Author's response to reviews

Title: Gene expression meta-analysis identifies metastatic pathways and transcription factors in breast cancer

Authors:

Mads Thomassen (Mads.thomassen@ouh.regionsyddanmark.dk)
Qihua Tan (qihua.tan@ouh.regionsyddanmark.dk)
Torben A Kruse (Torben.kruse@ouh.regionsyddanmark.dk)

Version: 2 Date: 17 August 2008

Author's response to reviews:

Response to reviewers

Reviewers comments are addressed below and all major changes in the manuscript are marked in bold.

Response to reviewer 1

1) We have performed global analysis of pathways involved in breast cancer metastasis. This will of course identify a number of well known mechanisms. In previous studies, these mechanisms have typically been identified one by one with different techniques. This study serves as an independent global validation of these mechanisms with a different technique. Some previous pathways are elaborated on e.g. metabolism where both upregulated glycolysis, downregulated amino acid and downregulated fatty acid is reflected in several pathways. To our knowledge this has not been reported for breast cancer before. Previous studies of metabolic changes associated to metastasis have typically focused on single enzymes e.g. hexokinase in fructose metabolism. Some new metastatic mechanisms are identified, e.g. a number of transcription factors not previously linked to metastasis. However, it would be beyond space limitation to discuss this. These things are now clarified in the discussion.

2) We agree that different subtypes of breast tumors e.g. molecular subtypes may metastasize by different mechanisms. It would indeed be very relevant to perform pathway analysis of metastasis within molecular groups. However, in this paper we focus on the general pathways and it is beyond the scope of the paper to subgroup tumors. The result is that identified pathways and transcription factors are biased towards mechanisms that are common across tumor subgroups. This is actually also a strength of the paper, because sub grouping would introduce more degrees of freedom, especially if different clinical subgroups where also considered. These aspect are now included in the discussion.

3) We agree that validation of the data would be valuable and are planning to this in future studies. Regarding the significance of the findings, there are several
advantages of the statistical method we use. By analyzing gene sets instead of single genes, the complexity of data is reduced and the multiple measurements for each gene set strengthen the gathered results. The meta-analysis identifies gene sets that are concordantly differentially expressed in 8 data sets performed with different clinical groups, outcome and platforms. This means that the test is conservative because only pathways that are strongly associated to metastasis will be significant despite these biases.

GSEA rank gene sets according to the normalized enrichment score (NES) and the significance of each gene set is estimated by the GSEA algorithms. However, this information is not used in the downstream analysis. Only the ranking is used. This actually makes the permutation test very conservative because significance level from GSEA and GenMAPP analyses would if accumulated from 8 data sets increase the gathered significance. It has not been possibly to integrate the statistical information in meta-analysis because of the different distribution of the parameters (NES or Z score) between the data sets.

Response to reviewer 2

1) Statistical method
   Our null-hypothesis is that the expressions of genes in pathway gene sets are unrelated to metastasis. This means that the ranking value for a given gene set in a given data set is expected to be a random value between 1 and the maximum number of gene sets analyzed. To simulate the distribution of mean ranking values across the 8 data sets fulfilling the null-hypothesis, random drawing of 8 ranking values were performed 106 times and the mean value was calculated each time. A null distribution of mean ranking values was generated from these results. To test the significance for a given gene set, the observed mean ranking value was compared to the null distribution. This is now explained in the method section.

2) GSEA parameters
   It is correct that GSEA rank gene sets according to the normalized enrichment score (NES) and the significance of each gene set is estimated by the GSEA algorithms. However, this information is not used in the downstream analysis. Only the ranking is used. This actually makes the permutation test very conservative because significance level from GSEA and GenMAPP analyses would if accumulated from 8 data sets increase the significance (as also explained to reviewer 1). It has not been possibly to integrate the statistical information in meta-analysis because of the different distribution of the parameters (NES or Z score) between the data sets.

3) Inclusion of patients with quite diverse characteristics
   Patients with different outcomes (relapse, metastasis, distant metastasis and death from breast cancer) are included in the study. We completely agree that this may bias the results because relapse includes local recurrences and these may be the result of suboptimal surgery or mechanism of spreading different from distant metastasis. Regional metastasis, i.e. recurrence in lymph nodes may
also be the result of different mechanism compared to distant metastasis.

To estimate the occurrence of local and regional metastasis in a typical data set, we have now looked into a large Danish tumor bank located in Odense, Funen. We use data from a tumor bank instead of data from the entire patient cohort to mimic the typical samples used in the studies. A typical difference between these groups is lack of very small tumors in tumor banks.

From 1989 to 1999, 2263 patients underwent surgery at Funen. 253 samples from patients that experienced recurrence are frozen. Out of these, 43 had local recurrence (17 %), 11 had regional recurrence (4 %), and 199 (79 %) had distant metastasis. The bias from approximately 21 % recurrences that are not distant metastasis in 3 out of 8 data sets is minor. Although it would be ideal to remove local and regional recurrences from the data sets, this has not been possible because this patient information is not available. In Uppsala data set the outcome is death from breast cancer, but this is almost synonymous with distant metastasis.

The meta-analysis identifies gene sets that are concordantly differentially expressed in 8 data sets performed with these different clinical groups, outcomes and platforms. This means that the test is conservative because only pathways that are strongly associated to metastasis will be significant despite these biases.

4) language
Several corrections have been performed. For simplicity these are not marked in bold.