Reviewer's report

Title: Integrated mutation, copy number and expression profiling in resectable non-small cell lung cancer

Version: 1 Date: 30 September 2010

Reviewer: Ruprecht Kuner

Reviewer's report:

The authors present a large collection of results from genomic and transcriptomic data in a limited NSCLC collective. In principle it is very useful to collect and compare data from different molecular levels. A major weakness of the manuscript is the presentation of results without weighting their significance and relevance in this dataset and for the cancer entity. Furthermore it is hard to distinguish novel from known findings. Many comparisons are strongly limited by the number of samples in one group and by the high risk for confounding effects from other parameters. These sections should be reduced, even the outcome carefully stated. It is impossible to rule out the focus of this manuscript. I think the authors should avoid to sum up all results, but explain the most reliable results more in detail.

1. The supplemental data for the microarray and CGH platform and data is not standardized. Please submit your data to GEO or ArrayExpress according to the MIAME rules.

2. The authors should give more details to the t-statistic. Did you use a modified t-Test like SAM or LIMMA including multiple testing correction? What are the limitation of your methods to screen for mutations (false positive, false negative rate)?

3. Figure 1 is needless, Table 1 is not stated in the method section

4. Table 2 and result section Aneuploidy: How are the breakpoints defined?

5. CNV is difficult to interpretate from Figure 1. There should be a list of all significant aberration including the frequencies in AC and SCC.

6. Result section:

Correlation in the headers is the wrong vocabular. Regarding the limited size of the investigated collective statements about associations between molecular findings and clinical parameters should be significant or better omitted. Very often detailed information are missing (e.g. group size, distribution of other parameters)

7. The authors should clearly assign all group comparison.
7.1 In general comparison of very small sample size below n=5 should be avoided by reason of the potential bias from other parameters (e.g. Subtype).

7.2 For example, TP53 signature is affected by the NSCLC subtype. This was stated from the authors themselves. Such analyses or interpretations should be avoided anyway.

7.3 The number of sample size should be clearly stated, e.g. KRAS signature in AC and LCC, may this be biased by the subtype? Similar for metastatic profile (n<5?), recurrence and survival. How are these profiles affected by other parameters?

8. External dataset: In principle two different gene lists have been compared, the overlap was applied to each dataset. The overlap of 40 from 60 genes looks promising, how long was the list for GSE11117? How did you define your subgroups according to the signature (classification, clustering)? The authors should present informations and illustrations for that.

9. Integration of datasets:
I did not really see integration rather than comparing gene lists or loci. However genomic and transcriptomic data seem to be strongly limited with respect to the used platforms. Single genes have been further prioritized by a potential second event (e.g. CNV). Associations with mutation and clinical parameter are too weak in this approach.

10. Validation experiments for findings/signatures using other methods are missing.

**Level of interest:** An article of limited interest

**Quality of written English:** Not suitable for publication unless extensively edited

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

**Declaration of competing interests:**
'I declare that I have no competing interests'