# Exploring Gene Expression with the DISCOVERY Platform

**Zuyderduyn S, Varhol R, Oveisi-Fordoei M, Rusaw S, Jones SJM**

British Columbia Cancer Agency
Vancouver, British Columbia, Canada

Genome Sciences Centre

www.bcgsc.ca
info@bcgsc.ca

## Introduction

The study of gene expression can provides the ability to elucidate the molecular characteristics of a cell. These studies can concentrate on the temporal or spatial characteristics of gene expression, or abnormal pertubations of a cell's transcriptome brought on by disease.

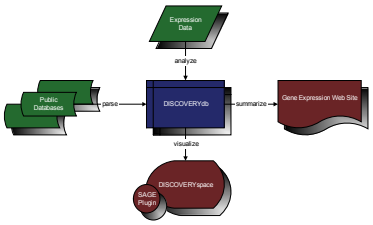
However, the transcriptome of a cell represents complex molecular interactions and machinery. This complexity introduces challenges in deriving the significance of gene expression changes.

We have attempted to build a system whereby the analysis of gene expression data can be made comprehensive, flexible, and visual. Our system, the DISCOVERY platform, embraces several philosophies to achieve this:

+ large-scale storage of existing biological knowledge
+ a system to rapidly characterize and incorporate new knowledge
+ the ability to apply and make available the results of new algorithms and analysis techniques
+ providing visual context to raw knowledge and analysis
+ the ability to rapidly develop software plugins to address the particular needs of given type of experimental approach

The DISCOVERY platform has been used to assist investigations of cell death in *Drosophila* early development, mechanisms of aging in the *C. elegans* model organism, telomerase-induced cell immortality, gene expression in embryonic stem cells, and the study of early-stage lung cancer.

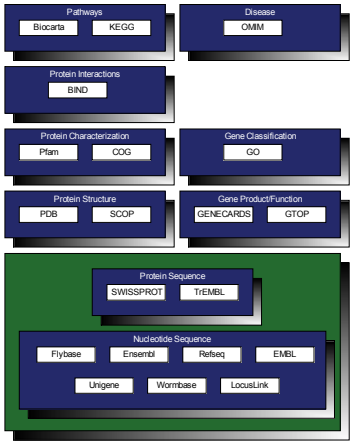
## Overview




The DISCOVERY platform consists of a) DISCOVERYdb where parsed public databases and analyzed expression data are stored, b) a web site where database contents are summarized and tracked, and c) the DISCOVERYspace application, where expression data and biological information can be interrogated and visualized.

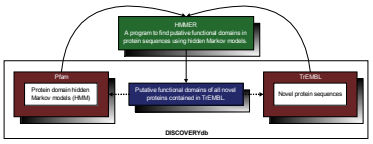
## DISCOVERYdb

The DISCOVERYdb system is a collection of parsing and schema characterization tools used to transform widely varying public databases into a well-formed MySQL database. This collection can be rapidly expanded to meet the specific needs of an investigator.




A partial list of DISCOVERYdb's collection of public databases.

The DISCOVERYdb system makes well-formed connections between these data sources. Although natural cross-references often exist between different data sources, large-scale computational analyses can be performed to generate new relationships.




A large-scale analysis of proteins to find functional domains.

As new sources of biological knowledge are made available or generated, the DISCOVERY platform can rapidly characterize and incorporate them into the system.

DISCOVERYdb also includes a wealth of publicly available SAGE data, as well as providing our expression laboratory with the means to store, and assess the quality of, our own SAGE data.

### Dictionary

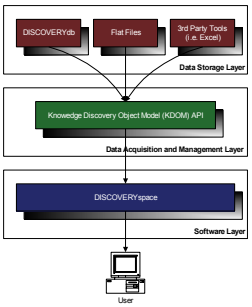
**SAGE:**
An experimental technique that splices a short (10bp) nucleotide fragment from the 3'-most NlaIII restriction site of cDNAs derived from mRNA transcripts. These tags are randomly concatenated to generate long chains suitable for high-throughput sequencing. The sequenced SAGE tags thus provide a profile of a cell's transcriptome.

**MySQL:**
A free, popular relational database system.

**Java:**
A programming language developed by Sun Microsystems. Java is popular for it's object-oriented philosophy, cross-platform compatibility, and web-friendly characteristics.

## DISCOVERYspace

The DISCOVERYspace software is a Java application designed to facilitate fast, flexible and intuitive interrogation and visualization of genomic and experimental data.




The DISCOVERYspace architecture.

The display and organization of data is user-directed. For example, the investigator may be interested in a number of interesting biological pathways. The user can query a pathway database for this information. Then, one can acquire additional available biological knowledge to further elucidate the pathway information.

A window showing gene components of a number of apoptosis-related pathways.

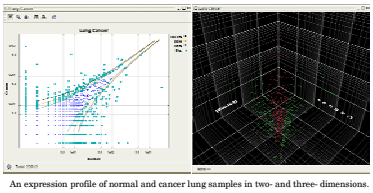
When appropriate, specialized visualization components can be programmed to provide additional functionality.

A specialized browser where pathway images from a number of sources can be displayed.

The application contains a number of useful features: including, keyword searches, cut-and-paste functionality to common office software suites, and local disk storage of interesting information for fast resumption of analysis tracks.

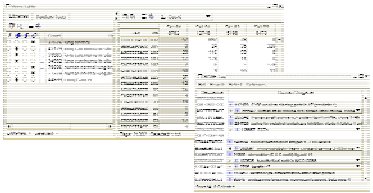
## Exploring Gene Expression

The serial analysis of gene expression (SAGE) technique uses short pieces of transcribed genes to create a near-global snapshot of a cell's transcriptome. We are involved in the analysis of a large number of cancer tissue SAGE libraries. By identifying genes that are abnormally abundant or absent in these samples, we hope to elucidate the molecular mechanisms of cancer and find possible diagnostic or treatment targets.



An expression profile of normal and cancer lung samples in two- and three- dimensions.

We have created a SAGE extension/plugin for DISCOVERYspace. These tools provide additional functionality useful for analyzing SAGE data. In particular, an investigator can create pair-wise comparisons of SAGE data and identify statistically significant observations. Biological inference can be assigned to these observations by utilize the core DISCOVERYspace functionality. For example, one can isolate SAGE data indicating up-regulation, and acquire gene assignments for each. These genes can be further elucidated with, for example, functional assignments, pathway membership, or subcellular localization predictions.



A component used to find SAGE tags of interest in multiple samples, and a window displaying gene assignments for a list of select SAGE tags.