Reviewer's report

Title: A new molecular breast cancer subclass defined from a large scale real-time quantitative RT-PCR study

Version: 1 Date: 26 November 2006

Reviewer: Luc Dirix

Reviewer's report:

General
This study on 199 primary breast tumors claims to have identified a number of new subgroups based on the quantitative expression of 47 selected candidate genes, potentially able to define subgroups in human breast cancer.

The heterogeneity of the sample of patients is a potentially enormous confounder in this study; all but one have been treated with a combination of (heterogeneous) adjuvant regimens; follow-up is short (median 5 years) for a population that is in majority node negative. This should be taken into account when suggesting that subgroup 8 is particular innovative good prognosis group. After a median follow up of 5 years (the overall relapse rate including local relapses accounts for only some 15% (34/199, underscores the need for longer follow-up. Subgroup is mainly ER+ (77%), LNN - (64%), T1 (45%), grade 1 (54%) and no grade III tumours.

-------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

My main criticism is the basis of the division of the clustering dendrogram for tumour samples in these numerous different subtypes. Why have subgroup 3 and 4 been separated, there is no evidence in the dendrogram for this separation. Also for the division between subgroups 5 and 6, 6 and 7, 7 and 8, 8 and 9, 9 and 10, 10 and 11, 11 and 12 no or poor evidence is provided in the dendrogram. Anyway, a robustness of clustering analysis should be performed here.

The strategy of using the entire data set to identify subgroups and then using the entire data set as an internal validation set to test the robustness of the classification is not without risk. A superior strategy would be to divide the data set (n=205) into a training set (n=n/2) and a test set (n=n/2). The unsupervised hierarchical clustering can be used to determine relevant clusters. Then, gene expression signatures can be associated with each of the clusters which are then tested on the validation set. The robustness of the classification can then be tested on an external test set.

By turning continuous variables in discrete variables, a substantial amount of information is lost. An alternative would be to select genes based on ROC curves to take into account sensitivity and specificity. Selected genes should be validated against a bootstrapped data set. Next, selected genes for each subgroup can be used to calculate centroids and these centroids can then be used to classify the remaining samples by calculating Pearson or Spearman correlation coefficients.

The set of genes in this study has been randomly chosen. It contains a lot of ER target genes as well as genes present in the intrinsic gene set of Perou. Therefore, it remains surprising that the luminal tumours are scattered throughout the dendrogram. The expression of ER regulated genes is one of the major determinants of the clustering pattern in gene expression studies using breast tumour data sets. The selection of the intrinsic gene set, based on a greater variability between samples of different patients then between samples of the same patient is biologically more relevant then randomly selecting genes to classify breast tumours. Therefore, the current classification should be compared to the classification using only the 15 intrinsic genes.

-------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

Is there a strong association between the different cell-of-origin subtypes and the tumours from the Sorlie
and Van 't Veer data set clustering in subgroup 8. Was a strong or a weak correlation coefficient observed?

Might it be possible that group 3 and 4 are the luminal A breast tumours and group 1 and 2 are the luminal B breast tumours?

Who are the 14 samples from Dr.Katsaros? All patients were treated in Motpellier or not?

Discretionary Revisions (which the author can choose to ignore)

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article of importance in its field

Quality of written English: Needs some language corrections before being published

Statistical review: Yes

Declaration of competing interests:

I declare that I have no competing interests