Author's response to reviews

Title: Gene signatures ESC, MYC and ERG-fusion are early markers of a potentially dangerous subtype of prostate cancer

Authors:

Morten B Rye (morten.rye@ntnu.no)
Helena Bertilsson (Helena.Bertilsson@ntnu.no)
Finn Drabløs (finn.drablos@ntnu.no)
Anders Angelsen (anders.angelsen@ntnu.no)
Tone F Bathen (tone.f.bathen@ntnu.no)
May-Britt Tessem (May-Britt.tessem@ntnu.no)

Version: 3
Date: 6 June 2014

Author's response to reviews: see over
Dear editor,

Please find enclosed a revised version of manuscript MS: 1896525600122317, entitled “Gene signatures ESC, MYC and ERG-fusion are early markers of a potentially dangerous subtype of prostate cancer” previously submitted to BMC Medical Genomics.

Included below is our reply to the reviewers’ comments.

All reviewers

A general concern amongst the reviewers, which was also pointed out by the editor, was the lack of additional validation in an independent dataset. We thus address this issue in general first, and then address other issues from each reviewer afterwards.

As was also pointed out by one of the reviewers, finding a dataset meeting all the necessary requirements for complete validation would be very challenging. Nevertheless, we were able to locate a dataset with many of the requirements. These were: i) sufficient number of samples, ii) detailed assessment of sample composition, iii) Gleason score and iv) follow up data on recurrence. We were not able to locate a dataset to assess within patient homogeneity. We downloaded microarray expression and sample composition data from GEO accession GSE8218, and patient data on Gleason and survival from a recent follow up study on the same data (Chen, 2012, PLOS ONE). We have thus used the GSE8218 to validate our results regarding subtype classification with respect to sample composition before and after subtraction of normal tissue component as well as relations of signatures and gene-set scores to patient Gleason scores. In general, the results and observations were similar or comparable to observations in our own data. The additional validation with respect to survival was inconclusive, in particular due insufficient data from patients with good prognosis signatures. Of the samples assigned with bad prognosis together with a Gleason score of 6 or 7, 13 out of 29 patients showed recurrence. The signatures related to general recurrence and survival in general were already validated using a second dataset in the original study by Markert et al., so this has not been the main focus for this study. Since our study focus on signatures in early, mostly localized PCa, use of recurrence data for early localised PCa with low Gleason scores are likely also less informative, since successful surgery will mostly prevent recurrence, even for potentially dangerous tumors. Actually, a gene set or signature with perfect correlation to recurrence would also be of less use for early marker purposes, since it would already be too late for the patient when this marker is present. We have thus chosen to put less emphasis on recurrence in our study, and focused on potentially dangerous subtype features present in early localised PCa with low Gleason. Such features could be very useful for withdrawing patients on active surveillance. The analysis results from the additional validation dataset have been summarized in Supplementary Text 1, and referred to in the main manuscript when appropriate. We have also extended the discussion regarding recurrence data. In general, there have been huge challenges in trying to validate results based on prostate cancer tissue samples. In this respect, the stable presence of the subtype signatures in samples from several studies should represent a promising approach to improve on this issue.
As for the individual comments for each reviewer we have tried to respond to all questions and comments, however, given the time-frame of this revision, we have not been able to follow up on all points, but selected those we decided to be most relevant.

Reviewer number 1

According to the title “Gene signatures ESC, MYC and ERG-fusion are early markers of a potentially dangerous subtype of prostate cancer”, it seems that the main finding of this paper is the three signatures. But in the original paper (Markert et al., 2011), which defined the PCa subtypes used in this paper already indicated that patient groups with poor survival are characterized by ESC and ERG-fusion.

This is correct, but Markert et al. also highly emphasize the P53 and PTEN signatures (and strangely, hardly mention the enrichment of MYC-signatures). Our results are not in conflict with these results. However, we are able to highlight a group of samples with low Gleason score which are not specifically enriched for P53 and PTEN (which, in our opinion, are lesions occurring at later stages), but rather have combinations of ESC, MYC and ERG-fusion signatures.

In the abstract, the authors indicated “Our study validates a previous molecular subtyping of PCa in a new patient cohort, and identifies a subgroup of PCa samples highly interesting for detecting high risk PCa at an early stage”. But the authors also indicated “Our data did not distinguish between four different subtypes of PCa as previously published, but rather highlighted two groups of samples”. The authors should clarify the confusion of these statements.

The subgroup referred to in the first sentence is not directly one of the four subtypes from Markert. So we were not able to distinguish the four subtypes, but were able to identify a subgroup based on our own data, with features related to (but not conflicting with) Markerts subtypes. We have changed the phrasing in the abstract to avoid confusion.

For subtype assignment, p < 0.05 is used. Is the multiple testing correction performed?

We have assesses false discovery rate by i) randomly shuffling gene-set scores among samples, and ii) randomly shuffling the order of genes in all samples. The number of false positive predictions for good and bad prognosis was 5 for the first randomisation and 3 for the second randomization (p<0.05). Thus assignments are very far from being random.

2. For “Gene Set Enrichment Analysis” in Methods: (1) dsig(j) should be defined; (2) in the original paper (Subramanian et al., 2005), the Enrichment Score is defined as the maximum deviation from zero of Phit - Pmiss. What’s the relationships between dGS(j) = square_root((N-S)/s) and Phit, dGS(j) = -square_root(s/(N-S)) and Pmiss?

The “dsig(j)” designation was a typing error and has been fixed. For GSEA analysis we have used the implementation described in supplementary material for Markert et al, which operates with positive and negative GSEA-scores as the maximum and minimum values obtained in a summation scheme. Then the GSEA scores are taken as the maximum deviation of zero between these two. We have tried to compare the formulas in Subramanian and Markert, but it is not completely clear whether the positive and negative Markert scores
exactly correspond to \( \text{Phit} \) and \( \text{Pmiss} \), though they seem to be very related. However, the general strategy is similar.

- Minor Essential Revisions

1. “Prognoses” should be “prognosis”.

*This has been corrected*

2. The authors should indicate which kind of correlation (Spearman correlation or Pearson correlation) was used in this paper.

*Pearson correlation. This has been added to Methods.*

3. Legend of table 1: Table 1: Number of samples assigned exclusively and significantly (\( p>0.05 \) and \( p>0.25 \), dependent correlations by Steiger [36]). Should it be \( p < 0.05 \)?

*This has been corrected.*

**Reviewer number 2**

…there is no independent statistic for the significance of the composition of a cluster. Alternatively the false discovery rate (FDR) for the classification (clustering) should be determined. This might be done by using a random selection of values from Figure 1A not including the chosen profile or even use random values within the range of values shown in Figure 1A and repeating the clustering process and counting the percent of the 54 patients clustered together where a patient is a member with a value above or below a suitable threshold based on the current cluster. There is a facility in R among other packages for this process where it can be repeated a large number of times, say 100, to provide an average for the number of the 54 patients that are clustered together as one estimate of FDR. Several alternatives can be imagined. More rigorous would be to a random selection of the genes of Markert et al. to cluster their cases and use exact values in place of the visual estimates. An estimate of the FDR is essential.

*We have assesses false discovery rate by i) randomly shuffling gene-set scores among samples, and ii) randomly shuffling the order of genes in all samples. The number of false positive predictions for good and bad prognosis was 5 for the first randomisation and 3 for the second randomization (\( p<0.05 \)). Thus assignments are very far from being random, and about what you would expect at a \( p \)-value cutoff of 0.05.*

Did the authors demonstrate that the new subgroup can cluster ‘Unassigned’ samples in Figure 1B into Bad or Good Prognosis groups?

*Some of the samples in the cluster would be the unassigned ones from Fig 1B.*

Correlation is one way to help with the classification. Were other methods used such as Euclidean distance with consistent results?
Pearson correlation was chosen to focus on signature profile shape rather than absolute score values. Our experience with other projects has shown that Euclidean distance does not work very well for profile shape, although we did not try it here. In addition, the statistical test would not be valid for Euclidean distance.

It is not clear how “normal” tissue was defined and chosen which should be described in detail especially any distance separation from tumor.

The normal tissue was sampled as far away from the tumor as possible on each tissue slice. This has now been mentioned in Methods. Actually this probably explains the consistent depletion of bad signatures in our normal tissue samples, which is in contrast to the validation data which seem to have sampled normal tissue closer to the tumor.

The term ‘diagnosis’ and ‘prognosis’ are occasionally used as equivalent, e.g. line 41 which should be corrected throughout as the focus in prognosis only.

This has been corrected.

Minor points are:

Line 86: ‘poly clonal composition’ vs. ‘multi-nodality’.

In our opinion a tumor can be multi-nodal without necessarily being polyclonal. We have tried to be more consistent on this terminology.

Line 357: ‘expectedly’ vs. ‘As expected’.

This has been corrected.

Reviewer number 3

The clinical importance analysis was not validated by clinical follow up data in the present cohort, instead was dependent of previous studies. This weak point should be addressed already in the introduction and method sections.

This has now been mentioned already in the abstract.

Major questions

1. The sample assignment or classification is not only influenced by sample composition but also by different selections of genes or gene sets. Can selections of gene sets lead to the phenomenon of the finding that the molecular similarity between tumors with different Gleason scores in the same patient (prostate) is stronger than between tumors with the same Gleason score in different patients (prostates)? Is this phenomenon observed when the all samples are analyzed by the original microarray dataset with all genes?

This is an interesting question. We have done analysis using all genes and compared them with analysis using the gene-sets, and we generally observe that the general overall molecular characteristics (all genes) seem to be more correlated to gene-set scores than
Gleason scores, that is, the answer to the reviewer’s last question is partly yes. It is of course very likely that by selecting specific genes or other gene sets one can improve the correlation to Gleason score, if this is the goal of the analysis. We have not elaborated more on this here, though.

2. Peng Z et al (Prostate Cancer and Prostatic Disease (2014) 17, 81–90) recently described an ESCGP signature. What are the similarities and differences between this ESCGP and the ESC signature?

We have investigated this new signature (491 of 641 genes present in our data). Interestingly, the ESCGP signature differs quite a lot from the other 3 ESC signatures. While the average correlation among the other three ESC signatures is 0.77, the average correlation of the ESCGP signature to the other ESC signatures is -0.17. The highest correlation of the ESCGP signature is observed with the PTEN+ (0.79), RAS (0.78) and the Mesenchyme (0.72) gene-sets. We have not included this in the manuscript.

Minor Comments:

The PRC signature is misspelled to PCR signature in several places.

This has been corrected

Reviewer number 4

Next, the authors demonstrate the significant impact that sample composition has on subtype assignment, reporting that unassigned samples were composed of nearly equal parts cancer and stroma while samples of higher cancer to normal ratios were more often assignable. This is somewhat troubling, as 31/40 normal samples were assignable, yet heterogeneous samples, whose overall gene signature is highly influenced by the normal expression signature, were paradoxically obscured from assignment.

In our opinion this is not paradoxical. The normal samples are not confounded by cancer tissue, and are thus easier to assign, while cancer samples confounded with a high amount of normal tissue will have signatures consisting of mixtures of cancer and normal signals, and are thus more difficult to assign.

An important finding, and one that seems to be reiterated in the field is that gene set signatures related to poor and good prognosis show little relationship to Gleason score. The robustness of this concept is demonstrated by the fact that distinct samples from the same patient but with different Gleason scores show high gene signature correlation (though we are not told if these distinct samples are assigned to the same subtype)

If two samples have a similar signature, they are also very likely to be assigned to the same subtype. A signature in one sample may of course be stronger than in another sample (for example due to tissue composition), leading to one sample being assigned, and one that remains statistically unassigned. However, we very rarely observed conflicting signatures in the same patient in our data (though we did observe this on a few occasions in the new, additional validation data)
1) How were the original 21 gene sets organized in the final 15 gene sets?

*This organization is illustrated in Supplementary S2. Scores were averaged over related gene-sets, as described in Methods.*

2) Within the gene sets you combined to form the final 15, is it possible that certain sets could have a stronger influence on the composite gene set score? Could this obscure the ability to assign samples to a subtype when assignment is possible if the 21 gene sets were utilized as originally proposed?

*We do not think this is likely. Gene set scores from related gene-sets were highly correlated across all samples (0.95 for our data, and 0.89 for the additional validation set), so combining them should make sense.*

3) It seems to me that the approach to correlating sample signatures to subtype signatures might result in artificial subtype assignment. Assigning a sample to a subtype based on the correlation with that subtype being “significantly better” than the correlation to other subtypes does not seem rigorous. Did you explore an independent correlation strategy?

*We agree that there are probably exists many more assignments strategies, with various levels of sophistication. We have not explored on this further, but the best method for assigning samples in a clinical setting is an important topic which should be investigated. We have assessed FDR-levels by random shuffling of genes and gene-set scores.*

4) I would also like to see the authors explore a normal signature subtraction method using an “individual normal signature” in cases where a matched normal sample is available for a particular cancer sample. The method utilized seems problematic, as Figure 1C (Top – Normal) displays a dramatic left skew with outliers that could negatively impact the assembly of an average normal signature. Perhaps characterizing the gene set signature differences in normal samples from patients with different Gleason scores would lead to a more refined method.

*We actually tried to subtract the signature for matched normal samples, but this did not give any better results. This may be due to noisy variations in each individual sample, while the noise is averaged out when all samples are considered. We have not investigated whether removal of the three outliers would improve the result. However, we think that three outliers in 40 may not influence the results dramatically.*

5) Throughout the results section, it would be appropriate for the authors to describe how many of the samples they are describing are from a single patient. It is critical in allowing the reader to judge the robustness of the approach, and readers should be able to identify redundancies within the data.

6) Regarding Comment 5, the impact of this work is severely limited by the small size of the cohort, which is in reality only 41 patients. I would imagine that several of your significant results would be lost if you performed these analyses based on a single, composite gene set signature (normal and cancer) for each patient. For example, clusters 1, 2, and 3 in Figure 2A are composed largely of multiple samples from the same patient. You would very likely lose cluster 3 and perhaps cluster 2 if you considered patients rather than samples.
The thoughts behind the study was partly to investigate how the signature would vary in different samples taken from the same patient, where each sample was classified with its individual Gleason score independent from the other samples, rather than a single Gleason score from each patient. The two smaller clusters were also emphasized to illustrate the within-patient homogeneity in contrast to between patient heterogeneity. Samples with similar signatures from the same patient are partly also visualized in the heatmap. We have not tried to make a composite signature for each patient here, though we agree that this would be a relevant issue to study.

The authors are very unlikely to find a validation cohort of appropriate size for which clinical follow up data and thorough pathologic analysis (such as that required for their subtraction method) is available.

This is true. We did identify a promising dataset, but not of appropriate size, and the results turned out to be inconclusive, especially regarding the good prognosis assignments.

The manuscript has been seen and approved by all authors.

Thank you for your kind consideration.

Morten Rye, PhD
Department of Cancer Research and Molecular Medicine
Faculty of Medicine
Norwegian University of Science and Technology
P.O. Box 8905, NO-7491 Trondheim, NORWAY
Phone: +47 72573415
E-mail: morten.rye@ntnu.no