Reviewer's report

Title: Histotype-specific copy-number alterations in ovarian cancer

Version: 1 Date: 27 April 2012

Reviewer: Charles Warden

Reviewer's report:

Huang et al. investigate copy number alterations in different histological subtypes of ovarian cancer from 3 independent cohorts, and they claim this combined dataset is crucial because of the high frequency of serous histotypes compared to the other histological subtypes. I think characterizing genomic alterations that are unique to various histological subtypes is an important research topic for personalized medicine, but I do have some concerns about some of the specific results in this paper. If these concerns are addressed and the results still stand as significant, then I think this will be a highly accessed publication.

Major Compulsory Revisions:

1) I think the authors should be commended for addressing two of my greatest concerns at the end of the discussion, but I think these issues need to be addressed more rigorously and discussed in the results section.

1a) The first concern is eliminating bias that may be introduced from combining data from 3 different cohorts. The authors qualitatively say that the PCA plot doesn’t show distinct copy number alterations for the different datasets, but I think this needs to be a quantitative rather than qualitative statement. For example, an ANOVA test could indicate if the copy number varied among groups for a particular gene / probe. Each gene of interest (like ERBB2) could have a p-value indicating the likelihood that there may be bias for each signal and/or you could check for the number of genes / probes that do show variation among dataset with an FDR < 0.05 (hopefully, this number would be close to 0). Alternatively, the authors could factor out any potential bias between the studies. It is mentioned that different reference groups are used for the different populations, but this doesn’t rule out the possibility of technical bias between the data sets.

1b) The second concern is whether or not the combined dataset truly overcomes the problem of having small sample sizes for the endometrioid, clear cell and mucinous tumors. In general, I think two strategies that would help would be to provide statistical analysis for the comparisons and to test other publicly available datasets (I give more specific examples throughout the major and minor revision recommendations).

2) Does ERBB2 show differential expression when comparing mucinous and serous tumors? The copy number and expression values are correlated with one another, but is the difference in expression between the two groups significant?
Given the low frequency of copy number alterations, it may not be. However, this needs to be reported either way.

3) It would be useful to test if the ERBB2 alteration can be validated in an external dataset. I think there should be some copy number studies with mucinous tumors, but you can at least look at expression studies for mucinous vs. serous differential expression if the appropriate copy number studies cannot be found. For example, I believe the expO dataset (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109) contains 10 mucinous tumors and the data from Hendrix et al. (http://www.ncbi.nlm.nih.gov/pubmed/16452189) contains 13 mucinous tumors. Given that the combined dataset in this paper still only contains 14 mucinous tumors (19, if you count the borderline samples), I think it is worth searching for more data to compare.

4) If the authors have access to the original samples for the 3 cohorts, why was qPCR validation only applied to 7 samples? I think all of the mucinous and serious tumors showing copy number alteration should be tested.

5) I do not think the data in “Serous tumors had the most overlapping genes with other histotypes” currently warrants its own section. As mentioned in the discussion (“the higher number of copy number altered genes in serous tumors could be attributed to the larger sample size in this collection”), I think this result is due in large part to the sample size for the different histological subtypes. So, I think some sort of statistical analysis need to be conducted to determine if these overlaps are greater than would be expected due to the different sample sizes. The non-serous tumors seem to have roughly equal sample size, so the differences between overlap with those subtypes may indeed be significant. However, I think there are also other ways to make this overlap analysis more rigorous. For example, I believe alteration in a single sample is sufficient for inclusion in the gene list for each subtype. I think the overlap also needs to be shown with a threshold for percent of tumors represented (like 50% of tumors, or at least 25% of tumors). Finally, different areas of the genome have high and low densities of genes. I think the fraction of genome altered should be presented for the 4 subtypes, and I think it would be nice if the authors could present some data to indicate how likely the increased overlap is to be biologically significant (for example, copy number alterations overlapping clusters of olfactory receptor genes may include a large number of genes in a relatively small region, but my guess is that this may not mean the larger overlap corresponds to a larger number of driver alterations). In fact, Figure 1 seems to indicate that the overlap of cancer related genes between serous and non-serous tumors is roughly similar.

Minor Essential Revisions:

1) Why is the total number of genes different for Figure 2 than Table S2? For example, Figure 2A indicates that there are 476 genes shared by clear cell and serous tumors but Table S2 indicates that there are 477 genes shared by clear cell and serous tumors. Also, I thought that Table S2 was a little confusing at
first. Maybe each category (Amplification, Deletion, and Overall) should only get one table where each cell contains the percentage of genes with the corresponding numbers provided underneath?

2) The authors describe the different frequencies of ERBB2 amplification and deletion at the end of the section “Distinct copy number alterations in EOC histotytes,” but there no associated p-value. Figure 1 contains $-\log(q)$ values, but I specifically would like to see the value for the different ERBB2 frequency between serous and mucinous tumors. Also, the authors cite that 5/19 tumors were mucinous, but I assume this counts the borderline cases. Is this correct? If so, I think the percentages for the confidently identified mucinous samples need to be provided.

3) Although it is OK to present the $-\log(q)$ values for the 4 histological subtypes (in Figure 1B), I think the frequency of alterations for each of the subtypes also needs to be displayed for each of the subtypes (like it is for Figure 1A). This could either replace Figure 1B or it could be a supplemental figure. I think the frequency of the alteration is an important factor for gauging the biological significance and needs to be presented.

4) In the interests of reproducibility, please make sure the methods are as detailed as possible. I think the authors generally do a pretty good job with this, but I think this is worth double-checking. For example, I thought the PCA plots might have been created in Partek, but I couldn’t find any description of what software was used to create those plots.

5) Need to review paper for typos and grammar errors. For example:
   a. Typo in title for Figure 2A
   b. I think “Till date” should be “To date”
   c. The 3rd to last paragraph in the abstract says “In mucinous where KRAS” but I think it should say “In mucinous tumors” or “In mucinous ovarian cancer”.
   d. First sentence in “Distinct copy number alterations in EOC subtypes: “showed” should be “shows”
   e. Last sentence in “Results” introduction: “expressions” should be “expression”

Discretionary Revisions:

1) The section “Copy number alterations in known cancer genes” describes correlations between microarray and qPCR data, but the microarray data has not yet been introduced. I think it would be better to describe these results after the “Identification of driver genes in EOC histotypes” section (which would have to be reorganized a bit).

2) The paper focuses significantly on the different frequency of ERBB2 duplications and alterations in serous and mucinous tumors. The discussion reviews previous studies that examined this gene, but I don’t believe the authors propose any reason why this gene would show different frequencies of alterations in the two histological subtypes. I personally would like to know if
there are any hypotheses about the underlying biological cause.

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

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 interests