Sample size considerations for the efficiency of extracting regulatory connections from a combined miRNA and gene expression data set

Smriti Shridhar, David P. Kreil. Chair of Bioinformatics, Boku University Vienna, Austria.

One of the key challenges in the analysis of the wealth of genome scale experiments is its meaningful interpretation. This is complicated by the fact that biological processes are regulated at different levels while also interacting with one another. The integrated study of complementary data promises to provide a handle to addressing this complexity. While the curse of dimensionality is a well known issue in the analysis of genome-scale data, it is compounded by the integration of multiple data tracks. The question therefore arises, what sample sizes are typically required to allow a meaningful multi-track analysis. We here try to address this by studying the performance of a joint analysis of gene and miRNA expression data in a large multi-patient study. In this abstract, we present preliminary results for a popular method for inferring regulatory factors of modules identified by a Gibbs sampling based bi-clustering. We will extend this preliminary study with a comparison to other analysis approaches for the conference.

We have applied the LeMoNe algorithm (Bonnet *et al*., 2010) to construct regulatory modules from a Glioblastoma multiforme subset of 534 miRNA and 11925 gene expression profiles for 232 patients. The dataset was subsampled to sizes reduced by multiples of two for a smallest data set of 7 samples. Each subset was sampled 9 times for an assessment of robustness. After bi-clustering of the joint gene & miRNA expression profiles, the miRNAs were tested as potential regulators.

We report on two measures of biological relevance:
a) the enrichment of miRNAs that have been reported to be Glioblastoma associated ('known') in the set of identified regulator miRNAs (Ruepp *et al*., 2010), and
b) the enrichment of predicted gene targets according to MicroCosm (Kozomara & Griffiths-Jones, 2011) matching the 'regulator miRNA' in its associated modules of regulated genes. Enrichment significance was computed by Fisher's exact test.

Figure 1.

Significance of enrichment of known Glioblastoma associated miRNAs in the set of identified regulator miRNAs.

The x-axis shows the number of patient samples used, the plus symbols mark the significance of enrichment for the nine independently drawn subsamples (Fisher's exact test). Symbols falling below the red line (denoting p=5%) are considered not significant. The right most symbol represents results for the full dataset.

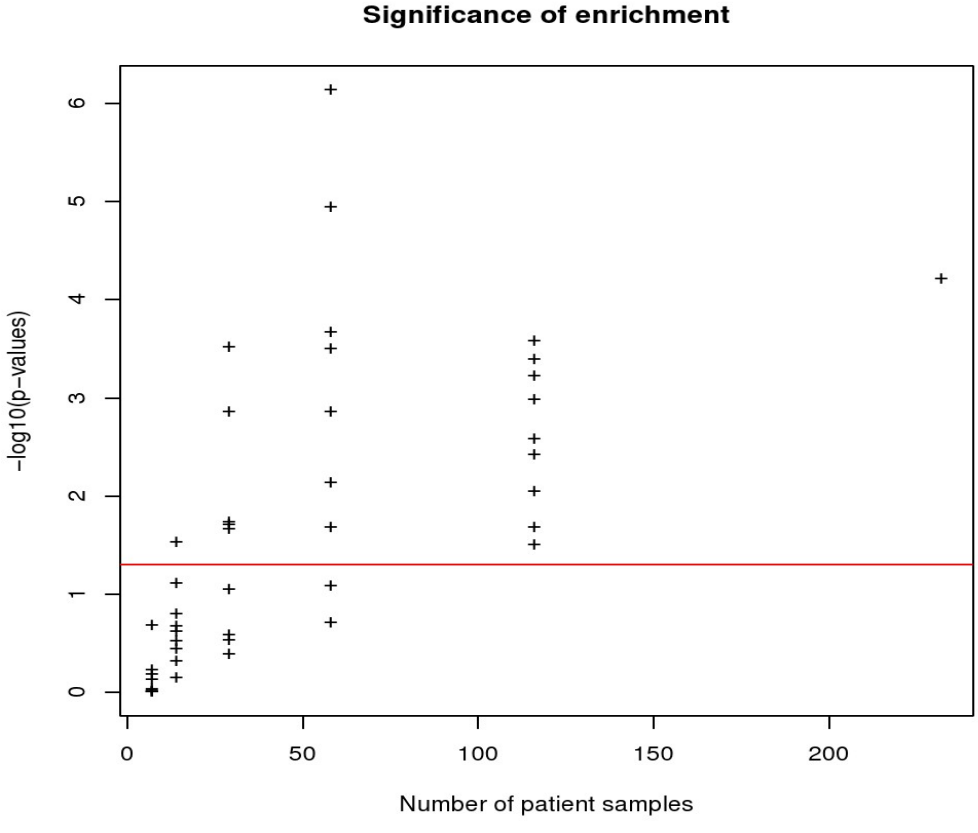
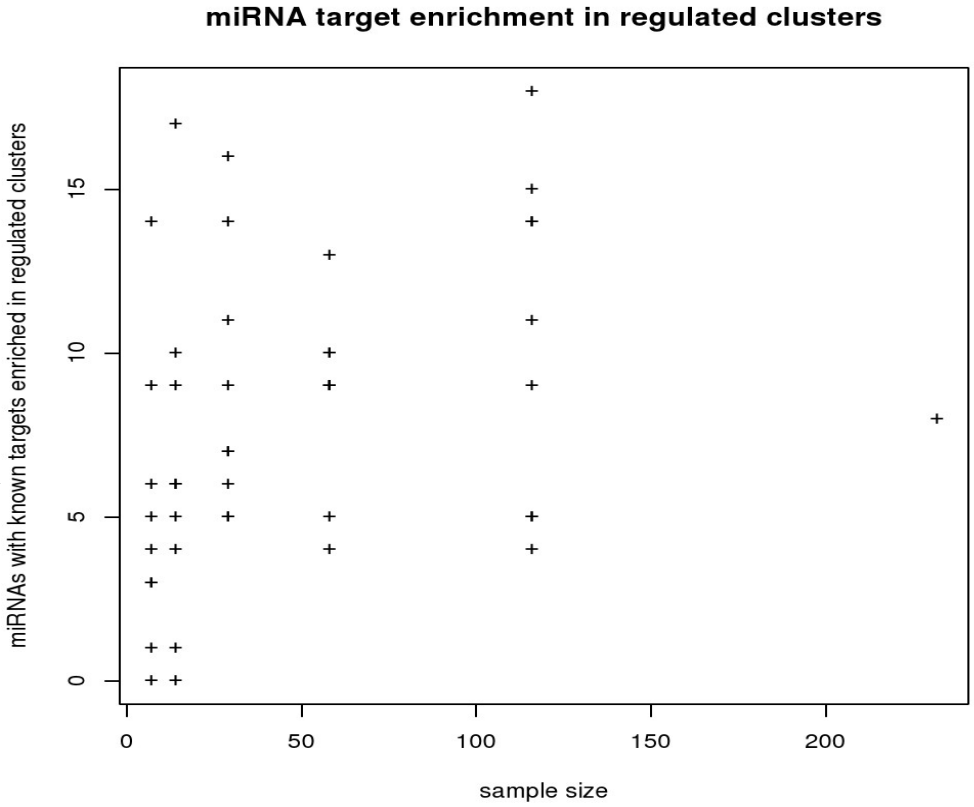


Figure 2.



Number of miRNAs for which a significant enrichment of their predicted targets could be identified in their regulated clusters. An enrichment was considered significant for a *p*-value threshold of 10%.

The *x*-axis shows the number of patient samples used, the plus symbols mark the number of miRNAs with significant target enrichment. Results are shown for the nine independently drawn subsamples, as well as the full dataset (rightmost plus).

A significant enrichment could be observed both of Glioblastoma associated miRNAs in the set of identified regulatory miRNAs and of predicted miRNA targets in the regulated clusters. While good results can already be obtained with moderate sample sizes, the studied approach begins to break down for the smaller sample sizes tested. In particular, for samples of 7 or 14 patients, enrichment of Glioblastoma associated miRNA s was not significant, and no significant target enrichment was found for some subset samples. This suggests that about 30 samples are required to take advantage of this integrated analysis approach for the identification of miRNA/gene regulatory modules.

We will extend this study to compare the robustness of the standard LeMoNe approach with alternative approaches for the conference.

References:

Bonnet, E., Michoel, T., Van de Peer, Y. (2010) Prediction of a regulatory network linked to prostate cancer from gene expression, microRNA and clinical data. *Bioinformatics* **26**, 638-644.
Kozomara A, Griffiths-Jones S.(2011) MiRBase: integrating microRNA annotaion and deep-sequencing data. *Nucleic Acids Research* **39**, D152-D157
Ruepp A, Kowarsch A, Schmidl D, Buggenthin F, Brauner B, Dunger I, Fobo G, Frishman G, Montrone C, Theis FJ.(2010) PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes**.** *Genome Biology* **11**, R6