Author's response to reviews

Title: Discriminating lymphomas and reactive lymphadenopathy in lymph node biopsies by gene expression profiling

Authors:

To Ha Loi (t.loi@amr.org.au)
Anna Campain (annac@maths.usyd.edu.au)
Adam Bryant (a.bryant@amr.org.au)
Tim J Molloy (t.molloy@amr.org.au)
Mark Lutherborrow (m.lutherborrow@amr.org.au)
Jennifer Turner (jtur8838@bigpond.net.au)
Yee Hwa Jean Yang (jeany@maths.usyd.edu.au)
David DF Ma (d.ma@amr.org.au)

Version: 3 **Date:** 30 November 2010

Author's response to reviews: see over

Blood, Stem Cell & Cancer Research Unit
Department of Haematology,
St Vincent's Hospital
390 Victoria St
Darlinghurst NSW 2010 Australia

29th November, 2010

Dear Editor of BMC Medical Genomics,

RE: Submission of revised manuscript 1852645016427570

Discriminating lymphomas and reactive lymphadenopathy in lymph node biopsies by gene expression profiling

I would like to take this opportunity to thank the reviewers for their time and insightful comments in reviewing and improving my manuscript. Their comments are addressed in a point-by-point response on the next 7 pages. For your convenience, each reviewer comment has been numbered and inserted into this document in *italics*.

The manuscript has been altered accordingly in the form of track changes. Any changes to tables are highlighted in yellow. As requested, the GEO record GSE23647 has also been altered and can be accessed for private view using the following link:
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=rvkfjeigsoscini&acc=GSE23647

Please do not hesitate to contact me if there are any problems.

Sincerely,

David D.F. Ma
Professor of Haematology

REVIEWER 1: Paul Boutros

1. *The clinical annotation provided in the GEO records are insufficient. A lot of simple clinical covariates have not been provided: treatment, age, gender, stage/grade, and array-batch. There may be other important clinical covariates in lymphoma that should be added as well. It should be determined if the classifier is biased towards/against any of these factors.*

RESPONSE: We recognise the need for a more detailed clinical annotation. As suggested, the age, gender and array-batch for each biopsy analysed has been added to the GEO record. Grade and subtype information for FL, DLBCL and cHL samples has also been added. Treatment information was not included as 1) we only have access to this information for few samples, and that 2) this manuscript only focuses on the ability of microarray gene profiling to diagnose LN biopsies.

To show that the identified classifiers are not biased towards/against any of these factors, we firstly clustered the relevant samples using the classifiers identified in this study. We then colour coded the samples below each dendrogram according to gender, age, hospital, date sample was arrayed, and subtype/grading. In all dendrograms, samples were found to be randomly spread according to the different covariate features thus indicating that the identified classifiers are not biased towards/against any of these factors. The dendograms for the FL versus DLBCL and the cHL versus remainder comparison are shown as examples in Figure 1 (below). The statement "The clustering of these identified classifiers is not influenced by clinical covariates such as the age or gender (data not shown)." was added to page 11, paragraph 1.

2. *The paper focuses on potential clinical applicability, so it would be useful for the authors to comment on whether or not they feel it will be challenging to obtain flash-frozen samples (rather than FFPE) as part of routine practice. For some tumour-types this is thought to be a major confounding issue for microarray-based classifiers.*

RESPONSE: We absolutely agree that practical considerations are of the upmost importance when envisaging introduction of a new test into routine clinical practice.

This study was conducted prospectively, whereby portions of the samples were routinely allocated to microarray at the time of initial specimen collection. Allocated sample was therefore able to be immediately placed in RNAlater solution (RNA stabilising solution, Ambion). This required a slight change in practise whereby the samples were delivered fresh (ie not in formalin) to the onsite pathology laboratory where the allocation of tissue took place. This process proved to be quite practical in our clinical setting, however it is agreed that such practical issues need to be addressed for this service to become routine:

a. The laboratory needs to have an operating procedure whereby the freshly received sample is promptly allocated as required, including the portion for microarray to RNAlater solution.
b. If the surgery takes place at a site remote to the pathology laboratory it may be necessary for immediate allocation of a portion into RNAlater and the remainder to formalin.

We do not envisage that any of the above points would be a major practical hurdle, nevertheless we have added a comment to the discussion that alludes to the necessary practical considerations (page 13, paragraph 1).

3. *The authors should note that their LOOCV is actually quite heavily biased because of the large number of analyses done on that dataset. Aside from running three separate classification strategies, feature-selection was also performed on this sample cohort, meaning that the LOOCV* *results are very heavily biased upwards. This makes it unsurprising that on the (much smaller) training dataset accuracy drops. The authors should comment on these issues, and on whether or not other validation-cohorts are available.*

RESPONSE: In light of this reviewer comment, we feel that we did not sufficiently explain the process of obtaining the classification rule, the LOOCV error rate and the independent data set error rate sufficiently. Consequently this section in the methods part of the manuscript (page 7-8) has been changed to the following:

"The classification power of the determined optimal set of genes was then tested on the independent test set sample. Firstly the results for each classification built from training datasets are expressed in terms of a classification accuracy rate (%), which represents the similarity between the pathological clinical diagnosis and the microarray diagnosis [14]. The accuracy rate of training datasets was determined by subtracting the LOOCV-error rate (%) from 100%.

Then a separate dataset was used to obtain an independent error rate and accuracy. A DLDA classification rule was constructed from the complete training set data using the optimal number of genes estimated via the LOOCV stage of the analysis. This classification rule was then used to classify the independent data."

4. *How were the training/testing sets identified? Did they come from one large cohort, or are they from clinically separate series? If the former, how was the cohort divided? If the latter, what distinguishes those clinical groups?*

RESPONSE: This study is comprised of two batches of arrays. Array data from batch-1 formed the training set whilst the testing set was identified by batch-2 arrays. Thus the two groups were chronologically defined, but were drawn from the same case-mix of patients. This information used to identify the training and testing sets has been added to the text under sub-heading "GEP classification analysis" of the materials and methods section (page 7, paragraph 1).

5. *A comparison of clinical covariates between the training & testing cohorts (including array batch) needs to be given*

RESPONSE: We agree that it is important to demonstrate that the test and training cohorts are populated by a similar cross-section of patients. To address this point, the gender and age of patients at diagnosis have been added to "Table 1 Summary of the biopsies in each disease category examined by microarray", and shows similar distribution in the training and test set. In addition, two dendograms generated from the unsupervised clustering of samples from batch-1 (training set) and batch-2 (test set) (see Figure 2 below) show the random clustering of samples colour coded according to age, gender and diagnosis. This indicates that the samples from either array batch are not biased towards particular clinical covariates.

6. *The names and version numbers of software packages used needs to be given*

RESPONSE: The following has been added to the methods section (page 7, paragraph 1): "The statistical analysis was performed using the R statistical software version 2.8.1".

7. *26/142 arrays were excluded (18%) -- this seems quite high. Can the authors comment on this rate, and what might cause it? If microarrays were to be clinically useful, would an 18% failure rate be problematic? Are the samples that failed clinically biased (should be shown in supplementary)?*

RESPONSE: While the technical exclusion rate of 18% is somewhat high, it is not inconsistent with that seen in many two-colour printed microarray facilities. We believe that while the exclusions do not affect the findings of this pilot study, there is no doubt that routine application of this test would require an improvement in this rate lest an unacceptable proportion of patients miss out on complete molecular characterisation of their tumour samples.

These failure rates are likely to improve with increased experience with microarray processing. Laboratories with low caseloads may find that the most practical solution would be to refer testing to a centralised laboratory. It is likely also that refined array platforms with decreased complexity (eg: mini arrays) will have lower technical exclusion rates. Furthermore, alternative newer platforms such as Illumina bead arrays may increase the robustness of gene expression profiling.

Given the importance of this comment from the reviewer, a paragraph in the discussion (page 13, paragraph 2) has been reworked, with the insertion of comments specifically addressing the point.

8. *There are several tables of genes. The authors could useful provide some univariate statistics other than fold-changes (eg p-values) on each*

RESPONSE: We agree that enhanced statistical information would be a useful addition to the provided gene tables for the specialist reader. To address this we have provided an additional file-4 “Complete list of classifier genes” which includes the advanced statistics ratio of between sum of squares to within sum of squares (*bss/wss*) values that were used as the feature selection criteria. This statistic was used for feature selection as it produced greater accuracy in a classification context, than other univariate methods.

9. *Lists of all genes in every classifier should be provided as supplementary, so that a reader can faithfully reproduce the classifications*

RESPONSE: We agree that this would be useful to the reader and consequently the information has been included as “Additional file 4 –Complete list of classifier genes”.

10. *The abstract states "Classification error rates were determined by leave one out cross-validation and the identified gene classifiers then evaluated on independent array data sets" but it seems that only one such dataset was used (and it's not clear if it's from an independent clinical series or not)*

RESPONSE: We acknowledge the ambiguity of this statement in the originally submitted manuscript. Therefore paragraph 2 of the abstract (page 2) has been altered as follows: “The independent error rate was then evaluated by testing the identified gene classifiers on an independent (test) set of array data.”

REVIEWER 2: Patricia JTA Groenen

Minor Essential Revisions:

1. *The authors should provide information on the size/amount of the tissue specimens that is needed for the analysis. In addition, how much RNA is used in the analysis? Did the authors check for RNA degradation in the samples? Or was there another quality control step to exclude samples from the analysis?*

RESPONSE: We agree that more information about the specimen collection and processing would be useful to the reader. The approximate size of the biopsies obtained has been added to text under sub-heading “Patient samples” page 5. The use of denaturing agarose gel electrophoresis to check for RNA degradation and the amount of RNA used for analysis has been added to text under sub-heading “RNA and microarray assays” page 5 of the materials and methods section.

2. *More information on the correction for batch effect of arrays is needed.*

RESPONSE: Thank you for identifying this error in the manuscript where the appropriate information was erroneously excluded in the pre-submission editing process. We have corrected this error with the replacement of the text “To correct for batch effect the entire dataset underwent a transformation allowing both batches to have the same signal distribution” by the following text “The batches were analysed and normalised separately to maintain the independence of the two datasets.” (see “preprocessing of data” page 6).

3. *I would like to see more details on the samples used in the study, to rule out misclassification effects due to: 1) the hospital were the samples came from 2) grading of the FL-cases and the GC/ABC type of DLBCL cases 3) tumor load.*

RESPONSE: As requested the GEO records were updated with information relating to the hospital (all samples) and grading/subtyping for FL and DLBCL samples. The following 3 cases of FL from the training set were incorrectly classified: LN176 - grade 1, LN177 - grade 1, LN179 - grade 2. Please see below for our comments on “tumor load”.

4. *Was there a minimal tumor load of the lymphoma cases in this study?*

RESPONSE: Only samples with an unequivocal and non-compound diagnosis were included in the study.

5. *I assume that well-defined and classified cases are included in the present study. Would this technology be able to successful discrimination of cases that are difficult to classify by a pathologist? Did the 5 GC- cases of GCB cell type have the BCL2 translocation and /or the BCL6 translocation?*

RESPONSE: We concur with the reviewer comments here. As mentioned above in the response to reviewer comment 4, this pilot study only included samples with a well-defined unequivocal diagnosis. We hope that the reviewer agrees with our comments in the discussion (page 12, paragraph 2) which addresses the limitations of this pilot study. Certainly, the ability of our identified classifiers to discriminate difficult to classify cases is yet to be tested.

The 5 GC cases of GCB cell type were all BCL6 positive by immunohistochemistry (IHC), however only 3 of the 5 cases had cytogenetics performed and none of the 3 cases had a karyotypically detectable translocation involving BCL6 (or BCL2). Only 2 of the 5 cases were assessed for BCL2 by IHC and were both negative.

6. *The accuracy of RL versus lymphoma cases is limited (80% in the test set). As such, there is a false positivity of 20% of patients receiving the diagnosis of lymphoma instead of having a reactive lesion. And 20% of patients who have a reactive lesion which would be molecularly diagnosed as lymphoma. Can the authors provide more details on the reactive lesions, which* *may partly explain their findings. How can this accuracy (bening-malignant) be improved according to the authors?*

RESPONSE: We agree that this pilot study indicates a somewhat high false positive rate. Further details relating to the reactive lesions of the 23 cases of RL have been added under subheading "Patient samples" (page 5). Only 2 cases of RL (with reactive hyperplasia) in the test set were misclassified as lymphoma. The accuracy to distinguish benign from malignant may be improved by increasing the number of cases used to build the classification, especially since there is an imbalance in the number of reactive biopsies (23) compared to the number of cancerous cases (93). To address this reviewer point and other limitations of microarray technology, we have expanded the discussion (page 12 paragraph 1, and page 13 paragraph 2).

7. *More information on the 31 (of the 38 annotatable) genes that are used for classification*

RESPONSE: We agree that it is important to expand on this information and as such more detailed information has been added to the spreadsheet labelled "common genes" of "Additional file 4 –Complete list of classifier genes".

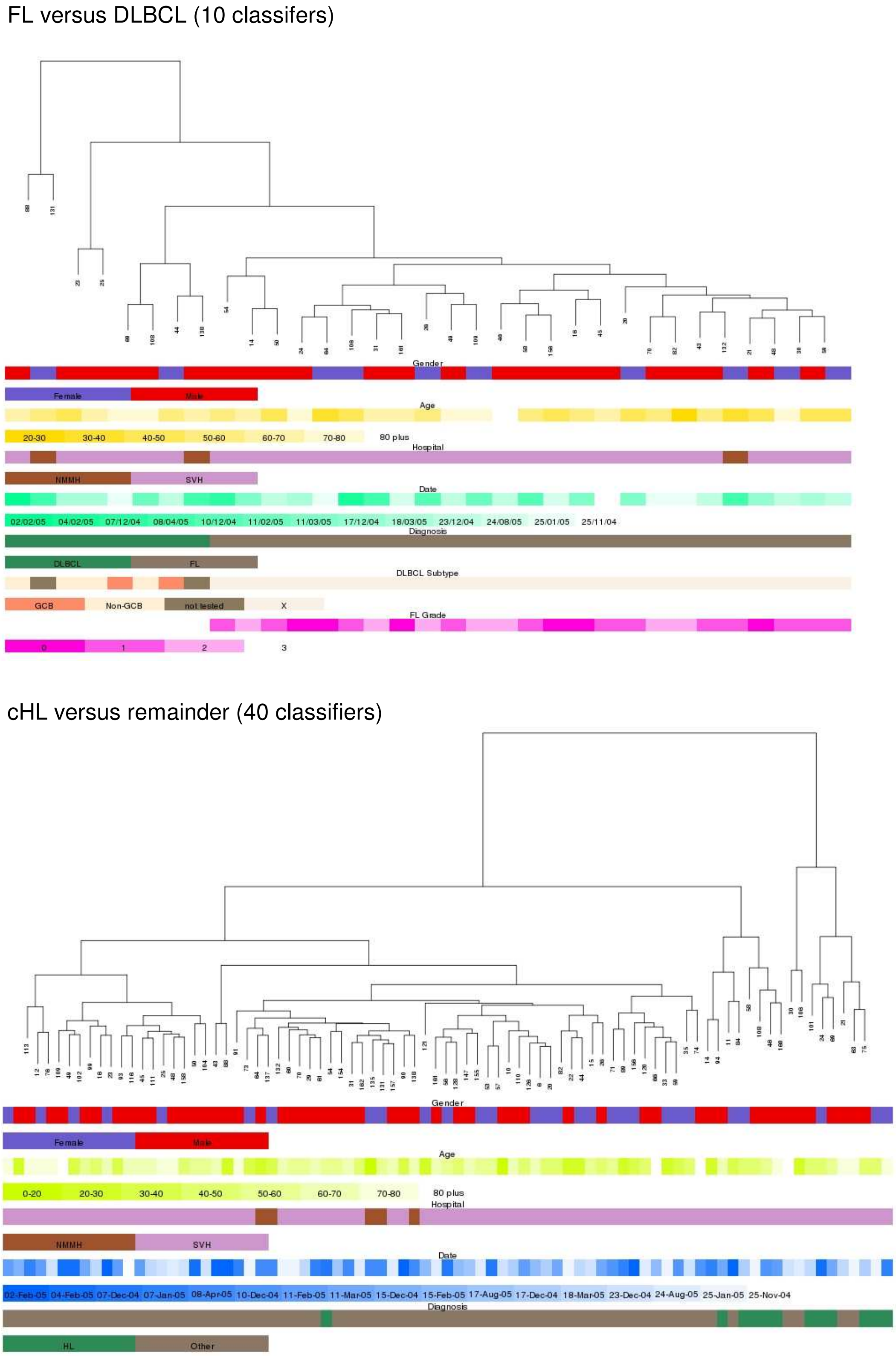
Discretionary revisions

8. *What is the future perspective for this approach? Would this approach be valid also for consultation cases, coming from another hospital, also fresh frozen samples.*

RESPONSE: We have attempted to address these issues by the expanded text in the amended discussion (pages 12-13, paragraphs 2 and 1-2).

9. *The lower expression of particularly the light chain immunoglobulin genes in reactive lymph node is intriguing. Do the authors have information on the light chain use in their reactive lesions and the lymphoma samples, that may explain this phenomenon.*

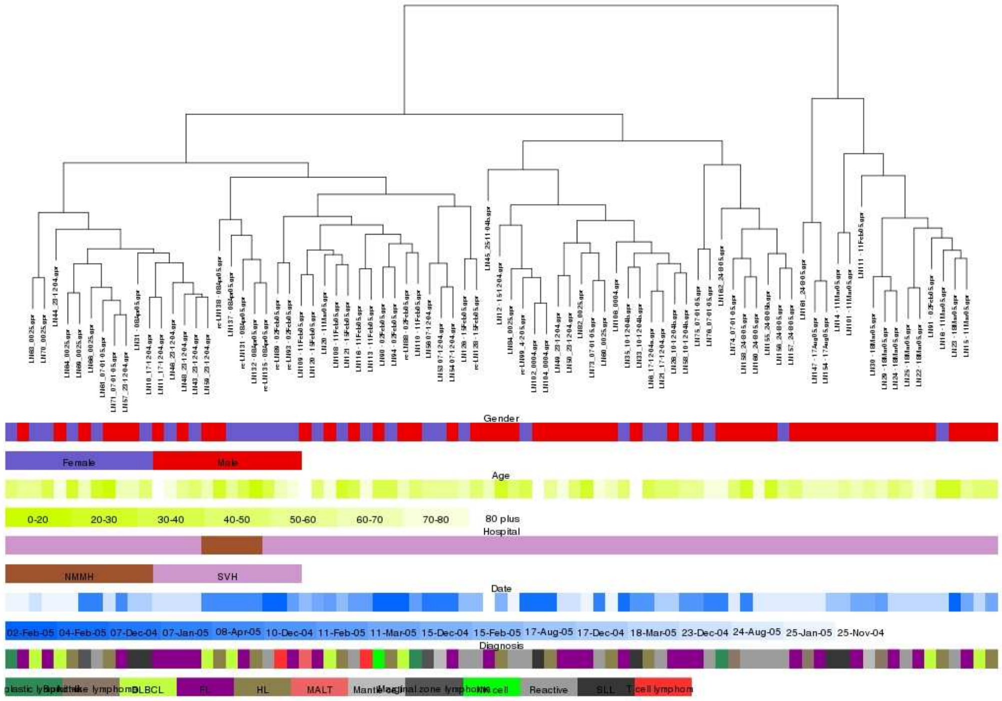
RESPONSE: Unfortunately we did not include this potentially very interesting information in our analysis. This information will be included if the study is expanded beyond the pilot stage.


FL versus DLBCL (10 classifers)

cHL versus remainder (40 classifiers)

Figure 1: Analysis of clinical covariates below dendrograms. Samples from the FL versus DLBCL or cHL versus remainder comparison were clustered using the classifiers identified from these comparisons. Depicted below the dendrogram is the random clustering of samples colour coded according to gender, age, hospital (tissue derived) and date sample was arrayed, FL grading or DLBCL subtype

Batch 1 arrays (training set)



Batch 2 arrays (Testing set)

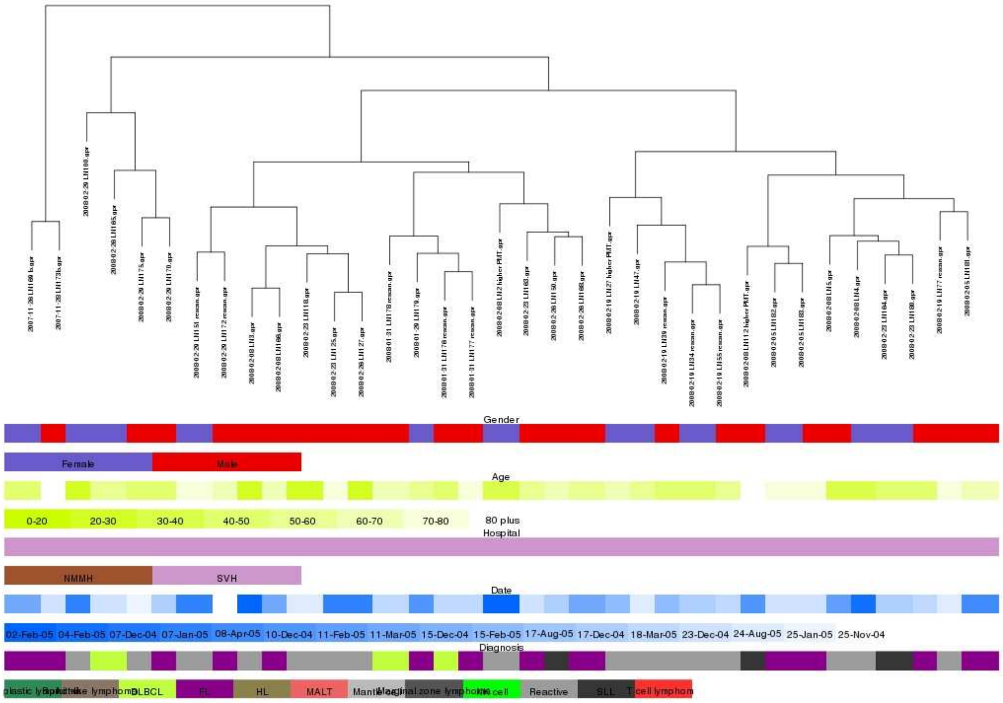


Figure 2: Analysis of clinical covariates below dendrograms generated from the clustering of samples from batch-1 and 2 arrays.