Reviewer’s report

Title: Identification of Transcriptional Regulatory Networks Specific to Pilocytic Astrocytoma

Version: 2 Date: 10 December 2010

Reviewer: Pierre Bushel

Reviewer’s report:

The manuscript by Deshmukh et al. entitled “Identification of Transcriptional Regulatory Networks Specific to Pilocytic Astrocytoma” and submitted as an original research paper to BMC Medical Genomics for review describes the use of a so called prediction strategy and the use of a network analysis on gene expression data from pilocytic astrocytoma and non-malignant brain tissue (frontal and prefrontal) to identify transcription factor-gene and gene-gene interactions that are specific to the pathogenesis of pediatric brain tumors. The following revisions are suggested to improve the manuscript.

Major revisions

1) The authors claim that their analysis strategy identifies “genes to predict a PA-specific gene network based on transcription factor gene regulatory interactions”. This is misleading and a gross overstatement since a prediction was not performed in the sense of the word. To truly claim that the genes predict the PA-specific gene network, the components of the network (genes and interactions) would need to be permuted and then the most likely network would need to be tested for consistency across several independent PA-data sets. The author need to clarify what they actually did and should find a different way to state their claim if a prediction was not actually done.

2) The iterative and independent parallel data set framework (Figure 1) that the authors use to identify gene expression signatures has a caveat to it which can potentially introduce bias\an overfitting so to say in the genes. Since only one split of the data into a set for training and testing is performed, the true distribution of the samples is not incorporated to test the selected genes upon. An outer loop would be needed to resample the data a fair number of times (i.e. 100) in order to split the data into training and testing data sets. Simply performing the inner loop on the training data set 10 times and then looking for the “extent of overlap” with the test data set is ignoring the randomness and variation in the population. At the very least, the authors need to acknowledge this in the manuscript and guard the reader in terms of the limits to the interpretations gathered from the results and the shortcomings of strategy.

3) The methods section is lacking a fair amount of details for one to replicate the results. For instance, what identifiers (GeneBank, RefSeq, UniGene, etc.) were used for mapping the probes on the arrays to the transcription factors (TFs) and genes? Also, what version of TRANSFAC was used for the TF and gene
mapping? How many TFs were actually mapped and where is the mapping of the TFs to the downstream target genes??? What Affymetrix array was actually used and how may probes on it\them? Was multiple testing correction employed??? The authors need to add more details to the methods section for clarification.

4) It is not clear why both the TRANSFAC and the Promoter Analysis Pipeline (PAP) application were needed for the TF-gene mapping. Does PAP leverage TRANSFAC? The TF-gene mapping can be obtained just from TRANSFAC. For instance, in Figure 2., step 1 is from the DEGs and step 4 is from the pruning. But what is steps 2 and 3 from??? Just TRANSFAC and\or PAP?? This clarification would be good in the body of the manuscript as well as in Figure 2. The authors claim that PAP ascertains TF binding motifs in the genes represented by the probes on the arrays but it is not clear how reliable the predictions are, what if any TF position weight matrix is used or why genes not differentially expressed were not included in the network analysis. It is well-known that from gene expression array analysis, TFs are often times expressed at low levels, not differentially expressed but have a large impact on pathway\gene network signaling. Therefore, modules of TFs and genes whether differentially expressed or not are typically used for network analysis so that central hubs in the network, that often times are TFs which are not differentially expressed but highly connected, can be identified. The authors need to give evidence and support for the limited network analysis which appears to be based just on the core\focus DEGs and TFs.

Minor Revisions

1) The subnetworks generated from the analysis are essentially repressed but an indication of that hasn’t been addressed in the manuscript. What interpretation does this have? Could not the authors provide a list of the genes up and down regulated and also provide a notation of how many genes are differentially expressed (up and down)?

**Level of interest:** An article of importance in its field

**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