The importance of normalization on large and heterogeneous microarray datasets

Ed Perkins, Tanwir Habib, Lyle Burgoon, Stephen Edwards, Francesco Falciani, Alex Loguinov, Dan Villeneuve, Chris Vulpe, Natalia Garcia-Reyero

DNA microarray technology is a powerful functional genomics tool increasingly used for investigating global gene expression in environmental studies. Microarrays can also be used in identifying biological networks, as they give insight on the complex gene-to-gene interactions, networks and pathways, thereby enabling the exploration and examination of how chemicals cause toxicity. Gene expression analysis is a multi-step process and there are many sources contribute to systematic variations that can affect the measured gene expression levels. Normalization is one step to minimize the systematic variations in the measured gene expression levels. Appropriate normalization procedures must be implemented so that the expression levels can be effectively compared across biological samples within an experiment and between different experiments.

In this study we used the dataset composed of 1,472 single color 15k Agilent arrays from different experiments. All experiments were performed in the same laboratory, on the same tissue (fathead minnow ovary), and with a range of treatments for which we can hypothesize the original assumptions to be correct. Non-linear normalization methods: quantile normalization and a slightly modified cyclic loess normalization, fastlo normalization. We applied two network inference algorithms, based on mutual information, Accurate Cellular Networks (ARACNE) and Context Likelihood of Relatedness (CLR) to infer the network model from combining the datasets. Results indicated that fastlo was found to be best normalization as the correlations between interacting genes were enhanced and models obtained from combined datasets revealed that the networks were associated with specific biological processes or potential relevance for ovary biology.