Breast Cancer Genomics Informatics

Kaustubh Shinde[1], Xiangdong Liu[1,2], Maureen Sartor[1], Prachi Kothiyal[2], Mario Medvedovic[1,2]

[1]Department of Environmental Health, University of Cincinnati Medical Center, PO Box 670056, Cincinnati, OH 45267-0056

[2]Division of Biomedical Informatics, Cincinnati Children's Hospital Research Foundation, Cincinnati, OH 45229

We describe the project to provide bioinformatics support for breast cancer related genomics initiatives and research projects affiliated with the BCERC at University of Cincinnati. The goal of the project is to provide data management solutions, comprehensive analyses and the portal for disseminating data and results of the analyses for genomics data generated locally. In addition to organizing and analyzing locally generated data, one of our aims is to construct and analyze a comprehensive database of publicly available breast cancer related microarray data and incorporate data generated by other BCERC centers. There are several aspects of this project that distinguishes it from other well-known initiatives for organizing cancer genomics data (http://www.oncomine.org). First of all, we are focused on breast cancer genomics which will allow us to more quickly process all target datasets. Oncomine has a long backlog of breast cancer genomics datasets. Second, we will include data from model organisms and provide means to link orthologous genes between different species. Third and most important, we will apply unique analytical tools that we developed and continue to develop for modeling gene expression data across diverse microarray datasets (http://eh3.uc.edu/gimm). The methodological developments are currently funded by an R01 grant from NHGRI and they are intimately linked to BCERC. The goal of methodological efforts is to develop a comprehensive mathematical framework and computational tools for identifying statistically significant patterns in functional genomics data. The mathematical framework is based on context-specific Bayesian infinite mixture models (http://eh3.uc.edu/gimm/csimm) and the resulting computational tools will be disseminated as open source software and a Bioconductor package. The validation of the procedures will include identification of putative regulators of transcriptional program in initiated mammary epithelium through the large scale computational analysis of public microarray data. The wet lab validation will be performed by siRNA-based knock-down experiments involving putative regulators identified in computational analysis. Consequently, the exhaustive collection of breast cancer microarray datasets will serve for validation of the computational tools while, on the other hand, the ability to query results of such analyses as well as ability to centrally query large number of breast cancer genomics datasets will offer a unique tool for the wider research community.