Genedata **Expressionist®**

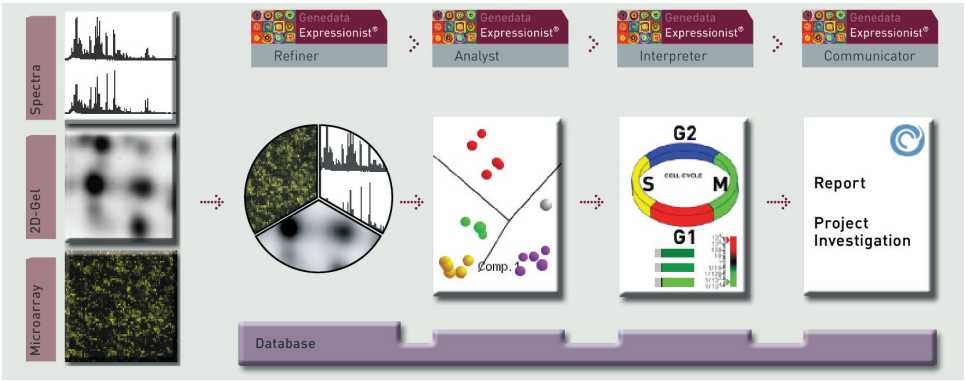
Biomarker Edition
A discovery tale

Genedata
solutions in-silico

Combining systems biology results to identify suitable cancer markers

Growth in genomic information and advances in fundamental research hold great promise for drug discovery and development. But translating basic research findings into next-generation diagnostics poses an enormous scientific challenge.
This dossier describes a collaboration between an elite pharmaceutical research team and Genedata, who will assume the role of expert partner in bioinformatics and computational biology. The scenario plays out in an oncology biomarker discovery team tasked to develop cancer biomarkers. The team will combine transcriptomic and proteomic data and use this information to guide their search for suitable metabolites for use as cancer markers.
We follow the team's progress from the acquisition of clinical samples, to data analysis and result sharing and finally, describe how they identify suitable metabolites for the next phase of their investigation.

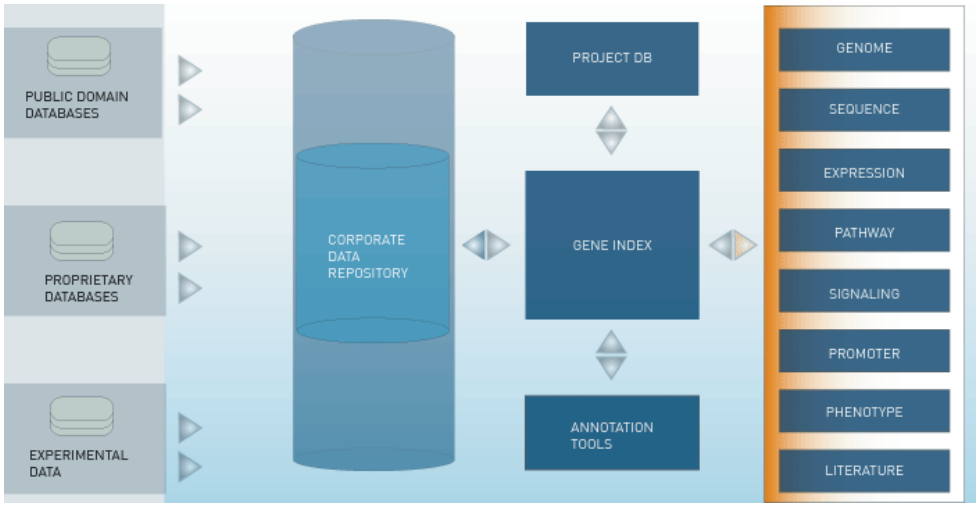
- Genedata's computational biology solution is tailor-made to support translational research. Users are placed at the center of a scientifically validated analysis and result management platform designed for developing next-generation therapies and diagnostics.
- The platform enables combined analysis of transcriptomic, proteomic and metabolomic results. High throughput quality evaluation tools assure that only high quality data are considered in downstream analysis.
- Knowledge is shared readily within the team and the platform integrates smoothly with existing IT infrastructure, software tools system biology content.



Clinical samples obtained from a major teaching hospital are prepared for analysis using RNA microarray and 2D gel technology.
Sarah is analyzing the transcriptomic results. Working with the team's biostatistician, a **Project** is created in the Database that will serve as a single point from which the team shares all project-related information.

A standardized data quality evaluation is performed and poor quality results are removed from downstream analysis. Sarah uses the quality control software module to perform this step. The module automatically placed the data in a **relational database**. A comprehensive quality report is generated automatically.

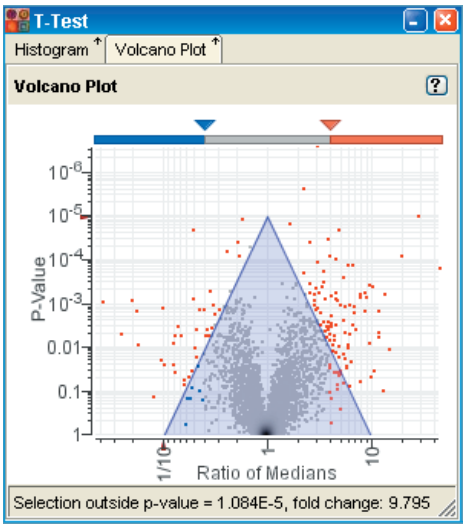
- **Projects** reside in the database and function as a single repository for raw data, analyzed results and for details of the experimental design.
- Advanced sample tracking and result handling functions form a bridge between research sites that may have very different organizational cultures and infrastructure: eg hospital databases, research LIMS at a high throughput facility. A high level of automation standardizes processing and ensures consistent reporting and auditing.
- The platform is based on industry standard data management technology (RDBMS).



Sarah finds 270 marker gene transcripts in a comparison between normal and cancerous tissue samples. Mappings between the microarray probe set identifier, the RNA it targets and its corresponding gene name place this result in biological context. The mapping also facilitates comparison with results from the 2D gel experiments. Sarah can pass the first major hurdle towards interpreting the results.

The entire analysis is performed in silico using Analyst's session functionality. Sessions can be paused and resumed, archived, and copied to colleagues. More than merely a report on findings, team members can open a session and perform their own analyses, sharing their results in turn. Face-to-face meetings can then focus on data interpretation.

- Mappings between probe ID and gene name guide the search for the biological information needed to compare expression results obtained from different technologies.
- A session is an electronic record of all analysis steps. It can be shared and archived.
- Inside a session: The interactive Volcano plot (below) is used to identify genes that meet a combination of n-fold and statistical criteria in the comparison between normal and cancerous samples. Transcripts with high statistical p-values in the t-test and substantial n-fold change are highlighted in red. These are marker genes.

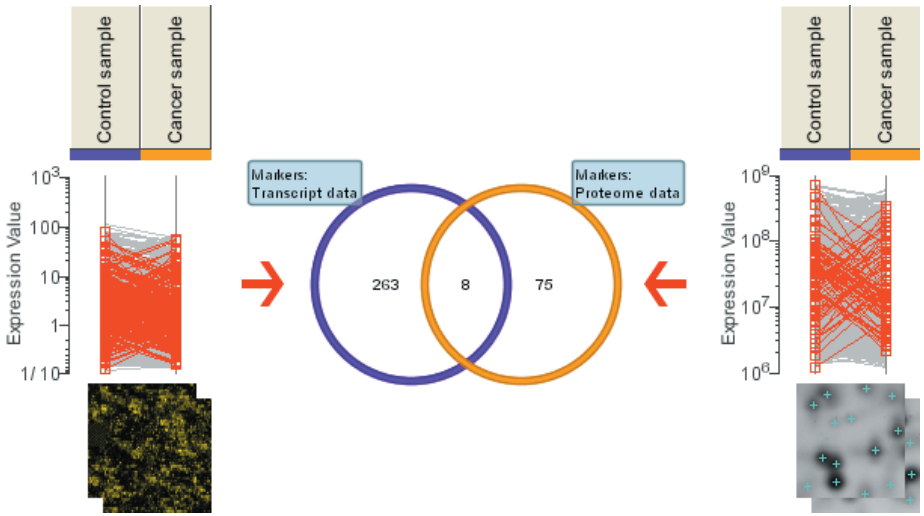


Frank prepares a list of markers from the 2D gel experiments. Starting with 300 spots with unusual expression patterns in the cancer samples, Frank sends 83 excised spots to the company's high throughput mass spectrometry facility. Peak lists are then sent for protein identification analysis and the results placed in the database.

Matching the identified proteins with the corresponding gene, Sarah and Frank explore the overlap between the results of the proteomic and transcriptomic experiments. They prepare for the meeting by reviewing each others analysis sessions and browsing the Project as the individual pieces of the dataset come together. The transcriptomic and proteomic expression profiles can be combined into a single session and compared directly.

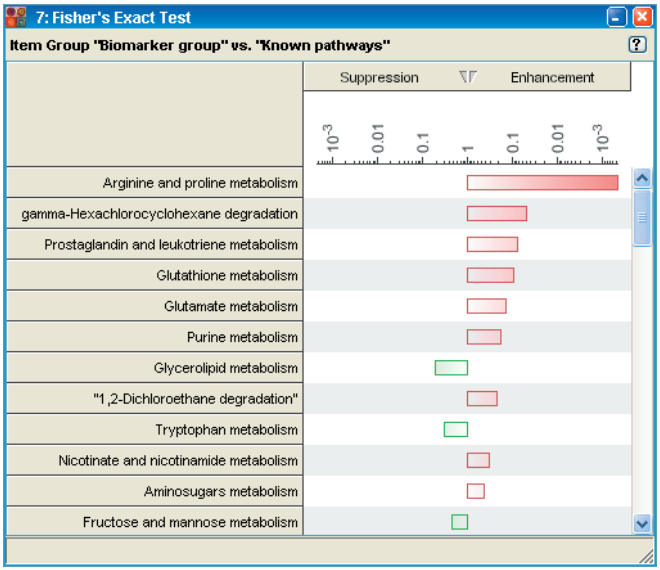
- The complex and multi-step task of identifying proteins from 2D gels is made tractable by a high degree of integration between databases and analysis tools. LIMS integration keeps track of the identity of excised spots sent off for MS analysis while the interactive visualizers of the 2D gel quality evaluation module are invaluable for high throughput quality evaluation. This makes it possible to identify poorly matched spots even when there are 10's–100's of gels.

- The Venn visualizer (below) indicates 8 genes identified in both the proteomic and transcriptomic marker lists. A larger list of approx. 250 marker is constructed based on these genes and from selected genes identified from the transcriptomic and proteomic results.



Prior knowledge about metabolites is critical to the effective use of MS technologies. Sample preparation methods and identification algorithms must be chosen carefully to ensure adequate sensitivity of the technique. Sarah and Frank are now confronted with a sizable challenge. They must use knowledge gained from the transcriptomic and proteomic experiments to guide the search for suitable metabolites to investigate using MS. Data must be placed in deeper biological context to meet this challenge. Biological computation experts at Genedata can help by annotating the biomarker list with scientifically validated information about the metabolic pathways to which each gene is associated. Tight integration between result information and advanced statistical tools makes it possible to explore the association in a quantitative fashion. Genedata suggest using Fisher's Exact Test to determine the most likely pathway.

- Using Fisher's Exact Test, over (or under)-representation of marker genes associated with a particular pathway is tested against the prediction that they are observed no more frequently than would be expected by chance.
- Markers involved in Arginine and proline metabolism are statistically over-represented (below).
- The Genedata analysis platform features a suite of sophisticated tools to test biological hypotheses, including multiparametric statistical tests and machine learning classification algorithms.



The biomarker team can now work with the MS facility with a suggestion as to which metabolic pathway might be involved.
Michael, chief scientist at the facility, brings specialized knowledge of how to most sensitively assay metabolites of a given mass. Reviewing the masses of metabolites associated with arginine and proline pathway, Michael selects sample preparation and identification methods appropriate for the target metabolites.
This will greatly improve the sensitivity of MS and improve the chances of finding metabolite markers that change measurably under diseased conditions.
Sarah and Frank complete this first phase of the investigation by submitting a report on these recommendations.

The database automatically keeps track of intermediate processing data such as MS peak and Scorer lists. Being able to drill down to this information is crucial for quality evaluation and the overall reliability of results.

Pathways involving arginine and proline metabolism are shown below.

Summary

Research teams face many hurdles in the race to gain clinical advantage from advances in fundamental biological research. As more and more biological information becomes available, bottlenecks in research process can prove extremely costly and overwhelm traditional approaches to data sharing and analysis.

We describe an approach to fruitful collaboration between an elite oncology research team and Genedata. Genedata's computational biology platform and mature collaborative approach offers substantial value those active in translational medicine and biomarker discovery. This approach can streamline operational aspects of research and solve specific scientific and bioinformatic problems.

Finished.

Direct comparison of z-transformed expression results from transcriptomic, proteomic and metabolimic technologies. From left to right, the choice of suitable metabolites is achieved through a consideration of the metabolic pathway most likely to show measurable effects in disease state. The rightmost plot shows twometabolites that have been chosen for further investigation using MS.