Reviewer's report

Title: SNP Microarray analyses reveal copy number alterations and progressive genome reorganization during tumor development in mice breast cancer

Version: 2  Date: 29 March 2012

Reviewer: Ramon Diaz-Uriarte

Reviewer's report:

Major compulsory revisions:

1. I have a hard time understanding how the comparisons have been made. Please, provide them. These are some of my problems:

1.1 Sample size, and statistics of SNPs, etc:

So you have a total of eight mice, 2 for each group: 2 normal, two WAP-SVT/t on first day of lactation, 2 WAP-SVT/t that had developed cancer, and two from 762TuD? (this is material and methods). Now, do you use any statistical method to compare CNAs? You say "all values in the range of -0.2 <= x <= 0.2 were considered as unchanged", why do you use those values? Then, you say "copy number values were ranged into five groups": what kind of method is this? I mean, we have t-test, ANOVAs, etc. Why would you do this?

Next, you say "no significant changes in CNA were found...". Significant? How do you assess significance and of what? It looks like you have categorized the data, and then applied some test for differences in frequencies? If you have done so, please explain. And if you have done so, explain why you have done this because categorizing is something that should not be done, and if strictly necessary only done very carefully.

After that, you talk about an increase in copy number: how can you say that? How do you make the comparison that leads to that conclusion? Comparing the number of SNPs in different categories? Again, explain the method you use and justify it.
1.2 Analyses of segmented data.
You say "the number of segments in both samples also differ significantly": how do you assess this?

My problems with the analyses continue on p. 4. The section with heading "Percentage of segment CN" says "the percentage of changed segCN is significantly higher ...": how do you assess this?

At the end of that paragraph, you say "it is obvious that tumor samples are heterogeneous". How do you reach that conclusion? How do you define heterogeneity? How do you test or make inferences about it?

1.3 Fragmentation patterns:
The authors say things such as "It is not only the case, that the number of segments is higher, or that new segments can be detected from the normal to the transgenic and the tumor samples, but also, the segments detected in tumor samples are mostly fragments from segments found in the transgenic sample. These segmentation patterns indicate predisposed chromosomal breakpoints (...)

These would all be very neat, but it seems basically a description of what they get from looking at figures. I can find no systematic procedure nor a statistical approach to arrive at those conclusions.

1.4 Segmentation and gene expression:
"We compared the impact of the top 500 differentially expressed genes (based on the false discovery rate adjusted p-value, FDR p-value) in the two normal and the two tumor samples with the copy number of related SNPs."

What does that mean? What is "comparing the impact"? How, exactly, did you compare the groups? Are these tests of differential expression? It seems so, since in p.8 you refer to using limma. But this is unclear.

"Furthermore, we observed that the biggest amount of SNPs were located in non-coding chromosomal regions. " So? How is this relevant for the previous paragraph? For what?
2. Organization and sequence of analyses and results: sometimes the paper is unnecessarily hard to follow.

2.1 SNPs and segmentation: the left side of page 2 talks about copy numbers from SNP data, without segmentation. Then, on the right, you segment the data. This seems strange: why didn't you just segment the data? What do exactly you gain from the first analyses? Some of the answers are provided later on, but a guide and rationale for your approach would have helped.

2.2 Fragmentation patterns:
When I get here, I realize that the authors are approaching a similar set of questions from slightly different perspectives and going step-by-step. This is fine, and possibly very interesting, but a road map of what and why is done would have been good.

3. Sample size: In the conclusions, the authors themselves recognize: "The limitation of this work was the small number of samples for the comparison of copy numbers and gene expression, making it difficult to determine (...)

Yes, that is certainly a very serious problem of this paper. Maybe there is little (nothing?) that can be said and done, unless the sample size is increased appropriately.

Minor essential revisions:

- Page 2, reference 2: include, please, more recent references too. Maybe now people think otherwise.
- Please, use "cf." properly. "cf." means "contrast with", and denotes a contraposition. For instance, in page 2, when you say "(cf. figure 1)" you really mean "(see figure 1)".

**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