Author's response to reviews

Title: Molecular differential diagnosis of follicular thyroid carcinoma and adenoma based on gene expression profiling by using formalin-fixed paraffin-embedded tissues

Authors:

Aleksandra Pfeifer (apfeifer@io.gliwice.pl)
Bartosz Wojtas (bwojtas@io.gliwice.pl)
Malgorzata Oczko-Wojciechowska (gosiaoczko@io.gliwice.pl)
Aleksandra Kukulska (akukulska@io.gliwice.pl)
Agnieszka Czarniecka (aczarniecka@io.gliwice.pl)
Markus Eszlinger (Markus.Eszlinger@medizin.uni-leipzig.de)
Thomas Musholt (musholt@uni-mainz.de)
Tomasz Stokowy (tomasz.stokowy@polsl.pl)
Michal Swierniak (mswierniak@io.gliwice.pl)
Ewa Stobiecka (e.stobiecka@io.gliwice.pl)
Dagmara Rusinek (drusinek@io.gliwice.pl)
Tomasz Tyszkiewicz (ttyszkiewicz@io.gliwice.pl)
Monika Kowal (mkowal@io.gliwice.pl)
Michal Jarzab (mjarzab@io.gliwice.pl)
Steffen Hauptmann (steffen.hauptmann@patho-ao.de)
Dariusz Lange (dlange@io.gliwice.pl)
Ralf Paschke (Ralf.Paschke@medizin.uni-leipzig.de)
Barbara Jarzab (bjarzab@io.gliwice.pl)

Version: 2 Date: 4 September 2013

Author's response to reviews: see over
Dear Editor of BMC Medical Genomics,

Thank you for considering our manuscript for publication in BMC Medical Genomics. We are grateful for the Reviewers’ comments. We have carefully looked through all of them and have altered the manuscript accordingly. The major changes that we introduced are:

- We deleted the section on validation of FNAB samples. We agree with the Reviewer that this makes the paper more concise than earlier.
- We deleted the entire section in the Discussion describing the genes as the discussion was too long and we moved that part to the supplemental info.
- We compared our 5-gene classifier to others with similar numbers of genes: the 3-gene classifier of Weber et al. and 5-gene classifier of Foukakis et al.
- The title has been corrected by the native speaker.
- We also introduced dr Michal Swierniak as a co-author, who participated in the analysis.

Below, we include point-by-point responses to the reviewers’ comments.

We believe that the present version of the manuscript satisfies the spirit of the journal and reviewers’ views and merits publication in BMC Medical Genomics.

Yours sincerely,

Barbara Jarzab
Reviewer 1 (Dr. Rehannah Borup):

**Major Compulsory Revisions**

1) One major issue is that the authors were unable to validate the defined 8-gene classifier in the FFPE samples which is the main purpose of the study, since only 5 of the 8 selected genes were amplified in the qPCR reaction. Furthermore it is unclear throughout the manuscript which classifier is the final, the 8-gene or the 5–gene classifier, since the two classifiers are referred to as the final interchangeably. This should be clarified and corrected.

We apologize for the lack of clarity. The 5-gene classifier is the final one; we have stressed this information in the corrected manuscript. We consider the three genes not amplified in FFPE material as not suitable for degraded samples.

2) Another issue is the small preoperative FNAB-based dataset used for validation purpose when Pfeifer et al. explicitly stresses that the approach for this study is different from other studies, in that the focus in this study is on the postoperative diagnostic dilemma (p.5, l.15-17). The paper could benefit from deleting the FNAB sections entirely to make it more “to the point”.

Thank you for this very valuable comment. We agree that deleting the FNAB section would improve the paper and we have accordingly deleted it.

**Minor Essential Revisions**

**Abstract**

1) p.3, l.5-8: a rather disorganized sentence that have two aims, where in fact the study focus on a single aim; the differentiation of FTC and FTA, and not so much the preselecting of specimens with increased risk, which could easily be deleted. Moreover the sentence is missing the post-operative focus which should be included in the aim.

We agree with the Reviewer’s comment. The aim of the study is mainly to differentiate FTC from FTA in the post-operative setting. The sentence has been corrected according to the Reviewer's advice.

2) p.3, l.19-20: it is controversial if ITIH5 is significantly expressed in FFPE samples (according to Table 3).

Thank you for this valuable comment; we have now deleted the word “significantly” from the sentence.

3) p.3, l.23-25: The final conclusion does not support the results since the performance of the classifier is likely to be inferior to the performance of an experienced pathologist. The sentence should be rephrased.

In accordance with the Reviewer’s suggestion, we have rephrased the conclusion from the original wording of “The proposed molecular classifier may be used for preselection of thyroid tumors for the judicious histopathological examination” to the following: "The proposed approach could support histopathological examination: 5-gene classifier may aid in molecular discrimination between FTC and FTA in FFPE material".
Background

4) p.4, l.6: why has been?

We agree with the Reviewer; the discrimination between FTC and FTA still is the most difficult aspect of thyroid pathology. We have now corrected this in the manuscript.

5) p.5, l. 12-15: please verify if this paragraph is the opinion of Pfeifer et al. (in which case it does not belong under the “background” section) or in agreement with the paper by Li H, Robinson KA, Anton B, Saldanha IJ & Ladenson PW 2011 Cost-effectiveness of a novel molecular test for cytologically indeterminate thyroid nodules. Journal of Clinical Endocrinology and Metabolism 96 E1719–E1726, which could serve as a reference.

The paragraph is the opinion of Pfeifer et al. For clarity, we added the phrase “in our opinion” to the sentence. We moved that part to the Discussion section, according to your advice. We would like to stress that pharmacoeconomic assessment of Li et al. is calculated according to the US estimated cost of procedures and does not apply worldwide, however we do not discuss it more widely in the manuscript.

6) p.5, l. 24-26: should to be rewritten and backed up by references or deleted.

This text has now been deleted.

7) p.6, l. 2: delete additionally by qPCR analysis and replace with in frozen...

The validation of FNAB samples has been deleted from the manuscript. Therefore, the sentence mentioned above has also been deleted.

Methods

8) p.6, l. 4-14: before describing the array analysis, perhaps make a paragraph “Patients”, hereby separating the two entities. This may make the array paragraph more easily read.

We have now changed this according to the suggestion of the Reviewer.

9) p.6, l. 10-12: how many samples with concordant diagnosis of two pathologists?

In training dataset B (13 FTC and 13 FTA samples), concordance in relation to malignancy was required, and present, for all the samples.

In the testing dataset D (14 FTC and 12 FTC), the diagnosis was based on the optimal consensus diagnosis of the pathologists. We considered the initial diagnosis from the reference centers led by the expert pathologists in the study and their re-assessment, if available. The full clinical material for all cases in test set (especially slides from initial assessment) was not available to us, and the material left in the FFPE block in some cases was not sufficient to provide enough material for the reliable re-assessment.

To summarize the re-assessment in the test set of samples:
Thirteen of all cases were not available for analysis by expert, thus the initial clinical diagnosis was used.

Eight samples, assessed by two experts, had concordance with respect to the malignancy, but the type of the benign lesion was discordant (adenoma vs. other benign lesion). The type of benign lesion was out of scope of the manuscript, we considered it as sufficient if at least one of the experts diagnosed adenoma.

In 3 samples (3 FTC), a definite diagnosis was not possible during the re-assessment phase according to both pathologists (not sufficient amount of material), initial clinical diagnosis was used.

1 sample was diagnosed as FTA by one and as minimally invasive FTC by the second pathologist. The sample was considered FTC, taking into account the definitive character of cancer diagnosis by one of the experts.

1 sample was diagnosed as FTA by both pathologists, but one commented that it was possibly minimally invasive FTC. We decided to set the diagnosis as FTA, taking into account the uncertainty of that suspicion.

As one may see, we used all the samples with ideal concordance to build the training set and used the samples with less thorough examination for the testing set. Nevertheless, as the assessment is based not only on our testing set, but also other external sets, with concordant results, we consider that a reliable approach. Please also note that two independent experts from two centers are the procedure more stringent than in other available studies, thus verifying in the test set only samples with such a concordance may be introducing a bias from routine clinical practice.

In past we carried out a large study of the concordance of pathological diagnosis in thyroid cancer, including follicular tumors. For the occurrence of inter-rater disagreement we refer the Reviewer to Lange et al., as referred in the article.

10) p.6, l. 15: an independent set of samples…..would it be more correct to write the remaining samples were used as an independent set of……?

We have now changed this according to the suggestion of the Reviewer.

11) p.7, l. 10-15: what is the purpose of including these samples? Consider leaving the FNAB section out of this study.

The section on FNAB samples has now been removed from the manuscript.


All the samples in the array experiment were of high quality. Therefore, the word has been removed from the sentence.

13) p.8, l. 2: 216 thyroid samples…..please add of which XX originated from online available dataset...

The sentence has been changed as follows: "In total, 199 thyroid samples were analyzed (123 of our own samples and 76 publicly available samples).” The total number of samples decreased after the exclusion of the FNAB dataset F.

14) p.8, l. 9-10: why not the most specific probe i.e. at_only probe sets versus _s or _f probe sets?

In past, we applied both strategies to select one probe set per gene. In our opinion, both solutions, i.e., choosing the most significant probe set or the most specific probe set, are
possible and appropriate. However, we prefer to select most significant probe set, then relate it to the relevant transcript by hand-curation and further validate it by an independent method (QPCR), usually targeting other regions of the transcript and spanning an intron. This is mainly due to the lack of scientific evidence that Affymetrix/Bioconductor/other mappings of probe sets provide the optimal sensitivity and specificity.

Results

15) p.8 l. 17: It is not clear whether the 5-gene classifier is the final classifier. This should be clarified.

We have now clarified that the 5-gene classifier is the final one.

16) P. 8 l. 25: insert: 5-gene classifier in “…validation of the 5-gene classifier was performed…

This has now been changed according to the suggestion.

17) P. 9 l. 6. According to Figure 1, the 5-gene classifier and not the 8-gene classifier was validated on FNAB samples. Decide which classifier is deemed the final one and correct accordingly.

The paragraph about FNAB has now been removed.

18) P. 10 l. 5-11: Several papers deal with the construction of classifiers with one or a small number of genes, for instance TSP (top-scoring pair(s) algorithms). In general these classifiers perform well if there are relative large differences in expression values between the genes in the two groups of samples that are classified, and some studies show that fold change effect the classification accuracy more than the number of significant genes included in the classifier. Furthermore, the stability of the genes selected across CV (cross validation loops) may be an important criterion for the evaluation of simple classifiers with small number of genes.

Since the 99 pre-selected genes have a log ratios above 1.5 with mean value above 5 in at least one of the groups in data set A, with corresponding log ratios levels in data set B, it is relevant to know the stability (percent occurrence) of the chosen genes within the CV loops as the number of genes is increased. The stability of the chosen genes for each CV loop may be relevant to include in the feature selection process, both for the formulation of classifiers with the pre-selection step and without the pre-selection as described in supplemental Information 1_ page 3.

Before choosing the t-test as a gene selection methods, we also tested other methods such RFE or TSP. However, all the methods gave similar results, so we chose the Student's t-test which is the simplest and the most frequently used one. Finally, we decided to apply the simple 3-step analysis: preselecting the genes on one dataset, then narrowing the number of genes and training the classifier on the second dataset and testing it on the third dataset. In fact, the information about gene fold change was incorporated in our analysis, as the genes selected on dataset A had a fold change above 1.5.

For the Reviewer's information, we calculated the stability of the genes in the cross-validation. We repeated the 10-fold cross-validation 10 times (100 iterations in total). For each gene, we calculated the stability as the percentage of iterations in which the gene was present among the top 8 genes. The 8-gene classifier based on the analysis of the 99 genes from dataset B (with a gene preselection step) consists of the following genes: ELMO1, CA4,
LRP1B, ITIH5, EMCN, PLEKHG4B, SLCO2A1, and KCNAB. In cross-validation loops, the following genes occur, with their corresponding stabilities:

<table>
<thead>
<tr>
<th>affyID</th>
<th>symbol</th>
<th>stability</th>
</tr>
</thead>
<tbody>
<tr>
<td>204513_s_at</td>
<td>ELMO1</td>
<td>97%</td>
</tr>
<tr>
<td>206208_at</td>
<td>CA4</td>
<td>97%</td>
</tr>
<tr>
<td>219643_at</td>
<td>LRP1B</td>
<td>94%</td>
</tr>
<tr>
<td>1553243_at</td>
<td>ITIH5</td>
<td>83%</td>
</tr>
<tr>
<td>227874_at</td>
<td>EMCN</td>
<td>82%</td>
</tr>
<tr>
<td>236255_at</td>
<td>PLEKHG4B</td>
<td>75%</td>
</tr>
<tr>
<td>204368_at</td>
<td>SLCO2A1</td>
<td>46%</td>
</tr>
<tr>
<td>210078_s_at</td>
<td>KCNAB1</td>
<td>43%</td>
</tr>
<tr>
<td>203029_s_at</td>
<td>PTPRN2</td>
<td>40%</td>
</tr>
<tr>
<td>218559_s_at</td>
<td>MAFB</td>
<td>34%</td>
</tr>
<tr>
<td>203980_at</td>
<td>FABP4</td>
<td>30%</td>
</tr>
<tr>
<td>202954_at</td>
<td>UBE2C</td>
<td>21%</td>
</tr>
<tr>
<td>205357_s_at</td>
<td>AGTR1</td>
<td>18%</td>
</tr>
<tr>
<td>202747_s_at</td>
<td>ITM2A</td>
<td>14%</td>
</tr>
<tr>
<td>205554_s_at</td>
<td>DNASE1L3</td>
<td>10%</td>
</tr>
<tr>
<td>229127_at</td>
<td>JAM2</td>
<td>3%</td>
</tr>
<tr>
<td>206529_x_at</td>
<td>SLC26A4</td>
<td>2%</td>
</tr>
<tr>
<td>213234_at</td>
<td>KIAA1467</td>
<td>2%</td>
</tr>
<tr>
<td>219877_at</td>
<td>ZMAT4</td>
<td>2%</td>
</tr>
<tr>
<td>201291_s_at</td>
<td>TOP2A</td>
<td>1%</td>
</tr>
<tr>
<td>202112_at</td>
<td>VWF</td>
<td>1%</td>
</tr>
<tr>
<td>206517_at</td>
<td>CDH16</td>
<td>1%</td>
</tr>
<tr>
<td>209735_at</td>
<td>ABCG2</td>
<td>1%</td>
</tr>
<tr>
<td>214735_at</td>
<td>IPCEFI</td>
<td>1%</td>
</tr>
<tr>
<td>218662_s_at</td>
<td>NCAPG</td>
<td>1%</td>
</tr>
<tr>
<td>219873_at</td>
<td>COLEC11</td>
<td>1%</td>
</tr>
</tbody>
</table>

As we can see, if we used the information about gene stability, we would choose the same genes as when using the t-test. The information about cross-validation occurrence was added to the supplemental file (additional file 2.xls).

Discretionary Revisions

Abstract

1) p.3, l.12: FFPE has not been written out previously.

This has now been corrected.

2) p.3, l. 17: other places in the manuscript the transcripts, named genes here, are written in italic. Please be consistent.

This has now been corrected.

Background

We have now corrected this sentence as follows: "This is evidently very important in the context of preoperative diagnosis, but it may not be the most efficient method for direct application to the analysis of postoperative formalin-fixed paraffin-embedded (FFPE) samples at the mRNA level."

Results

4) P.10 l. 6: “..., when over 20 genes were used”, change to: “..., when more than 20 genes were used within the preselected data set of 99 genes…”

This has now been changed according to the suggestion of the Reviewer.

Quality of written English: Needs some language corrections before being published

Thank you for this important comment. Language corrections have now been performed by a native speaker.
Reviewer 2 (Dr. Cristina Romei)

Major Compulsory Revisions

1. The experiments described in the paper have been well organized in group of cases for the testing and validation of data. Especially for those readers who are not very confident with these kind of experiments there could be some problems to understand. I would suggest to explain for example what is a "training or testing dataset".

   Thank you for the suggestion. We have now explained this a bit more in detail in the text in the revised manuscript.

2. Although it is explained in the supplemental materials, DLDA should be described also in the text.

   We briefly referenced DLDA in the main text of the manuscript:
   "Diagonal linear discriminant analysis (DLDA), a simple and reliable method based on linear combinations of genes, was chosen as the classification engine".

3. It is my opinion that the discussion is too long and in particular the description of the 5 genes might be avoided

   The part on the discussion about genes has been deleted from the manuscript and moved to the supplemental info.

Quality of written English: Needs some language corrections before being published

   Thank you for this important comment. Language correction has now been performed by a native speaker.
Reviewer 3 (Dr. Adetayo Kasim)

Major Compulsory Revisions:

1. A few genes are desirable, but Figures A and B SUPPL1 do not support your choice of 8 genes. 12 genes seem more appropriate based on the misclassification errors. The authors should justify why the number of genes selected is not arbitrary.

   We aimed to create a small and simple classifier, and the addition of restricted number of genes did not improve the classification accuracy significantly (even the 45-gene classifier gives the accuracy which is 4% higher than the 8-gene classifier). The classifier is limited both from the standpoint of cost-effectiveness and the limited amount of FFPE material obtained in the isolation from 5 slices. We have clarified this in the text: "In an attempt to create a classifier of low complexity and due to the limitation of the material, we decided to validate 8 genes".

2. I commend the authors for comparing their 5-gene classifier with the 76-gene classifier by Borup et al. However, there are other studies such as Weber et al with 3-gene and Hinsich et al with 5-gene classifier. It would greatly improve the manuscript if the authors could compare their 5-gene classifier with other classifiers with a similar number of genes. The authors should also discuss whether their 5 genes overlap with other studies (see Table 4 of Borup et al). I understand the discussion about pre- or post-operative gene signatures, but it would still be good to compare the different sets of genes.

   The only 5-gene classifier that we were able to find in FTC/FTA literature, is the one in the study by Foukakis et al. We included the the 5-gene Foukakis classifier and 3-gene Weber's classifier in our comparison. The comparison was performed in a manner analogous to the comparison of Borup's classifier. The results of these comparisons are included in Table 4, which has now been added to our paper. Furthermore, the corresponding paragraphs of the Methods, Results, and Discussion have been updated. Our classifier gives better accuracies than the classifiers of Weber and Foukakis.

   We also compared the 8 genes selected by us to the list of genes selected by other authors of papers with regard to follicular tumors. The results of the comparison are as follows:

   - **ELMO1** - differentially expressed in the study by Barden et al.
   - **EMCN** - not mentioned in any of reviewed papers
   - **ITIH5** - differentially expressed in the study by Borup et al. (but this is not an independent dataset as we used it in our analysis)
   - **KCNAB1** - differentially expressed in three papers: Takano et al., Weber et al., and Borup et al.
   - **SLCO2A1** - differentially expressed in the study by Borup et al.

   - **LRP1B** - not mentioned in any of reviewed papers
   - **CA4** - differentially expressed in the studies of Barden, Weber and Borup
   - **PLEKHG4B** - not mentioned in any of the reviewed papers

   **KCNAB1 and CA4 occur 3 times in the literature on the microarray analysis of FTC and FA differences. They are the most prevalent genes, as we did not see any gene that occurs more than 3 times, among these studies.**

   We have now added the above information to the discussion:
“Some of the genes from our 5-gene classifier have already been mentioned in other studies of high-throughput gene expression analysis of follicular tumours. ITIH5, KCNAB1, and SLCO2A1 are mentioned in the Borup et al. study. Besides, ELMO1 is differentially expressed in Barden et al. and KCNAB1 is differentially expressed in 2 other papers by Takano and Weber, respectively.”

3. The manuscript lacks any information about the uncertainty of identifying true gene signatures. Specifically, the 8 genes identified by the authors are not necessarily “the genes” but a set of genes to discriminate between FTC and FTA. It would be helpful if the authors could provide information about how these 8 genes are ranked in all the datasets in terms of fold changes and p-values. For example, are they consistently ranked in the top 8 across all the datasets?

Thank you for this valuable comment. Indeed, we consider our 5-gene classifier to be a set of cooperating 5 genes. We added the supplementary file “additional file 4.xls” in which we included information about those 5 genes (and 3 other genes, which didn’t amplify on FFPE samples): the results of Student’s t-test or Mann-Whitney test, fold changes, log ratios and the ranks of the genes according to the p-value in various datasets.

**Minor Essential Revisions**

1. The authors should use either “p” (page 3: p < 0.05) or “p value” (page 8: pvalue < 0.0005) and not both.

   In the corrected version, only the form “p-value” is used.

2. Page 8: p value < 0.0005 (i.e. 0.09 false positive expected). The 0.09 false positives should be clarified because the significant level of 0.0005 implies 0.05% chance of finding a false positive.

   We apologize for the lack of clarity. In the corrected version of manuscript, we have explained it more clearly. We selected the genes with p-value below 0.0005 in Student’s t-test. Then we applied the False Discovery Rate correction for multiple comparisons. Assuming that a selection was made from preselected 99 genes, all the genes with p-value below 0.0005 have FDR below 0.09, which means that we expect 9% of false positives among selected genes.

3. Check “!” in “0.11-0.69!1” on page 12.

   Thank you for this revision; we have now deleted this.

**Discretional Revisions**

1. How does the K-folds cross-validation results compare with LOOCV for dataset C and D?

   The results for 10-fold cross-validation and LOOCV are similar:
### dataset C:

<table>
<thead>
<tr>
<th></th>
<th>accuracy</th>
<th>sensitivity</th>
<th>specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>positive likelihood ratio</th>
<th>negative likelihood ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOOCV</td>
<td><strong>0.72</strong></td>
<td>0.71</td>
<td>0.73</td>
<td>0.67</td>
<td>0.76</td>
<td>2.58</td>
<td>0.40</td>
</tr>
<tr>
<td>10-fold cross-validation repeated 10 times</td>
<td>0.71</td>
<td>0.69</td>
<td>0.73</td>
<td>0.66</td>
<td>0.75</td>
<td>2.51</td>
<td>0.43</td>
</tr>
</tbody>
</table>

### dataset D:

<table>
<thead>
<tr>
<th></th>
<th>accuracy</th>
<th>sensitivity</th>
<th>specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>positive likelihood ratio</th>
<th>negative likelihood ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>learning on B and testing on D</td>
<td><strong>0.73</strong></td>
<td>0.71</td>
<td>0.75</td>
<td>0.77</td>
<td>0.69</td>
<td>2.86</td>
<td>0.38</td>
</tr>
<tr>
<td>10-fold cross-validation repeated 10 times</td>
<td>0.69</td>
<td>0.65</td>
<td>0.74</td>
<td>0.75</td>
<td>0.64</td>
<td>2.52</td>
<td>0.47</td>
</tr>
<tr>
<td>LOOCV</td>
<td>0.69</td>
<td>0.64</td>
<td>0.75</td>
<td>0.75</td>
<td>0.64</td>
<td>2.57</td>
<td>0.48</td>
</tr>
</tbody>
</table>