Author’s response to reviews

Title: Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients

Authors:

Petr Nazarov (petr.nazarov@lih.lu)

Anke Wienecke-Baldacchino (anke.wienecke-baldacchino@lns.etat.lu)

Andrei Zinovyev (andrei.zinovyev@curie.fr)

Urszula Czerwinska (ulcia.liberte@gmail.com)

Arnaud Muller (arnaud.muller@lih.lu)

Dorothee Nashan (dorothee.nashan@klinikumdo.de)

Gunnar Dittmar (gunnar.dittmar@lih.lu)

Francisco Azuaje (francisco.azuaje@lih.lu)

Stephanie Kreis (stephanie.kreis@uni.lu)

Version: 1 Date: 19 Jun 2019

Author’s response to reviews:

[Formatted PDF version of the response (including 2 illustrating figures) is attached as letter_response.pdf]

Dr. Yongsheng Bai
Editor of BMC Medical Genomics

Manuscript number: MGNM-D-19-00126

Dear Dr. Bai,

Please find enclosed the revised manuscript and a detailed point-by-point response to the constructive criticism of the reviewers. We think that including the reviewer’s suggestions helped us to further strengthen the important conclusions of our work and demonstrate the applicability of the approach. We thank the reviewers for this and we have addressed all points raised.
The reviewer’s comments are repeated below (marked by "->") followed by the answers (marked by "Answer:"). In addition to changes in the text, we included 3 new figures and 5 subfigures to the main text and 2 new figures in the supplements.

Many thanks for considering our revised paper. We hope that the changes and additions meet with your approval and look forward to hearing from you.

Sincerely yours,

Petr Nazarov
Multiomics Data Science Research Group, Quantitative Biology Unit,
Luxembourg Institute of Health
1A-B, rue Thomas Edison, L-1445 Strassen
Luxembourg
Tel: +352 26970-385
E-mail: petr.nazarov@lih.lu

==========================================================================================================

Point-by-point response

Enrico Glaab (Reviewer 1)

General comments:

->The authors present an approach using Independent Component Analysis (ICA) to decompose cancer-related transcriptomics datasets in order to address technical biases for cross-study data integration, train machine learning models to predict clinically relevant outcomes and identify tumor-associated cellular processes. The proposed approach is based on existing computational methodologies, but its application in the field melanoma transcriptomics data analysis is new and provides an interesting alternative to common integrative omics analysis approaches. My comments mainly refer to the discussion of the benefits of the proposed Consensus ICA method in relation to classical ICA, the comparison of the method to alternative dimension reduction approaches, and questions concerning the evaluation, e.g. the separation between samples from the reference dataset used for parameter optimization (choosing the number of ICA-components such that the classification performance on the reference set is maximized) and samples used for cross-validation based performance assessment.

Answer: We are very pleased that the reviewer found that our work is an interesting alternative to common integrative omics analysis approaches and. Also, we thank him for the important suggestions, which we implemented as detailed below.

Specific comments:

->1. The authors mention that for the ICA-based machine learning analyses the number of components was chosen based on the ability of ICA features to classify patients in the reference set (page 15). Since a leave-one-out cross-validation using Random Forest classification on the same reference set was also used as part of the performance evaluation (page 16 and Table 1 on page 18), it would be important to
ensure and clarify that the parameter optimization for choosing the number of components is conducted in a nested cross-validation or on a hold-out dataset partition that is separate from the data used for performance testing, i.e. making sure that the same outcome information used for parameter optimization is not used to assess the predictive performance of the method, in order to prevent biased accuracy estimates. Apart from the LOOCV analysis presented in table 1, the authors use further microarray and in-house clinical investigation data for the evaluation, but given the lower sample size of these datasets, an additional nested cross-validation on the reference dataset would provide a more robust evaluation. Since the class labels are not balanced for all considered outcomes and the performance estimates may vary across the different LOOCV cycles, I suggest adding sensitivity and specificity statistics to table 1, and the standard deviations of the accuracies.

Answer: We agree with the reviewer. Therefore, we improved the cross-validation and added the required information – confidence intervals for accuracy, together with sensitivity and specificity, into Table 1.

> In addition, as recommended, we performed the nested cross-validation to identify the optimal number of components in ICA and checked the reproducibility of the reported accuracies. For the nested CV, we excluded 20% of the samples, estimated the number of components and trained random forest on another 80% of the samples. Then, we predicted the classes of the 20% samples excluded. Five runs of this cycle were performed in order to ensure reproducibility. These results are now described in Supplementary Methods (Additional File 1).

Answer: We also introduced changes in the text of the Methods (pp.12-13) and Results (p.19) to reflect these improvements.

> 2. The manuscript proposes a consensus ICA approach as a robust alternative to the standard ICA method. This consensus ICA method involves repeated applications of the ICA method on reduced versions of the input dataset, derived from excluding random samples from the analysis, and then averaging reordered weight and metagene matrices. It is plausible to expect more robust results from this approach than for the standard ICA; however, a comparison of the results in terms of robustness or performance in subsequent machine learning analyses using standard ICA as a reference should be shown to enable an assessment of the extent and significance of the differences.

Answer: We are thankful for this important comment. Indeed, the importance of the consensus ICA was recently discussed by one of the co-authors of this paper in Cantini L et al., 2019, Bioinformatics [https://www.ncbi.nlm.nih.gov/pubmed/30938767]. The required reference was inserted into the text. In addition, here we independently addressed the effect of consensus ICA on classification and reproducibility of the results and added a paragraph to the Results section. We added the Supplementary Figure S4 showing this improvement and modified the text accordingly (p.16). The level of the improvement was measured by the coefficient of determination between metagenes in the repeated analyses (R2) and Jaccard indexes between the most contributing genes.

> Since various other dedicated cross-study normalization approaches have been proposed (e.g. ComBat, XPN), a short discussion should also be provided on the advantages and limitations of the ICA-based data integration method in comparison to these alternative methods, which do not involve dimension reduction but may provide competitive benefits in terms of technical bias removal.

Answer: As recommended, we investigated the effect of normalization using ComBat and XPN. Indeed, in terms of the first principal components, the new data become more similar to the reference
dataset. However, this did not improve the accuracy of classification. We considered PCAs of ComBat- and XPN-corrected data, together with other dimensionality reduction methods proposed below and added Figure 3. Corresponding modifications were added to Methods (p.12-13), Results (p.20) and Discussion (pp.29-30) sections.

> 3. The proposed approach is compared to a PCA dimension reduction, and the authors report that the ICA-derived features showed higher statistical significance in terms of ANOVA-based p-values, and higher AUC for gender and sample type prediction, and lower AUC for tumor subtype prediction. Given that the performance estimates may partly differ due to stochastic variation in these estimates, a statistical test (e.g. the Friedman test in combination with a post-hoc analysis) should be applied to assess the significance of the difference across the performance statistics. Since PCA has been superseded for many biomedical data analysis applications by alternative dimension reduction approaches (e.g. t-SNE, LLE, Isomap), I suggest including also a more recently developed dimension reduction approach in this comparison.

Answer: We would like to thank the reviewer for raising these interesting points. However, we cannot use Friedman's test here. It is a non-parametric test equivalent to 2-way ANOVA without replication. In our situation, one factor would be “groups of interest” (i.e., methods of feature selection), another one would be “blocks to be corrected for”, such as components. However, we cannot block independent and principal components as they are not linked. Instead, we used here three alternative approaches to support our claims.

First, we performed Wilcoxon rank sum test and reported the p-value in the titles of Supplementary Figure S3.

Second, as an alternative to Friedman's test, we performed 2-way ANOVA on log-transformed p-values. Although, x=log(p-values) does not follow a normal distribution, it is much less skewed compared to original p-values. In addition, application of ANOVA to non-normal data does not affect Type I error and results in a pessimistic estimation of the significance (e.g. Blanca MJ et al., Psicothema, 2017, 29:4 and Khan A, Rayner GD, Journal Applied Mathematics and Decision Sciences, 2003, 7:4). In our analysis, one factor was linked to the classification task (Gender, Sample type, Subtype) and another to the method: PCA or ICA. Only the second factor showed significance (p-value=0.0175). Post-hoc analysis with TukeyHSD confirmed that on average, ICA gave lower p-values.

Third, we compared several methods proposed by this and other reviewers (batch correction, t-SNE, LLE, Isomap, NMF, LRAcluster) by classification accuracy. The results were added in Figure 3. ICA showed high accuracy comparing to other considered methods of batch correction and is only in one comparison slightly lower than the results of NMF (0.869 +/- 0.001 and 0.881 +/- 0.002).

The text was modified accordingly in the Methods (pp.12-13) and Results (p.18-19) sections.

Minor points:

> 1.) Page 6, line 15: I suggest replacing "and other" by "and others" or by "among others".

Answer: Corrected

> 2.) Page 10, line 20: The code on https://gitlab.com/biomodlih/consica uses the sample-function but does not set a seed number for the random number generator to ensure reproducibility of the results. I suggest adding a line at the beginning which calls the set.seed() function with a chosen seed number for this purpose.
> 3.) Page 11, line 10: It is not clear what is meant with "16579 informative genes" here - how is the informativeness assessed and is the selection threshold for informative genes set appropriately (16579 genes is a large fraction of the entire genome, and probably not all of these genes are informative with regard to the biological outcomes of interest).

Answer: We agree with the reviewer. Here we mean genes that were expressed above the selected threshold (log2 > 5) in at least one sample of the reference dataset. The text was improved accordingly.

> 4.) Page 13, line 12: "through highest correlation" - does this mean highest absolute Pearson correlation? Please clarify.

Answer: The text was modified accordingly

> 5.) Page 28, line 12: I suggest replacing "Taken together, consensus ICA approach..." by "Taken together, the consensus ICA approach...".

Answer: Corrected

Shulan Tian (Reviewer 2)

General comments:

> The authors presented a nice workflow that combines consensus ICA with the associated downstream analysis, which enables the identification of tumour-related biological processes in patients based on previously collected large reference datasets. By analyzing three datasets (reference, validation and patient samples), they demonstrated that the workflow can potentially predict cancer subtypes and estimate the activity of key tumour-related processes.

To increase the prognostic power of individual components from ICA for survival prediction, they proposed a hazard score (HS) that integrates weights of several components. The applicability of the proposed score was checked using an independent validation set.

Finally, the authors concluded that the method can be used to map new transcriptomic data from cancer patient samples onto large reference datasets and provide prognosis for the patient survival. This is a comprehensive and well-documented work. The available scripts used for the analysis and the supplemental files are very helpful. The work is worth publishing in the journal as a valuable resource in systems biology.

The paper was well written and the analyses were fully documented. I only have a few minor comments below.

Answer: We would like to thank the reviewer for this kind summary.

> 1). In page 12, line 2, "A p-value was individually assigned to each gene/miRNA within each component, based on the probability that it came from a normal distribution with estimated parameter" What is the rationale for assuming a normal distribution? Better to explain this.

Answer: The text in the Methods (p.11) was improved in order to answer the question. The selection of
the normal distribution is linked to the fact that ICA by design tries to identify non-Gaussian signals in the Gaussian mixtures. Therefore, the genes with normal distributions in rows of S matrix are considered as "background" and do not belong to the top contributing genes of the component.

> 2). In page 12, line 8, it would be helpful to list (mention?) the number of genes made to the two lists of top-contributing genes in the main text (p-value (adj.p-value) <0.01)

Answer: Indeed, such information helps to better understand the method. We added a reference to Additional File 2 (analysis report) that contains the number of genes as well as the top genes. In addition, we added legends into Additional File 2.

> 3). In page 12, line 3, what is the confidence level (probability of outcomes) for RF to predict class for the new clinical samples?

Answer: We appreciate this suggestion. The information was added to the Table 1. This also explains why NHEM was considered as metastatic – it had a border probability of 0.51 (p.19).

> 4). In page 13, how robust it is to use the hazard score for survival prediction?
Suggestion: please extend line 15 "The applicability of the proposed score was checked using the independent validation set"

Answer: We improved the text as suggested, stressing that the validation dataset was based on an independent patient cohort, studied with a different technique: microarrays instead of sequencing. Text is added at Methods section (p.14).

> 5). In page 16, line 1, possible typo " Ne number of components"?

Answer: Corrected

> 6). In page 19, line 14, it would be nice to clearly point out whether all kept components that can be assigned to specific biological processes

Answer: Eight out of 11 significant RICs were linked to biological processes. However, 3 had a limited list of significantly contributing genes and we were not able to link them directly to biological functions. Nevertheless, the profile of their weights was similar to those RICs that are linked to immune response and allowed us to include them in the immune cluster (Fig.5D). The text has been improved accordingly.

Syed Haider (Reviewer 3)

> Authors propose an application of Independent Component Analysis (ICA) towards predicting clinical covariates of melanoma. They profile 3 primary melanoma samples and 2 control samples (called as Investigation dataset) and combine this with TCGA dataset (called as Reference dataset) to identify ICAs. Authors demonstrate insensitivity of ICA towards cohort of origin thereby identifying components which may be reflective of true residual signal. These results are compared to Principal Component Analyses.
Major:

> 1. Background page 6, line 9: The use of hazard score as a terminology is misleading and should be at the very least defined first. Ideally, authors should use a different terminology. Similarly, elsewhere in the manuscript, the use of terminologies such as Investigation, Reference and Validation sets are not ideal. Authors could use terms such as "Discovery" and "Validation" for instance.

Answer: Instead of "hazard score", now we use "risk score", which is a less ambiguous term in the context of our article. The definition of the risk score is given in the Methods section. In addition, we followed the reviewer's recommendation and replaced the term "Reference dataset" by the more standard "Discovery dataset" throughout the text.

> 2. Authors apply ICA to a combined (5 in-house patients + TCGA) cohort and compare the results with PCA, and demonstrate ICA is more sensitive to cell-specific signal than the cohort of origin. Although interesting, this is simply profiling differences between ICA and PCA. It remains insufficient to compare ICA with PCA on one dataset and draw conclusions; and more importantly it is unclear what is the scope of this study? If it is meant to benchmark these methods, why not compare the two methods, alongside a panel of other variation and model-based methods and across a number of datasets.

Answer: We present a method that can be used to map new transcriptomic data from cancer patient samples onto large discovery datasets. The method corrects technical biases, extracts features that can be used for classification of the new patients, helps characterizing activity of biological processes or cell types in the new samples and provides prognosis for patient survival. The purpose was not to benchmark ICA against PCA but to demonstrate that ICA is applicable for the listed purposes. However, to address this point, we have included a new Figure 3 where we show the results of benchmarking of ICA against 8 other dimensionality reduction methods in terms of feature selection. From our experience (in parallel projects we worked on glioma and lung cancers), the results of ICA are more biologically interpretable and ICA produces better features for classification of the samples. Related modifications were introduced to pp.12,13,29,30,32

> Also, would PCA be more similar to ICA if the 5 in-house samples were not analysed with TCGA for the PCA analysis?

Answer: ICA and PCA capture different properties of the data. In particular, PCA aims at the direction of the highest variability, while ICA tries to decompose mixed signals into statistically independent sources. Both methods can be used for dimensionality reduction and feature selection, but the direct comparison between them may be questionable. However, we mapped PCA- and ICA-based components using R2 between profiles over the samples (components with the highest R2 were linked) and did not see the improvement.

> 3. In section "ICA yields clinically relevant information", it appears authors apply the features found in the combined dataset (Investigation+Reference) to Reference dataset itself to assess predictive potential. This is insufficient validation, authors should use an independent cohort to validate their findings.

Answer: We addressed this critics by two modifications. First, as recommended, we checked the predictions made by the random forest classifier on the validation dataset. Predicting patient gender showed high accuracy of 0.977 (only one female was
Testing sample type (primary/metastatic) for this validation cohort resulted in 34 samples classified as metastatic and 10 as primary (accuracy = 0.773, as all validation samples coming from metastatic). However, as the precise excision location of the tumours is unknown, we cannot exclude that some metastasis came from skin. Indeed, 7 of 10 misclassified samples showed high expression of keratinocyte marker genes (KRT5, KRT14).

Second, we performed additional validation by a nested cross-validation on the discovery dataset. At each iteration, 20% of the samples were excluded from the analysis. We defined the number of ICA components using the remaining 80%. Then, we performed ICA on the entire dataset and predicted labels for the excluded samples. By this, we ensured that no bias is introduced to accuracy calculation by optimization of the meta-parameters, such as number of components. The obtained accuracy results were very similar to those reported in Table 1. These results are now described in Supplementary Methods (Additional File 1).

We introduced changes in the text of the Methods (p.12) and Results (p.19) to reflect these improvements.

> 4. In section "ICA provides prognostic features linked to patient survival", authors find that the univariable prognostic features identified in the Investigation+Reference dataset did not validate in the Validation dataset. To overcome this challenge, authors derived HS (hazard score), which also did not validate in the validation set (Note: on page 17, line 21, HR=0.97 is rather meaningless in this context. I am also concerned how this was found to be significant). This questions the scientific advance of this section of the study.

Answer: Indeed, here log-hazard score was presented. We changed confusing abbreviation to LHR to avoid misunderstandings. We also noticed a mistype in the main text, as HLR was 0.87 with a confidence interval of 0.28 – 1.45 and a p-value of 1.3e-3 (the correct value was shown in the supplementary figure, which is now Figure 4). Therefore, the calculated risk score can be used for building prognosis in this independent validation cohort.

> Would this be because the validation cohort is entirely Metastatic patients?

Answer: The main predictive components were linked to immune system. Both primary and metastatic patients in our discovery cohort showed the same tendency: higher immune response positively correlated with survival. This was also showed for the validation cohort by Bogunovic et al. Therefore, here we rather face a well-known challenge of reduced reproducibility of survival markers between independent studies, which can be at least partially explained by the difference in the analysis platforms.

> It is also crucial to clarify whether the metastatic patients' biopsies were from the primary tumour or the metastatic site.

Answer: The precise excision location of the samples is unknown, but they were collected from metastatic sites. Many samples came from lymph nodes. We clarified this in the Methods section (p.7).

> 5. Page 19, section "Immune components" lines 8-10: Hazard Ratio (HR) is expressed as -0.89. This is incorrect as by definition, unless it is in logarithmic scale and if that's correct, it should be written appropriately. This is the case in several sections. This makes it in unclear to review this section.

Answer: Indeed, it was log-hazard ratio. The abbreviation is now corrected accordingly throughout the
> 6. Page 18, section "Independent components provide information about biological processes in tumours - General Strategy": The automatic reports which enable ICA mapping to biological processes is potentially an interesting piece of generic software which authors should make available as a part of this manuscript.

Answer: As was mentioned in the section "General Strategy", the automated report is a part of Supplements and is provided in Additional File 2 (Supplementary results). We improved this file by adding a page with legends.

> 7. Page 18, section "Independent components provide information about biological processes in tumours": Throughout this section, authors refer to 'new samples' which makes it difficult at times to understand whether they are referring to the "validation" cohort OR "Investigation" cohort.

Answer: We corrected the text accordingly, replacing "new samples" by the "investigation" cohort.

> If it is the latter, then why none of the data shown here was validated against the truly "Validation" cohort? Authors should show that data in order to demonstrate robustness of their findings beyond the "Investigation+Reference" cohorts which were used to discover the findings.

Answer: Validation cohort was introduced to test the risk score performance on an independent dataset. The sample acquisition for this dataset was less controlled compared to our clinical samples (investigation dataset) and samples are less annotated: apart from gender, age and survival there are no information available. In addition, there are no miRNA data associated with these samples. Therefore using this dataset to discuss biological processes that are captured by the components is not optimal. However, following the recommendations, we added Supplementary Figure S7 (Additional file 3) and discussed behaviour of several components in the corresponding section (pp. 24,26,27).

> 8. Authors claim that ICA deconvolution can identify components representing tumour or stromal cells; however this is inconclusive from the data shown here. Authors should at least compare correlation between identified tumour ICA components with tumour purity estimates of the same cohort.

Answer: We fully agree with this important recommendation. We compared the components with immune and stromal score calculated by the ESTIMATE tool (Yoshihara et al, Nat Commun 2013, https://bioinformatics.mdanderson.org/estimate/index.html) and confirmed that our immune components strongly correlate with ESTIMATE results and added this Figure 5 F,G. The immune signal found by ESTIMATE is, in fact, captured by a single component RIC25, which is linked to T-cells activities, most probably. However, our method allows identifying additional immune components, such as RIC2 and RIC27 (B cells), RIC57 (monocytes) and RIC37 (IFN signalling pathway).

The Results section was improved accordingly (p.22, 23).

Minor:

> 1. Abstract: 'tor' should be 'for'.

Answer: Corrected
> 2. Background: Until the line 16, authors have made a lot of statements without any citations. This is rather inappropriate to make such claims without citing the original studies.

Answer: It is not clear to which page the reviewer is referring to, but we added references both to page 4 and 5, related to heterogeneity of cancer and the in-silico deconvolution methods.

> 3. Background page 5, line 12: 'combing' should be 'combining'

Answer: Corrected

4. Results Line 1, page 15: 'Ne' is incomplete for?
Answer: Corrected ("The")

Jin Gu (Reviewer 4)

> This manuscript uses ICA to separate the mixture signals in bulk omics data. Results show that the components can capture biological meaningful patterns. I agree that ICA is a useful method to extract principal patterns in high dimensional omics data. Some major issues:

Answer: We thank the reviewer for the comments and tried to implement them in the new version of the manuscript.

> 1) ICA has different assumption of components than PCA. I suggest that the authors should discuss this point in depth and analyze theoretically the advantages/disadvantages of ICA;

Answer: Such discussions could already be found in literature, for example in the previous work of our co-authors: https://www.ncbi.nlm.nih.gov/pubmed/23261450 https://urszulaczerwinska.github.io/DeconICA/DeconICA_introduction.html and other research groups, e.g. https://www.ncbi.nlm.nih.gov/pubmed/15022635

We feel that adding theoretical discussion and more deep comparison of ICA to PCA would reduce the focus of the paper. Nevertheless, we added several dimensionality reduction methods, analysed their performances and added the data as a new Figure 3. From our experience in this and other projects, the results of ICA are more biologically interpretable than PCA and other considered dimensionality reduction methods.

Modifications were made to Methods (pp.12-13), Results (p.20) and Discussion (pp. 29,30,32) sections.

> 2) Except ICA/PCA, low-rank matrix approximation & non-negative matrix factorization is also useful to find omics patterns. Please include these kinds of methods, such as LRAcluster and jNMFMA;

Answer: We thank the reviewer for this recommendation. We compared our method to LRAcluster, developed by the group of the reviewer, and included this method into a set of other methods tested.

The results – accuracy of classification using other methods, – are shown in the new Figure 3. LRAcluster showed better results than PCA, but still lower than ICA.

We were unable to investigate the second method, as the jNMFMA webpage is no longer available.
Instead, we considered a well-established NMF method (R package `NMF`) and obtained results comparable to ICA in terms of classification accuracy. However, based on the recent work of one of the co-authors (https://www.ncbi.nlm.nih.gov/pubmed/30938767) and other evidences, NMF is less stable compared to ICA. It is hard to expect high reproducibility from NMF unless initial estimations for the decomposition matrices are fixed or/and additional limitations (e.g. – signatures of cell types) are used. At the same time, consensus ICA showed high stability of the majority of defined components (see increase of stability in Supplementary Fig.S4). We adapted the Methods (pp.12-13), Results (p.20) and Discussion (pp. 29,30,32) accordingly.

> 3) The results are too preliminary. The manuscript only included three rude figures. At least, the signature genes of the selected components and the p-values of the enrichment analysis should be presented as tables & figures. The correlations of identified signatures should be presented as figures. I think that it should design minimal 6-8 figures or tables.

Answer: We improved the manuscript and have addressed this concern by putting more validations and essential figures into the main part. The manuscript includes 6 figures now. The signature genes and enrichment analysis with corresponding FDRs are available as Additional Figure 2. The correlation between signatures (in terms of S matrix) is non-informative as all correlations between columns of S-matrix are around 0. Instead we put r² for M-matrices, based on which we connected the components, into Fig. 5 (here r² is more preferable than simple correlation for visualization reasons: it gives "clearer" grouping of the components). In addition, in Fig 1 we presented a global overview of the expression data, S and M matrices.

> 4) Some typos exist.

Corrected