyond the scope of this manuscript. In addition, given the fact that there is no overlap of the top discriminators for some of the subgroups a combination of both classifiers would increase the number of probe sets, rather than
reduce the number of discriminating genes that are needed for accurate class assignment, which was one of the aims of this study.

5. The authors lumped the ALL samples from patients with 21 cell lines for analysis. Since the expression profiles of cell lines are likely to be significantly different from ALL patient samples, it is not legitimate to combine the two groups as a single entity. Data on the two groups should be shown separately.

**Action/Comment:** The reviewer is concerned about analysing microarray data from cell lines together with those obtained from primary patient specimens. Several independent studies, including our own, have compared the microarray expression profiles of cell lines and primary leukemia specimens (e.g. Andersson et al. 2005, *Leukemia*; Kees et al. 2003, *Molecular Cancer Therapeutics*). These studies have shown that cell lines accurately reflect the profile seen in primary patient samples, and moreover, that primary leukemias co-segregate with cell lines carrying identical genetic rearrangements. We therefore believe that it is unnecessary to conduct a separate analysis for cell lines and primary specimens.

**Reviewer THREE**

**Minor essential revisions**

1. The title seems to propose a diagnostic test for acute leukaemia, which is still far from the results of the manuscript. The authors should probably precise better the message of the manuscript.

**Action/Comment:** The title of the manuscript has been changed to “Translating microarray data for diagnostic testing in childhood leukemia”. (Title page, page 1)

**Discretionary revisions**

1. Mention criteria used for leukaemia diagnosis.

**Action/Comment:** The introduction has been modified to read “Current National Cancer Institute (NCI) criteria for risk assignment utilise age and white blood cell counts (WBC) at diagnosis to stratify patients into standard risk (SR; 1-9.99 years of age and WBC<50,000/µl) and high risk (HR; ≥10 years of age or WBC≥50,000/µl) [4]. In addition, several structural and numerical chromosomal abnormalities are known as independent prognostic factors.” (Background, page 4).

2. Comparative results when comparing cell lines, PB and BM samples.

**Action/Comment:** See action/comments to concern 5 of Reviewer Two

We trust the modifications made to the manuscript meet your requirements and look forward to your response.

Yours sincerely,
Katrin Hoffmann  
Dr. rer.nat.