Reviewer's report

Title: Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data

Version: 2 Date: 6 January 2009

Reviewer: Phillip G Febbo

Reviewer's report:

Review of Gorlov et al manuscript entitled “Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data”

Major Compulsory Revisions.

1) The authors refer to statistical ‘true discoveries” on page 6 but perform no validation to demonstrate that their statistical method actually is valid. I.e., there is no quantitative PCR or immunohistochemistry in an independent dataset that confirms the consistent finding and demonstrates that the observed genes that represent “true discoveries” are differentially expressed.

2) After ranking genes based upon the P value, the authors used “top-ranked genes to analyze their clustering, pathway, molecular, and cellular functions”. If the authors are confident that their meta-analysis identifies “true discoveries” then why is it not best to use all genes?

3) In the supplemental materials, they tested multiple cut offs and used their functional annotation to determine the number of genes resulting in the highest number of discovered functions. They then go on to discuss the relevance of the functions and pathways identified. While their attempt to avoid arbitrary cut-offs is to be lauded, by testing the same data to identify the number of genes implicating the largest number of functions they detract from the validity of their findings. Ideally, the number of genes would be established in a training analysis the optimized parameters of which are used in an independent analysis.

4) The decreased expression of the focal adhesion pathway in the transition from normal prostate to local prostate cancer can be explained by the decreased percentage of stroma seen in tumor samples compared to local prostate cancer. This has been well described and as only one of the datasets analyzed was microdissected, the findings can all be attributed to tissue composition rather than biology. In fact, the authors state that the observed changes are not observed when differences in microdissected samples are looked at specifically on page 11. They propose that the observed changes are occurring within the fibroblasts. Without confirmation in microdissected samples or independent analysis that identifies the same changes while accounting for tissue composition differences, the value of these findings is unclear. The same is true for the transition from local to metastatic as there is often less stromal tissue with the metastatic samples that have been analyzed by microarray when compared to
the local cancers. This is my major concern with the paper.

5) The discussion and proposal of a "collagen hypothesis of prostate cancer development" is not adequately supported by the presented data without the authors addressing the above issues.

Minor Essential Revisions

1) The authors do not state the statistic used to generate a P value for their initial assessment of overlap. Because they refer to this as a “liberal P value of 0.05” it is not clear if their statistical measure takes into account multiple testings and is consistent with methods proposed for microarray data by Tibshirani et al and/or if the measure takes into account gene variation as well as fold change. It is important to use measures that account for variation as without such accounting analysis can be significantly impacted by spikes and/or outlying expression of genes in single samples.

2) The authors rank genes based upon a P value that represents effectively a sum of a gene’s Z score across all datasets divided by square root of the number of tests (k). It is not clear from the methods what “tests” were used for k.

3) There seems to be little difference between 500, 700, 1000, and 2000 genes with respect to the number of functions and pathways implicated. Is there significant overlap in the implicated functions and pathways? The impact of number of genes included on analysis on the results has to be described in detail.

4) In the first part of the results – the authors describe more methods under “proof of principle”. The analysis should be presented in the methods and the results here. In addition, they report only a single “null” testing of the lowest 500 genes. The more appropriate and more common approach is to perform multiple tests of randomly chosen genes and to use the distribution of results from usually 1000 randomly aggregated genes to the experimental group of 500 genes. It is insufficient to test to a single group.

5) The analysis of the overlap between NP-nMPC and those between nMPC and MPC assumes that all 17,859 genes are equal with independent expression and it is clear from the corpus of microarray work this is not the case. The authors should only use genes from the 17,859 matched for expression value and expression variation before they can argue that the overlap is greater than expected by chance alone.

Minor:

1) Typo in Supplemental materials page second page “expecte”

Level of interest: An article of limited interest

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.
Declaration of competing interests:

None to declare.