Reviewer's report

Title: Integration of Transcript Expression, Copy Number and LOH Analysis of Infiltrating Ductal Carcinoma of the Breast.

Version: 2 Date: 23 December 2009

Reviewer: Bauke Ylstra

Reviewer's report:

The authors present a paper which uses an integrative approach to pinpoint driver genes in a series of Ductal invasive breast carcinoma's. Therefore the authors use SNP arrays (Affymetrix) as well as expression arrays, both available for 12 samples.

Such analysis has been performed previously for 6 larger series (each > 40) [1-6]. The difference is that the authors also characterize copy neutral LOH.

Major Compulsory Revisions

The quality of writing is limited, statistics are poor and experimentally flawed. Most importantly and for different reasons stated below, I do not trust the calling of copy number aberrations (CNAs)

Some remarks to underpin my statements:

The abstract says 22 samples but in reality for only 12 samples RNA and DNA arrays were performed, this is misleading use of language.

12 (nor 22) samples is extremely low number and does not give sufficient statistical leverage for any of the analysis. As a result the copy number summary plot (Figure 3) appears spiky. (possibly also due to the low sample numbers)

Figure 3 is NOT representative of other studies; what happened to the frequent losses on 16, chromosome 2 is mostly stable in IDC, whereas these authors think they find many aberrations.

As a reference for normalization a large (>12) pool of normals performed in the same lab and equipment, preferably form the same sample types is recommended to be performed. This was NOT done for this study, rather a reference in the Partek software was used. This may further explain miscalling and non-representative results.

In more recent articles, which the authors in fact refer to, so must have known better, normality is performed in an even more sophisticated manner. A matched normal from the same samples is used, for breast cancer[6] and for Affymetrix arrays[7]. This may further explain spikiness in the plot Figure 3 since not only CNAs are called but also germ-line copy number variable regions. This is misleading

I do not trust the data pre-processing a. The authors work with cut-offs, whereas
sophisticated calling algorithms exist. The cut-offs are not None of the cut-offs are justified, why use 8 out of 12 to call it significant? Why not 7 or 9?

Sophisticated and published software exist, written in R and publicly available for integrative genomics (expression and CGH) as well as for calling of gains and losses.

Although LOH may seem an additional perk towards publication, the authors make no distinction between, gain and amplification. This is a serious and misleading flaw

Minor Essential Revisions

Much of the writing is redundant. ie in introduction integrative genomics is explained with too much text, rather than giving appropriate references.

Much writing is dedicated to all the different genes, this is mere literature review, which doe not belong in an article, particularly since no (biological) validation is done of any of the genes.

High resolution CGH; this is relative, much higher (10x) arrays are available.

700bp resolution; I do not know what that remark is based on, more info here [8]

Figure 3 and 4 can be combined, while adding gains in addition to the amplifications

I am sorry that I cannot be more positive at this point.

Yours,

REFERENCES


**Level of interest:** An article of insufficient interest to warrant publication in a scientific/medical journal

**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