Author's response to reviews

Title: Genomic investigation of etiologic heterogeneity: methodologic challenges

Authors:

Colin B Begg (beggc@mskcc.org)
Venkatraman E Seshan (seshanv@mskcc.org)
Emily C Zabor (zabore@mskcc.org)
Helena Furberg (furberga@mskcc.org)
Arshi Arora (aroraa@mskcc.org)
Ronglai Shen (shenr@mskcc.org)
Jodi K Maranchie (maranchiejk@upmc.edu)
Matthew E Nielsen (mnielsen@med.unc.edu)
W. Kimryn Rathmell (rathmell@med.unc.edu)
Sabina Signoretti (ssignoretti@partners.org)
Pheroze Tamboli (ptamboli@mdanderson.org)
Jose A Karam (jakaram@mdanderson.org)
Toni K Choueiri (toni_choueiri@dfci.harvard.edu)
A. Ari Hakimi (hakimia@mskcc.org)
James J Hsieh (hsiehj@mskcc.org)

Version: 2
Date: 13 November 2014

Author's response to reviews: see over
November 13, 2014

Arlene Pura
Journal Editorial Office
BioMed Central

Re: Resubmission of MS 1009958161417072

Dear Ms. Pura:

We have responded to the comments/concerns from the reviewers in the following ways:-

Associate Editor:

“provide a clear justification for...choice of 4 subtypes in this work.”
We have addressed this issue in some detail. First we describe more carefully the test we used to determine if the addition of a sub-type significantly increases the heterogeneity signal (lines 261-272). Repeated tests of this nature provide an obvious rule for determining the number of clusters. However, since there is a “true” underlying set of sub-types one would hope that the optimal number determined from this process would be congruent across all platforms. Data on these test results are now presented in a new Table 1 and in new text (lines 304-314), but unfortunately they do not point clearly to an appropriate unitary number of clusters. So as before we present in detail subsequently the results for the 4-class solutions for each genomic platform. This allows us to easily compare the classes based on the different platforms and with the published TCGA results which were based on 4 sub-types. These issues are also described now in the limitations section of the Discussion (lines 459-468).

“Formula of D- and its brief introduction should be given in the main manuscript because the analyses in this study used D- instead of D. (2) In order to analyze data in a case-only study, authors derived D- from D by assuming \{r_i/\mu\} = 1 in the formula of D. This is a strong assumption. This assumption should be evaluated and discussed in details and listed as a limitation of this study in section Discussion.”
We have revised and moved all of the material describing the calculation of the modified D-statistic in the context of case-only sampling from the Appendix to the main text (lines 229-260). We provide a strong heuristic argument as to why this is likely to lead to minimal (if any) differences in the results as would be obtained if controls were available. In principle the appropriateness of this substitution could be evaluated empirically using data from a case-control study where we could compare the rankings of D and D* but unfortunately we cannot do this since we have no controls. The issue is
also elaborated upon in the Discussion (lines 452-458). We really don’t view this as a “strong assumption”. The essential ingredients of the heterogeneity measure (in either the case-control or case-only setting) involve terms that capture differences in the risk profiles of the case sub-types. These differences do not involve controls at all. The controls only come into play because we have elected (for reasons explained in the text and in more detail in our earlier work) to use a population-based measure of heterogeneity that involves averaging over the population at risk. In the absence of direct information about the population at risk it seems eminently sensible to use a measure that captures the crucial information even though it implicitly weights the contributions from the cases somewhat differently. In other words one can argue that D* is an appropriate measure to use even if it doesn’t line up perfectly with D.

Reviewer: Debasish Ghosh

1. The lede is buried very much towards the end, but the authors themselves admit on line 403 that "we cannot assert with confidence that 4 is the appropriate number of subtypes." This suggests that the method is not very numerically stable. Some better justification for the number of clusters is needed. Much additional material has been included on this issue as described above in the response to the first comment of the associate editor.

2. lines 203 and 210: the authors should specify exactly how many terms are being added up in the equation. The current presentation is too casual. The formulas and text have been modified to clarify the number of terms in these equations.

3. line 241. The authors used what appeared to be a self-competitive test in the jargon of Goeman and Buhlmann. Can they explain why they used this rather than a self-contained test? We address this issue at the end of the Methods section with a reference to Goeman and Buelmann (lines 292-294). Some of the t statistics that form the raw data for this test necessarily arise from differentially expressed genes since the clustering is driven to identify sub-types with differential expression. However the purpose of the test is simply to see if the extent of such differential expression is related to the pathway under evaluation. As such this defines a competitive test.

4. lines 252 - 254: The authors should show a plot to verify this phenomenon. This is now addressed in the new Table 1 and explained in our response to the associate editor (above).

5. line 285: A useful rule of thumb for how many starts are needed for the k-means clustering should be given. The issue is now addressed on lines 188-190.

Reviewer: Qing Lu

1. Page 7, line 152– It seems that a substantial proportion of genes have been filtered out before the cluster analysis. Will the result change if a loose criterion is used and more genes are used for the analysis? We clarify our filtering strategy and its rationale in the section entitled “Data” (lines 149-161).
2. Page 7, line 158 – Has a filter algorithm also been applied to the sequencing data. If so, how many SNVs remained for the analysis? We used TCGA reported mutations without any adjustments (lines 161-163).

3. Page 9, line 187 – Can the authors explain in more detail how they evaluate the statistical significance of the increase in etiologic heterogeneity? This is now explained in detail in the “Analytic methodology” section and is further discussed in our response to the associate editor (above).

4. Page 12, line 252 – Analytical details regarding to the choice of 4 clusters are missing. Why 4 is selected as the best cluster size based on mRNA expression data? Did the authors obtain the same cluster size from the analysis of methylation, copy number of variation and exome sequencing data? See response to associate editor above.

Thank you.

Colin Begg
Corresponding Author