Author's response to reviews

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

Authors:

Ruby YunJu Huang (ruby_yj_huang@nuhs.edu.sg)
Geng Bo Chen (gengbo.chen@duke-nus.edu.sg)
Noriomi Matsumura (noriomi@kuhp.kyoto-u.ac.jp)
Hung-Cheng Lai (hclai30656@gmail.com)
Seiichi Mori (seiichi.mori@nus.edu.sg)
Meng Kang Wong (meng.wong@nus.edu.sg)
Ikuo Konishi (konishi@kuhp.kyoto-u.ac.jp)
Jean-Paul Thiery (jpthiery@imcb.a-star.edu.sg)
Liang Goh (liang.goh@duke-nus.edu.sg)

Version: 2 Date: 30 August 2012

Author's response to reviews: see over
Dear Dr Morin,

Thank you for the review of the manuscript. We are grateful for the insightful and constructive comments from the 3 reviewers. There are 3 main concerns highlighted by the reviewers: (i) batch effects from combining the 3 SNP6 datasets, (ii) the disproportional sample size in the histotypes of merged datasets and its effects on the analyses, and (iii) datasets selected for the study. I like to respond to these main points here in this summary followed by a point to point response to the reviewers.

**Batch Effects**

In assessing batch effects, we have employed principal component analysis to assess the effects. Although it is a qualitative approach, the PCA suggest batch effects are minimal. The 1st reviewer has suggested using ANOVA to assess the batch effects. One consideration for adjusting batch effects is comparable distribution from each cohort. The 3 cohorts have different distribution of histotypes, e.g. dataset 2 do not have endometrioid tumors, which is a consideration for batch effects adjustment. Although the PCA showed minimal effects, we have considered the possibility of bias and thus developed the 2-pronged workflow where a filtering criterion for concordance in at least 2 datasets was included. We do agree with the reviewer that a quantitative assessment of the batch effects is necessary. In this instance, housekeeping genes which are expected to be consistent across experiments can be used to assess batch effects. We performed ANOVA test on ~200 of these genes and observed no significance (FDR<0.05), thus supporting the PCA plot that batch effects was minimal.

**Effects of sample size**

Although merging the 3 SNP6 datasets resulted in larger sample size for each histotype, the disproportionate sample size amongst the histotypes still remains. The concern then is the effects of the disparate sample size on the analyses, especially when comparing between histotypes. To evaluate this and as suggested by the 2nd reviewer, we performed additional analyses by sub-sampling (with replacement) serous tumors from the merged dataset of 101 samples into multiple datasets of n=20 or 30. GISTIC was ran for each dataset and significant genes identified compared. In datasets of sample size 20 or 30, ~97% of significant genes found were also identified in the n=101 sample size dataset. This shows that despite smaller sample size, majority of the CNA genes identified within histotypes were robust candidates. However, the analyses also revealed that 57% and 43% of genes in n=101 dataset were not identified in the smaller datasets of n=20 and n=30 respectively. Hence, the reviewers were right that comparison of CNA genes between histotypes is premature. We have revised the manuscript accordingly to compare CNA genes within the histotypes and for between histotypes, we focused on genes identified in the non-serous tumors, where the sample size is more comparable. We also see fit to remove Table S1 and focused our report on known cancer genes. The sub-sampling analyses is included in Methods.

I would also like to highlight the following. In Table S2 (revised to Table S1 in revised manuscript), the total number of genes found in clear cell, endometrioid, and mucinous tumors were 566, 384, and 451 respectively for sample size of n=20-30. If sample size of these histotypes was n=101 (i.e. comparable with serous), we can speculate that we would identify approximately double the number, i.e. ~1000 genes individually in clear cell, endometrioid, and mucinous tumors. Compared with the >6000 genes found in serous samples, it suggests that serous tumors harbor more CNA genes.

**Datasets selected for the study**

The main criterion for selection of data for the study is homogeneity in microarray platforms (i.e. contain most of the histotypes) for both SNP and gene expression. Reviewers 1 and 3 have suggested some
datasets that could be included in the study: i.e. Anglesio et al. (2011), Hendrix et al. (2006), and GSE2109, which either do not contain most of the histotypes, or do not have corresponding SNP and expression data on similar platforms as our study, and hence was not included.

Attached in subsequent pages are the point by point responses to the comments raised by the reviewers. We thank the reviewers again for their valuable time and comments which have improved the manuscript greatly. We hope we have addressed their concerns in this revised manuscript.

Yours sincerely,

Liang Goh, PhD, Assistant Professor,
Cancer and Stem Cell Biology Program
Duke-NUS Graduate Medical School
Reviewer 1: Charles Warden
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.

We thank the reviewer for the encouraging comment and subsequent constructive inputs to improve the paper.

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.

We thank the reviewer for the suggestion. Pls see reply on page 1 for ‘batch effects’.

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).

Pls see reply on page 1 for ‘effects of sample size’. The sub-sampling analyses showed that ≥97% of genes found in smaller sample size datasets were also identified in the larger sample size dataset. We were conscious of the limited sample size and were prepared to identify a small number but high confidence CNA genes for non-serous tumors by setting stringent filtering criteria for calling significant genes. With regards to comparison with other studies, we have compared with 3 studies, TCGA (Cancer Genome Atlas Research Network 2011), as well as significant regions identified in mucinous tumors. In the comparison with 3 other studies (Gorringe, Jacobs et al. 2007; Gorringe, Ramakrishna et al. 2009; Haverty, Hon et al. 2009), 39% of genes reported in any of these studies were also found in our study. We have also included comparison with TCGA in revised manuscript. For TCGA (n=489), 27/42 (64.3%) amplified and 20/28 (71.4%) deleted regions were also found in our study (n=101). Although mucinous tumors has the smallest sample size of 19, the findings of ERBB2 amplification have been supported by several studies looking at mucinous EOC (McAlpine, Wiegand et al. 2009; Han, Hsu et al. 2010; Yan, Choo et al. 2011). Similarly, deletion in 9p21 of mucinous tumors was also reported (Campbell, Gareth Beynon et al. 1995; Devlin, PA Elder et al. 1996).

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?

In analyses of gene expression, we like to highlight that the datasets were not merged but were analyzed individually. Results of differential analyses between mucinous and serous EOC gene expression for each dataset is shown in the table below. There is indication that ERBB2 is differentially expressed between the 2 histotypes, with higher expression in mucinous EOC. Note that 3 mucinous samples in Dataset2 and 2 mucinous samples in Dataset3 do not have corresponding gene expression data.
Dataset | Mucinous (n) | Serous (n) | tstats | p value
--- | --- | --- | --- | ---
Dataset 1 | 7 | 31 | 1.61 | 0.116
Dataset 2 | 3 | 31 | 1.74 | 0.091
Dataset 3 | 6 | 37 | 2.66 | 0.011

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.

We thank the reviewer for the suggestion. As highlighted in the manuscript, we have compared ERBB2 amplification with several studies which looked at mucinous EOC specifically (McAlpine, Wiegand et al. 2009; Han, Hsu et al. 2010; Yan, Choo et al. 2011), all reporting ERBB2 amplification. In studies that looked at CNA of ERBB2 in EOC in general, the results were inconclusive (Gorringe, Jacobs et al. 2007; Nakayama, Nakayama et al. 2007; Pastor, Popovic et al. 2009). Hence, we were encouraged that our histotype-specific approach identified ERBB2 amplification in mucinous EOC despite sample size of 19. The 2 datasets suggested by the reviewer do not contain SNP data and were thus not included in the study. The objective of our study was to identify candidate CNA driver genes, thus without copy number data, it would not be possible for us to validate expression of ERBB2 was driven by copy number alterations in the 2 datasets.

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.

Datasets 1 & 2 were collected and profiled by us. After profiling, we only had samples left in Dataset1 which were subsequently used for qPCR validation. All the tumors identified with copy number alterations in Dataset1 were validated. Dataset3 is a public dataset.

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.

Pls see reply on page 1 for ‘effects of sample size’. We agree with the reviewer (and 2nd reviewer) that the comparison is premature. We have revised the section and manuscript accordingly.

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.

We agree with reviewers that the analysis on overlapped genes is premature and have revised this section accordingly. A frequency plot for the 4 histotypes is now included in Figure 1. The previous Figure 1b is now Figure 2 and the Venn diagram is Figure 3. Although using frequency threshold is one way to identify CNA genes, the setting of the threshold is an arbitrary approach. In addition, genes filtered via frequency can show both amplification and deletion. Our motivation was to identify significantly amplified or deleted genes. As GISTIC takes into consideration the frequency as well as LRR signals to compute the significance, we believe it is a more rigorous way to identify copy number alterations. Although analysis on genomic features would be interesting, as in the comparison of CNA genes between histotypes, we believe it is likewise premature. However, we agree with reviewer that pathways analyses on the potential CNA genes in each histotype would be of interest. We performed pathway analyses on the candidate genes in Table 1 which showed cell cycle, cellular development, growth and proliferation to be important molecular functions. We have decided to retain the Venn diagram in Figure 3 but limit comparison on CNA genes found in the non-serous tumors as their sample size is more comparable.

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?

In the Venn diagram, 476 genes overlapped between clear cell and serous, and 1 more gene overlapped between clear cell, mucinous, and serous, hence total 477 in Table S2. As suggested by reviewer, the tables have been merged. Table S2 is now S1.

2) The authors describe the different frequencies of ERBB2 amplification and deletion at the end of the section “Distinct copy number alterations in EOC histotypes,” 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.

The identification of ERBB2 comes from the 2-pronged analysis workflow which consists of 2 q-value filtering criteria: for merged dataset - q<0.05, and individual dataset - q<0.25. For ERBB2, the –log(q) for amplification in mucinous tumors is 10.9 and for deletion in serous tumors is 3.4 (merged dataset). We apologize for the previous Figure 1b plot, that the y-axis for mucinous tumors was truncated at 4. This has now been rectified. Of the 5 mucinous samples that harbored ERRB2 amplification, 1 is mucinous borderline. We have revised the frequency for mucinous samples (excluding the borderline cases) with ERBB2 amplification accordingly.

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.

We have generated the frequency plot and included it as Figure 1b. The original Figure 1b is now Figure 2. Determination of frequency requires setting a criterion for LRR which means a gene can show both amplification and deletion. As the motivation was to identify significantly amplified or deleted genes, we
believe the significance plot in Figure 2 is more informative in this respect. Comparing the frequency and the significance plots, the significance plot highlights interesting CNA regions more clearly, e.g. the ERBB2 peak is obvious.

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.

We have revised the methods section into 3 paragraphs to better describe the approach. The PCA plot was generated in Partek and is now stated in methods.

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”

The errors have been rectified.

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).

Thank you for this suggestion. The qPCR section has been shifted as suggested.

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.

At this point, I can only speculate that it could be due to the origin of histotypes in EOC, which is an ongoing study. Thus, ERBB2 alterations could be a reflection of the different origins; just as KRAS mutation differs across the histotypes – it is common in mucinous and low grade serous tumors but not in the other histotypes.
Reviewer 2: Terry Furey

The following are questions and suggestions for this manuscript:

Major Compulsory Revisions

1. Obviously, based on the frequency of each histotype, different numbers of samples of each are available. While the authors briefly discuss the implications of sample numbers on their analysis, I think this should be explored more fully. The serous histotype is the main outlier, though Clear Cell has 50% more samples than the other two as well. In figure 1, you see part of the effect of this by the y-axis range in b), so more samples clearly allow for lower q-value scores. What would happen if you sub-sampled 20 of the serous and clear cell samples and re-calculated significance? This could be done several times to get an average overall effect and help better understand how sample size affects the analysis. This could be presented in the results section.

We thank the reviewer for the suggestion. Pls see reply on page 1 for ‘effects of sample size’.

2. Related to the above, without understanding better the effect of the sample sizes on your ability to call copy number changes, then the discussion about how many “unique” alterations are present in each histotype is premature. It is not clear whether the enrichment in serous samples is simply due to greater power to detect these with more samples. The above might suggest what alterations our found in each at similar sample levels and whether one histotype appears to be more frequently altered.

We agree that CNA comparison between serous tumors and the other histotypes is premature and have revised the manuscript accordingly.

3. The pathway analysis should probably go in the results section as this is a separate analysis, or possibly combined within the discussion of the driver genes. How this pathway analysis was done needs to be described, possibly in the Methods. In the paragraph prior to this discussion, it says “This implies several canonical cancer pathways are involved in the pathogenesis of serous histotype.” What is the basis for this statement? Results of the pathway analysis?

Pathway analysis was performed using Ingenuity Pathway Analysis software and is now stated in the Methods section. The pathway analysis statements have been revised and shifted to the section ‘Copy number alterations in known cancer genes’ where Table 1 was first mentioned. The canonical cancer pathways refer to pathways involving known cancer genes such as PIK3CA and BRAF.

4. I think it needs to be clarified what you mean by a “driver” gene. In general, a driver gene is one for which there is evidence that this gene plays a role in the onset or progression of the disease. Evidence of driver genes is often based on expression levels correlated with a clinical or molecular phenotype, such as survival or progression. Since these data are not mentioned, then care must be taken to indicate in all discussions that these are “candidate” driver genes, and it would be good to clarify your definition of what a “candidate” driver gene is in your study.

We agree with the reviewer and have revised the manuscript to indicate that the driver genes are ‘potential’ or ‘candidates’. The definition for ‘candidate’ driver gene is also included in the section Background.

5. Along the same lines, the description of the driver gene analysis is confusing. I would assume that the first part results in a copy number status for a gene, though this isn’t completely clear. The ANOVA then was applied to find genes with histotype-specific alterations? What is being correlated for the Spearman test? I assume this was done for each histotype separately. In general, I think it would be best to split the Data Analysis section of the Methods into two parts – one focused on detecting copy number alterations and one on candidate driver gene analysis.
The reviewer is right; the ANOVA test was used to identify histotype-specific copy number alterations. The Spearman correlation was used to identify candidate driver genes, as shown in the scatter plots (Figure 4A) between copy number alterations and gene expression for each dataset. Analyses on gene expression data were assessed for individual dataset so the correlation was performed across all histotypes. Given the small sample size for the lower prevalent histotypes in individual dataset, it would not be possible or meaningful to assess gene expression for each histotype separately so the correlation was performed for all histotypes. The underlying assumption is that genes that are potential copy number driver across the histotypes are also potentially copy number driver within the histotype. We acknowledge that this is not the most ideal approach but given the limited sample size of the histotypes within each dataset, we have tried to adopt reasonable approaches that would optimize discovery. The data analyses section in Methods section is now presented in 3 paragraphs for (i) histotype-specific copy number alterations, (ii) potential driver genes, and (iii) pathway analyses and generation of plots.

Minor essential revisions
1. On page 8, in the Consistency with other studies paragraph, it needs to be clarified how genes were considered to be “found” in your study for this comparison. Are these all genes in any significantly altered region in any histotype? Were overlaps determined by gene name or genomic region?

   The genes considered “found” were significant genes or regions reported in the 3 EOC. For regions reported, we mapped it to genes within the region using hg18 Refseq genes. We have revised accordingly in the manuscript.

2. It would be good to note in the legend of Figure 1b that the y-axis is different in the histotype plots

   Figure 1b is now Figure 2. We have added the following “Note scale difference in y-axis for the histotypes.” in the legend.

3. Overall, the writing in the manuscript is good, but it would be good to have a native English speaker review this for grammatical corrections.

   The manuscript has been edited for English.
Reviewer 3: Åslaug Helland

Reviewer's report:

• Major Compulsory Revisions.

The choice of data to include in this study:
The study has tried to find aberrations characteristic of the more rare histologies, and by analysing all three selected cohorts, the number of for instance clear cell ovarian cancer amount to 29 samples. There are other studies with substantially more samples which also could have been used. For instance, I find it strange that they have not used the data published by Anglesio M et al, 2011, which have data on 59 clear cell tumours for instance. The TCGA data are also available for analyses.

For the study, the datasets used were on similar platforms, specifically Affymetrix SNP6 and GeneChip Human Gene 1.0ST arrays, to maintain platform homogeneity. Anglesio M et al profiled the gene expression on Affymetrix U133P2 which has different probe designs from the GeneChip 1.0ST array used in all our 3 datasets. TCGA focus on high grade serous data which in our study, there was already 101 samples.

The published cohorts analysed are two by Gorringe et al. Are some of the same samples analysed twice?

Gorringe et al published two papers on copy number using Affymetrix 500K SNP array and SNP6 array separately. For our study, we used the SNP6 dataset with no repeated samples.

As the abstract describes in silico analyses of datasets, the method chapter describes two datasets which seem to have been analysed in the lab and used as different datasets. Are this in the public domain, or are the lab-analyses performed for this particular study?

Three datasets were used in the study; Dataset 1 and 2 were collected and profiled by us while Dataset 3 is a publicly available data from Gorringe et al. The 2 datasets will be released with the publication. To improve clarity, we have shifted description of the 3 datasets to the beginning of Methods section.

The other aim of the study is to identify drivers. They report several possible drivers, by investigating the concordance between copy-number alterations and expression of the genes in specific altered regions of the genome. Many would say that this is a weak indication of driver-characteristics of an alteration. Some of the identified genes are known oncogenes, while others (like TP53) are tumor-suppressor genes. Other, more functional studies could be performed to validate the more unknown genes as actual drivers.

Association between copy number alterations and gene expression has been used by several studies to identify potential driver genes (Lando, Holden et al. 2009; Woo, Park et al. 2009). We agree that at best, these are candidate drivers and need to be further validated. We have revised the manuscript to indicate that the driver genes are potential or candidate drivers and need further validation. It is not within the scope of this manuscript to validate the driver genes though we have tried to shown for ERBB2 using qPCR; it appears to be a driver gene in EOC.
Reference


