Identification of Genes Expressed in Early-Stage Lung Cancers

S. Jones, P. Ruzanov, J. Asano, Y. Butterfield, E. Garland, N. Girn, R. Guin, L. Hsiao, M. Krzywinski, W. Lam, S. Lam, S. Lee, K. Lonergan, C. MacAulay, T. Olson
M. Oveisi, P. Pandoh, P. Saeedi, U. Skalska, L. Spence, D. Smailus, J. Stott, K. Teague, R, Varhol, G. Yang, S. Zuyderduyn, J. Schein, M. Marra

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Genome Sciences Centre

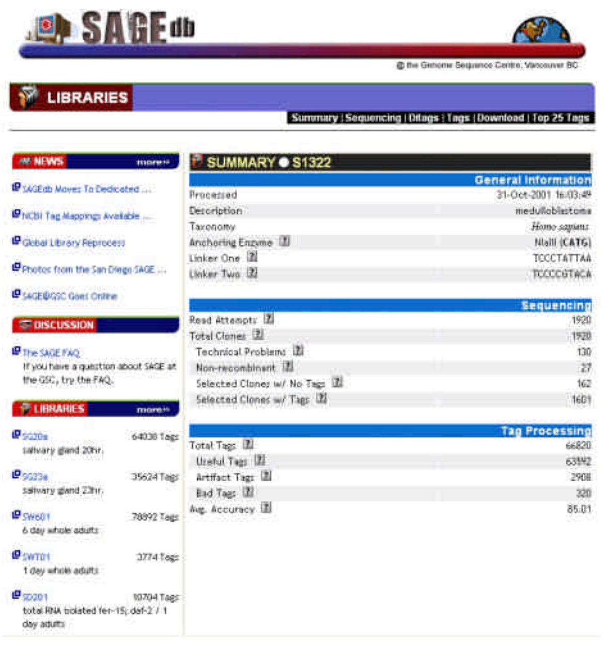
www.bcgsc.ca
info@bcgsc.ca

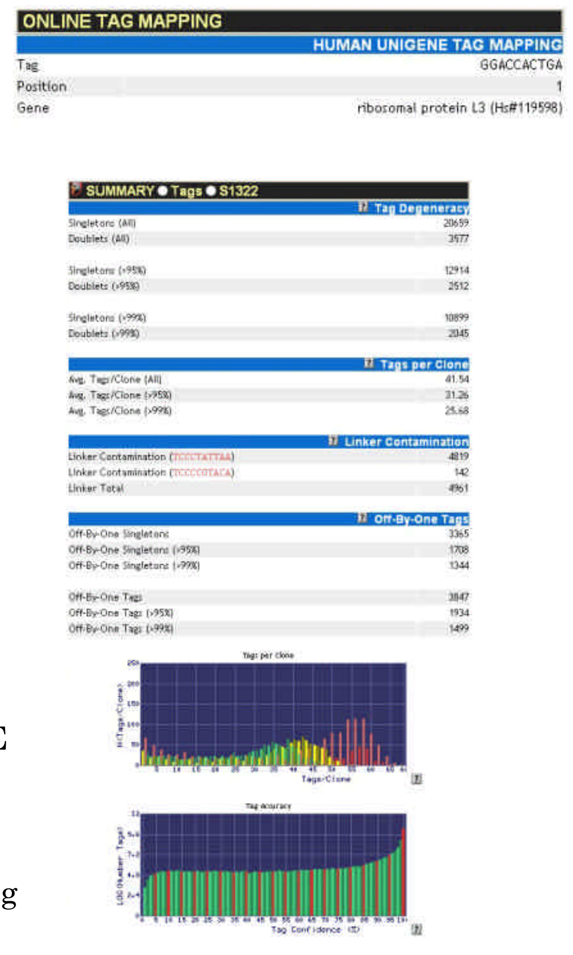
Abstract

We have established a multi-disciplinary approach to the large-scale high-throughput identification of genes involved in early stage cancers, under the auspices of Genome Canada. As part of this study we will be analyzing 166 Serial Analysis of Gene Expression (SAGE) libraries to profile mRNA expression in approximately 20 different tissue types. One particular focus has been the development of gene expression profiles for early stage lung cancer. By utilizing novel bronchoscopy techniques we have been able to obtain tissue samples from lung carcinoma in-situ. Although only small amounts of tissue are obtained, these are sufficient for the preparation of SAGE and micro-SAGE libraries. By sequencing the SAGE libraries we have been able to generate comprehensive and deep expression profiles for lung cancer. Ten libraries have been analyzed to date, five each for both cancerous and normal samples. We have identified, 14,624 different transcripts in these data. Eighty-four of these genes are consistently up-regulated within the lung cancer samples (58 of which are lung cancer specific and not observed in normal lung) and 122 genes have been identified as being consistently down-regulated in these early-stage lung tumors.

To facilitate analysis of our SAGE data we have developed an expression visualization tool, SageSpace and an analysis database, SageDB. Using these tools, we have been able to assess the heterogeneity between lung cancer samples. Through orthogonal comparison to other publicly available SAGE and EST data we have also been able to assess the representation of these lung cancer associated genes in other adult tissues. Our analysis indicates that four of the transcripts found to be lung cancer specific show expression in only one normal adult tissue. This approach has also allowed us to compare lung cancer expression with expression data from other cancers and other diseases. This information will be useful in determining the potential of existing drug therapies for application in lung cancer treatment. The SageDB system links expression data with a number of biological databases, including PFAM, SwissProt, OMIM, dbEST, BIND and KEGG. This functionality allows us to recreate expression profiles for specific biological pathways, e.g. apoptosis, allowing heterogeneity to be assessed in a pathway specific manner, the results being graphically visualized using the SageSpace viewer. This work is funded in part by the British Columbia Cancer Foundation, the National Cancer Institute of Canada and Genome Canada.

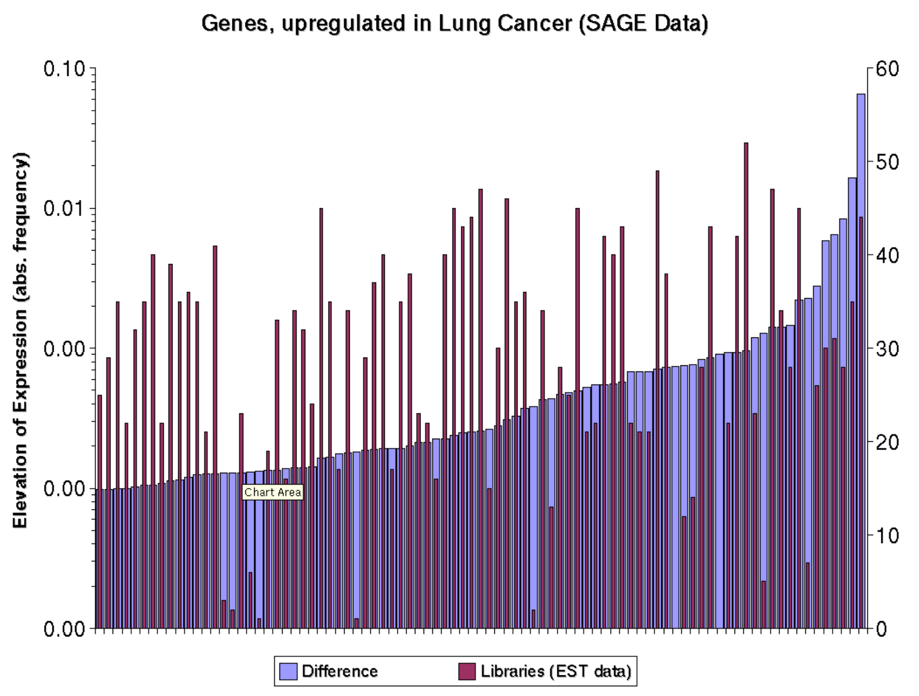
Data Management





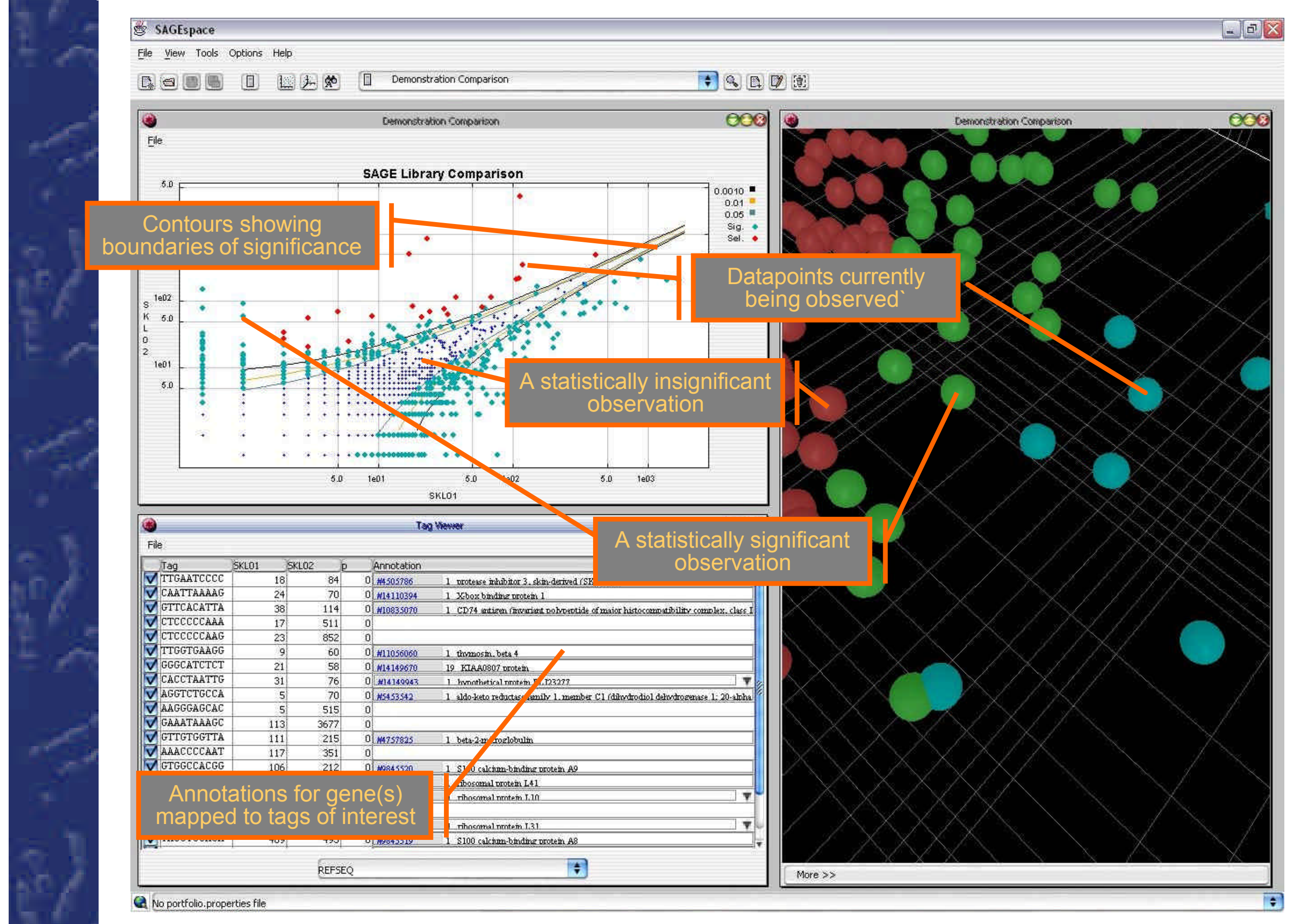
The panel above shows examples of online resources that Centre staff and investigators can utilize to view SAGE data. These tools are useful for ensuring quality control, viewing raw and processed data, making preliminary gene assignments, reading recent literature, browsing available software tools and following the progress of SAGE studies. Quality values for each sequenced tag are determined using PHRED and a probability for each tag being a correct sequence is derived.

Expression Analysis

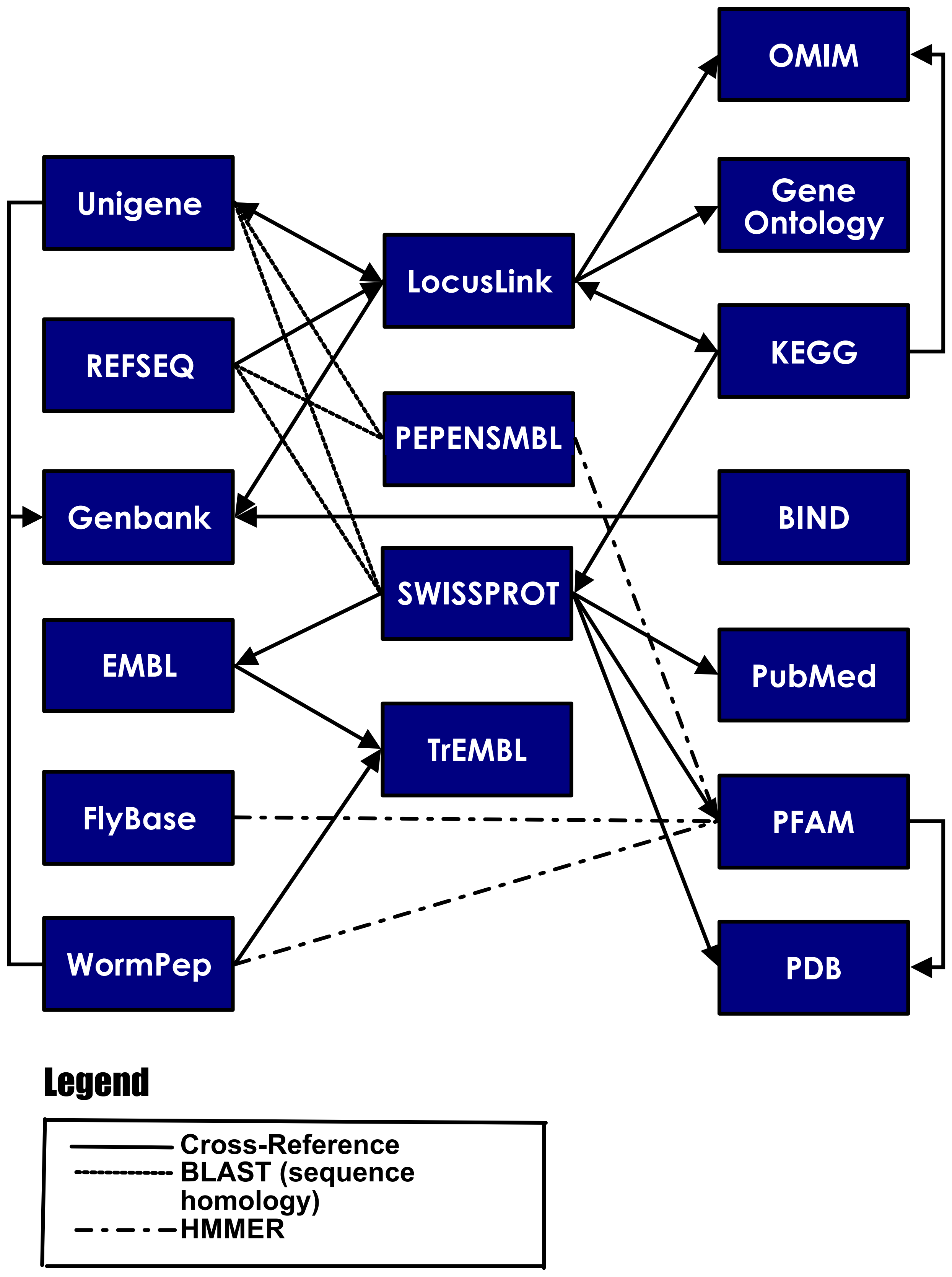


This graph shows genes up regulated in lung cancer (from 4 cancer libraries combined) as compared to normal lung tissue. We are also able to investigate the scope of the role of these in the human body by determining the number of EST libraries the genes are observed in.

SageSpace – The Analysis Environment


SAGEspace
File View Tools Options Help
Demonstration Comparison
SAGE Library Comparison
Contours showing boundaries of significance
Datapoints currently being observed
A statistically insignificant observation
A statistically significant observation
Annotations for gene(s) mapped to tags of interest
Tag Viewer
More >>

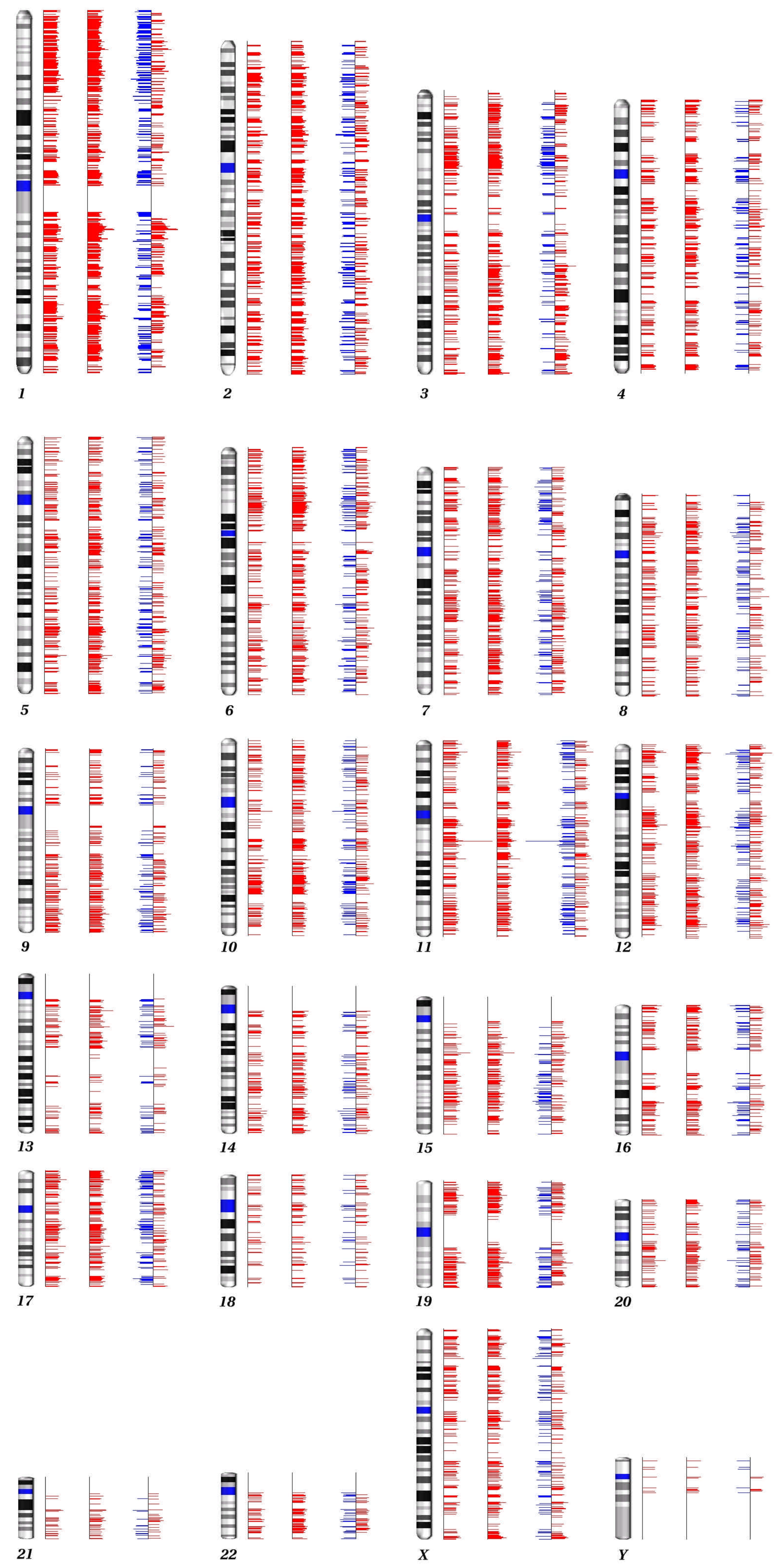
SageDB – The Analysis Database



There are many relationships between different databases that can be utilized to infer biological relevance from a SAGE experiment. The diagram above shows the relationship between some of the more common public databases. An important characteristic of our effort is to exploit the connections between disparate data sources. Where no connection exists, similarity can be inferred with various computational approaches. For example, BLAST will find similar sequences and HMM can classify protein motifs.

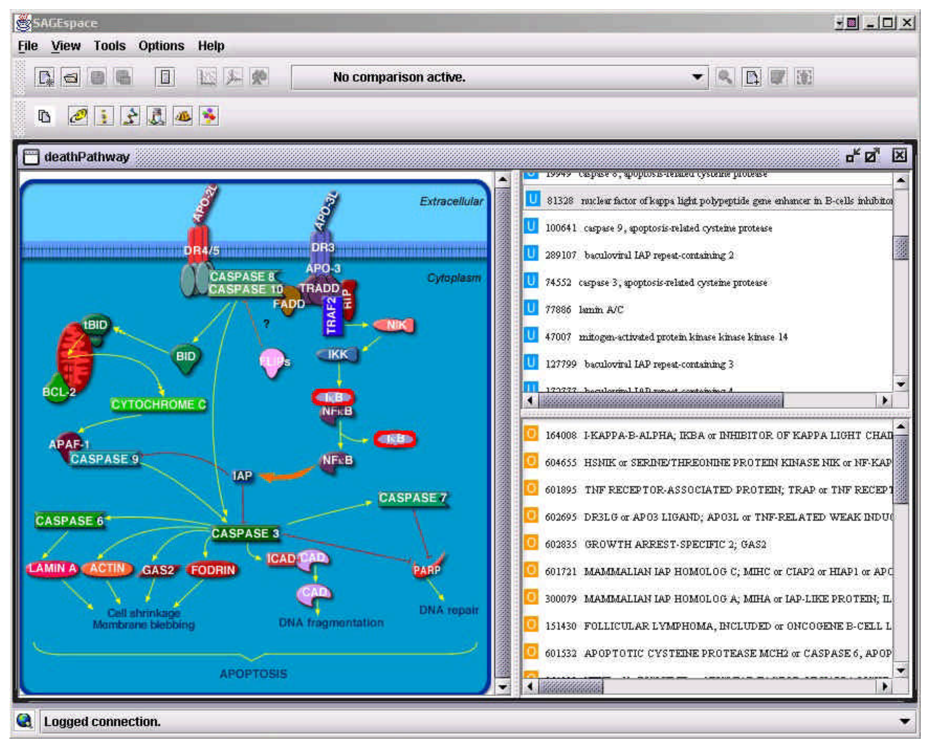
By utilizing information from a number of data sources, biological inferences can be assigned to the results observed in a SAGE experiment. Not only can particular observations be investigated, but targeted questions can be asked which will draw attention to a particular portion or portions of gene expression data (e.g. what kinase proteins show up regulation in a tumour).

An Expression Karyotype for Lung Cancer



Comparison of relative gene expression rates between normal lung and four combined lung cancer Libraries. The 1st and 2nd red columns indicate the relative level of expression in normal and cancer respectively. The 3rd column shows differences between the tissue states. Blue represents down-regulated and red represents up-regulated genes in cancer. Large gaps in expression data represent incomplete regions of the genomic sequence data.

Constructing Expression Profiles of Genetic Pathways



Expression profiles can be viewed and queried on an individual pathway basis. Shown here are the expression profiles of genes involved in Apoptosis. Pathway image courtesy of biocarta.com