Reviewer's report

Title: Gene expression profiles of lung adenocarcinoma correlate with histopathological grading and survival but not with EGF-R status: a microarray study

Version: 1 Date: 25 June 2009

Reviewer: Paul Boutros

Reviewer's report:

Major Compulsory Revisions

1. Annotated microarray data needs to be deposited in a public database
   As the authors note in their conclusion, it is required and highly beneficial to the research community to deposite raw data for all microarray experiments. Therefore, the raw (.CEL files) and normalized data needs to be deposited in a public database such as GEO (NCBI) or ArrayExpress (EBI). The authors should ensure that in addition to the already-given clinical parameters, percent tumour, IHC scores (membranous and cytoplasmic separately), and EGFR FISH signals are specified for each patient along with the microarray data.

2. EGFR mRNA/protein concordance
   The authors indicate that the EGFR mRNA (from microarray) and protein (from IHC) levels are well-correlated. They provide a graphic display (Figure 2), but do not appear to support this with a statistical analysis. A correlation (preferably Spearman's) between the IHC score and ProbeSet intensities should be calculated and reported, along with a measure of its statistical significance (i.e. p-value).

3. Validation of the Potti signature
   The validation of the Potti signature needs to be described in more detail. The description of that signature in the original NEJM paper is somewhat unclear, and therefore it is unclear which 114 genes the authors chose to include. Second, the Potti signature was validated via decision-tree analysis. It is remarkable that a simple unsupervised analysis recapitulated this signature, and the authors should comment on the possible reasons for that they were able to classify lung-cancer samples using a much simpler analytical technique than by most previous studies, include Potti et al.. Third, the Potti analysis was based on HG-U133 Plus 2.0 arrays, whereas the authors worked with HG-U133A arrays, which are a subset. The authors indicate that all *genes* were found on both platforms, but this does not indicate which ProbeSets were used. The authors should report what ProbeSet mappings they used between the two platforms.

4. Statistical analysis of survival data
   The authors should include a hazard-ratio and adjustment for histopathological
stage and grade

5. Statistical power of the study
   The discussion should contain a more explicit description of the potential challenges caused by the relatively small number of patient samples in this study.

Minor Essential Revisions
1. In some cancer the authors use "Lung cancer" in others "Lung Cancer" -- they should be consistent
2. page 4 "conferring to poor prognosis" should be "conferring poor prognosis"
3. the concordance in histopathological and grading between the three pathologists should be reported
4. page 6 "using the Qiashredder (Qiagen,)") should include the location of Qiagen (presumably Hilden)
5. The authors indicate that HG-U133A chip contains 14,564 human genes, but this number is dependent on the definition of gene (Entrez Gene, ENSEMBL, other?) and the version of the Affymetrix annotation (na22, na23, other?) used. These details need to be specified.
6. The version of the affy package used for analysis should be specified
7. page 9 "Grad 1 and 2" should be "Grade 1 and 2"
8. what statistical test was used to determine that the clusters differ with respect to histopathological grading?
9. The colours chosen to distinguish short- and long-term survivors in Figure 4 are hard to distinguish

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