Reviewer's report

Title: Genomic investigation of etiologic heterogeneity: methodologic challenges

Version: 1 Date: 12 October 2014

Reviewer: Qing Lu

Reviewer's report:

In this paper, Begg et al used genome-wide gene expression, methylation, copy number of variation and exome sequencing data to study the etiology heterogeneity of cancer. Based on a clustering algorithm, they identified 4 tumor subtypes that are etiologically distinct. Overall, the paper is well written, and is of current research interest. I have a few comments.

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?

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?

3. Page 9, line 187 – Can the authors explain in more detail how they evaluate the statistical significance of the increase in etiologic heterogeneity?

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?

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interest