Reviewer's report

Title: High-Resolution Analysis of Copy Number Alterations and Associated Expression Changes in Ovarian Tumors

Version: 2 Date: 16 December 2008

Reviewer: Giovanni Tonon

Reviewer's report:

The manuscript "High-resolution analysis of copy number alterations and associated expression changes in ovarian tumors" by Haverty et al. is analysis of the genomic alterations in ovarian cancer, considering and comparing different ovarian cancer subtypes. It should be mentioned that extensive analysis has been already conducted on this disease, but the Authors provide some interesting perspective and apply novel tools that make this manuscript interesting.

Major Compulsory Revisions

As a major comment, the data are not always properly presented and their conclusion adequately supported. On the overall, the material and methods section is not explained in enough details. Moreover, few references are provided in this section.

1. for the comparison of the aCGH data with the reported normal, germline CNVs, the Authors should include the most recent analysis published, that has extensively extended the number of germline CNVs.

2. It should be explained in some detail why “outliers in the segmented data were removed by merging segments of < 8 probe sets to the neighboring segment with the most similar ILR.” In this way, the most focal, potentially more interesting CNAs are excluded.

Also, given results in other papers, showing quite dramatic CNAs, it should be explained the reasoning behind choosing such a low thresholds for calling gains and losses.

3. I acknowledge that it is difficult to select among the expression probes. However, I am not sure whether the method chosen by the Authors is the most appropriate (“For each gene, the probe set with the highest median expression was selected unless noted otherwise.”). In my own experience, the most reliable Affymetrix probes are the ones showing the highest degree of variance.

4. The Authors should explain why they used MAS 5 for normalizing and processing their own data, and the more reliable RMA for the published dataset. I would have used RMA for both datasets. I am not convinced about the methodology for correcting for the batch problem, a notoriously impervious
problem. The Authors should provide references and more details on the methodology.
Also, comparing two different platforms poses some issues, that the Authors should discuss.

5. For comparing the expression of normal and tumor samples, the Authors have used again ANOVA and then corrected for multiple testing using FDR. Since most people use instead SAM, the Authors should explain why they choose this alternative method, and how it compares with the standard methodology. Also, the Authors should explain why they chose to use the fold-change as a threshold after correcting for multiple testing.

6. The authors should explain why, for the enrichment of cancer census genes, they have used “genes in 5Mb and 10Mb regions centered on the midpoint of each extended peak region.” in the main text they mention that this method was used to show how moving far away from the peak, the cancer gene enrichment disappears. However, since the size of the peaks is different, I am not sure that this method is reliable. I would simply compare cancer genes in the peaks versus cancer genes outside the peak, as proposed in other parts of the manuscript.

7. I am not sure how the clustering shown in figure 1c is informative. Clustering array CGH data is notoriously difficult. The results presented in figure 2c indeed suggest that more than anything else, there is a difference in the serous adenocarcinomas versus the others, that is mostly due to a different level of rearrangements (as shown in Suppl.Fig.1, where some form of statistical analysis should be included). Moreover, the pattern in the serous adenos is quite homogeneous, as shown by the GISTIC.

8. In terms of CNAs identified, a more thorough comparison between these Authors findings and what has been previously shown in the literature should be included, possibly in the form of a supplemental table.

9. On the overall, comparing the expression in tumors with the lesion, without the lesion and normals poses some problems. In particular, very often genes are dysregulated, in both directions, by means that are independent from amplification or deletions. Therefore, comparing the expression in tumors with the lesion versus tumors without it could indeed spuriously eliminate important cancer genes, as I suspect is the case of MYC in this report. More pertinent is the comparison with normal samples, given however all the caveats linked to the use of whole-ovary and not the somehow more specific epithelial cells.

10. The Authors provide an interesting association analysis between gains/losses. It is not clear to me why aberrations present on the same chromosome are excluded (they could be inversions).

11. some of the genes taken into consideration fail to show up-regulation relative to normal samples and in one case (CLDN11) there is even lower expression in comparison to normal. Authors should comment on the lack of overexpression of such genes. In fact, Claudin 11 has been recently shown to be upregulated in
ovarian cancer as well in the Author's analysis of their own data, may be the Authors should consider, in a few selected cases, to confirm the expression array results with quantitative PCR. They also presented data on LCM samples, proposing interesting question about cell-type specificity of the expression of the genes, but then any comment on that is missing.

Minor Essential Revisions

1) Data of the same type are presented sometimes as figure in the text and some other times as Supplementary figures (e.g. data on chromosome 3 and 20 in fig.4 and 5, respectively, and on chromosome 8 in fig.S4); I would suggest to put all similar data together in the text. Also data are often repeated between tables in the text and tables in the supplemental data. I could combine both data into a single table, to be included in the main text.

2) Data presented in the figures are sometimes not well explained and figure legends are not helpful in understanding of data (e.g. fig.3); I would suggest to explain in more details what is shown in the figures.

3) Several typos, and often the English is not clear. A more careful editing of the text is warranted.

4) The acronym ILR is explained only late in the manuscript.

5) This sentence is obscure: Affymetrix MAS 5.0 signal intensities were scaled to a mean, trimmed 2%, of 500.

Level of interest: An article whose findings are important to those with closely related research interests

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

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

Declaration of competing interests:

I declare that I have no competing interests