Using an integrated Stem Cell Knowledgebase for identification of novel "stemness" related genes and their targets.

Henry Yang, Team Leader Stem Cell Database Group, Bioinformatics Institute (BII)

Recently, stem cell biologists have intensively applied large-scale technologies (from MPSS to SAGE to microarray) for global gene expression profiling of stem cells in the hope to define the “stemness” and understanding the regulatory mechanisms for self-renewal and cell fate decisions. Unfortunately, the extreme discrepancies in the lists of putative “stemness” genes between different labs raise the doubt on integration of such large-scale high-throughput data from independent studies. It is possible that different strains and different culture protocols may be the main contributors to such discrepancies. It is also very likely that, since each lab used only a small number (2-4) of replicates for deriving the putative “stemness” genes, improper application of data analysis methods for a small sample-size problem can lead to high inaccuracies and large variance. Thus, integrated analysis of data from different labs can increase the sample size and may lead to more stable results in discovery of “stemness” genes. We have been developing a Stem Cell Knowledgebase integrating information from various published experiments to facilitate data integration and knowledge discovery in stem cell research. Applying novel data integration methods and pathway information/knowledge, we are able to determine stem-cell enriched genes and activated/deactivated pathways more accurately.