Extracting Biological Insight from Genomics Data: GeneSpring GX and Workgroup

Technical Overview

Author

Pam Tangvoranuntakul, PhD
GeneSpring Product Manager
Agilent Technologies, Inc.
Santa Clara, CA USA

Abstract

Agilent's GeneSpring Analysis Platform provides powerful, accessible statistical tools for fast visualization and analysis of microarray data. Designed specifically for the needs of biologists, both GeneSpring GX and GeneSpring Workgroup offer an interactive computing environment that promotes investigation and enables understanding of microarray data within a biological context. Regarded as the gold standard in expression analysis, GeneSpring GX allows you to quickly and reliably identify targets of interest that are both statistically and biologically meaningful. GeneSpring GX has over 9,500 references in Google Scholar, including over 1,600 in peer-reviewed publications. Developed on avadis from Strand Life Sciences, GeneSpring GX and the enterprise-level GeneSpring Workgroup are part of Agilent's GeneSpring Analysis Platform, an expanding suite of integrated software applications for systems-level research.

Agilent Technologies

Introduction

A key component of systems biology research involves producing heterogeneous, global data that measure various biological entities and events such as DNA, mRNA, proteins, microRNA, and exon splicing. GeneSpring GX allows researchers to perform integrated analysis of such heterogeneous data, enabling the identification of linkages and concordance of different data types that contribute to a more comprehensive understanding of the underlying mechanism of disease.

Integrated Platform for Multi-Omics Data Analysis

GeneSpring GX addresses the challenges in multi-omics analysis by providing comprehensive analytical and visualization tools for multiple data types within a single data analysis application. Each project in GeneSpring GX can contain one or more experiments of different data types, array platforms, or organisms. In this way, heterogeneous data such as gene expression, miRNA, exon splicing, genomic copy number, and genotyping, can be combined in a logical unit. Researchers can quickly analyze, compare, and view results from different experiments in a single user interface.

GeneSpring GX facilitates integrated analysis through its Translation Function tool, linking probes across data types, array platforms, and organisms that map to the same biological entity. Results from different experiments can be quickly compared in the Venn Diagram and the Find Similar Entity Lists tools. By supporting analysis and translation of multiple data types in a single application, GeneSpring GX sidesteps the issues of software interoperability and faulty semantic mapping to increase a researcher's ability to find linkages between data types.

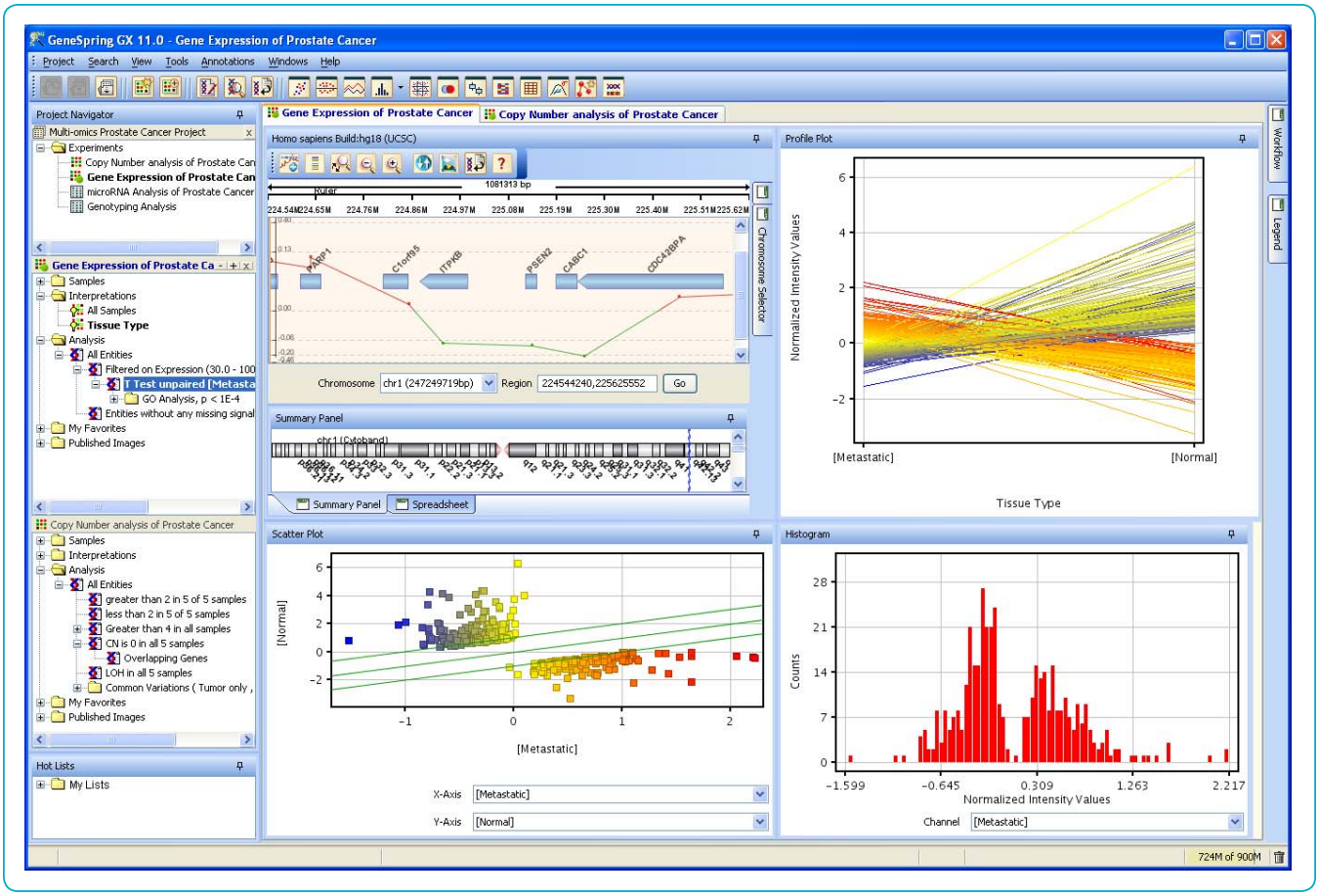

Figure 1: Project-based data organization enables you to store data from related experiments in a single workspace, regardless of data type, array platform, or organism. This organization facilitates comparison of results from related experiments.

Genomic Copy Number Analysis

GeneSpring GX expands data analysis capabilities to support the interrogation of genomic structural variations and their implications in disease susceptibility and progression. Providing workflows for paired and unpaired analysis of Affymetrix and Illumina genotyping array data, GeneSpring GX enables scientists to quickly detect regions of genomic copy number variation (CNV) and loss of heterozygosity (LOH) that are of potential importance to the biology under study. Once regions of interest are discovered, genes overlapping those regions can be identified and the biological impact of copy number variation can be assessed in downstream Gene Ontology (GO) analysis or pathway analysis. Features of the genomic copy number workflow include:

- Ability to create and use a custom reference in addition to packaged HapMap reference
- Batch effect correction method
- Circular binary segmentation
- Filters to identify copy-neutral LOH events and regions of allelic imbalance
- Identification of common variations across a set of samples

Genome-Wide Association Analysis

High-coverage genotyping arrays enable scientists to perform genome-wide association studies (GWAS) designed to interrogate associations of SNPs with qualitative or quantitative traits. GeneSpring GX provides a comprehensive workflow for case-control genetic association analysis of Affymetrix and Illumina genotyping array platforms. The flexible workflow supports a variety of experimental designs, as it offers a suite of statistical tests applied under various genetic models, multiple testing correction, and correction methods for population stratification. After identifying genes harboring

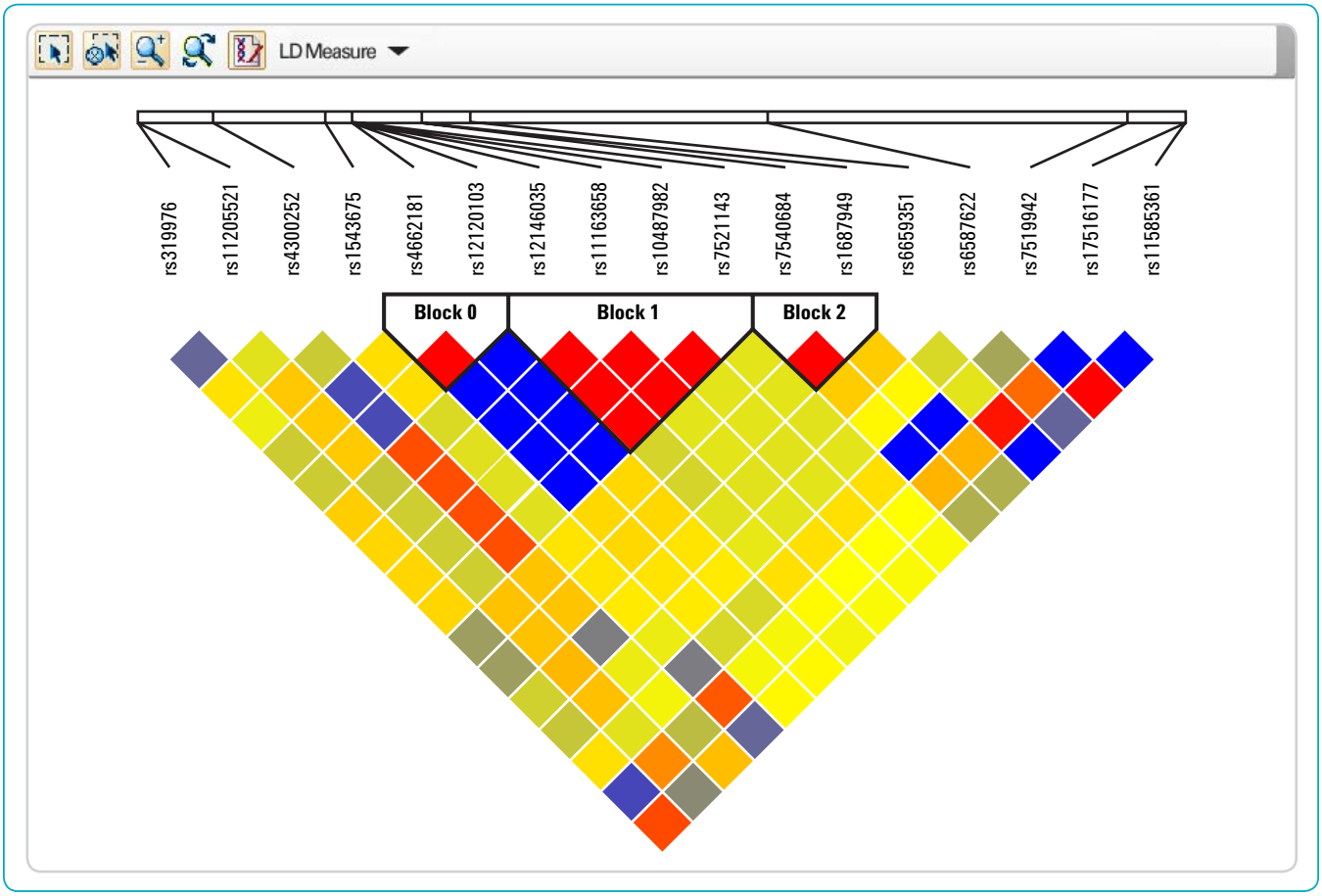

Figure 2: Linkage disequilibrium (LD) plot provides fine-grained representation of the LD blocks, which are colored according to the pair-wise correlation provided as D′ or R^2 values. SNPs within regions of strong LD can be saved as entity lists for further analysis.

SNPs or haplotype blocks associated with a trait, researchers can perform GO analysis and pathway analysis to determine what biological processes and pathways may be involved in the disease under study. Other key features for the genetic association workflow include:

- EIGENSTRAT and genomic control population stratification correction
- Tag SNP identification
- Haplotype inference and haplotype trend regression
- Pearson's chi-square, Fisher's exact, Cochran-Armitage, and chi-square correlation
- Logistic and linear regression
- LD plot

Transcriptomics Analysis

GeneSpring GX provides flexible and comprehensive workflows for a variety of transcriptomics applications. Depending on the researcher's level of expertise, data analysis can be performed in GeneSpring GX using either the Guided Workflow or Advanced Analysis mode. A broad spectrum of data pre-processing, linear and non-linear normalization methods is available for both one- and two-color expression data. Quality control can be performed using metrics specific to the microarray platform, enabling researchers to optimize pre-processing steps before statistical analysis. Other key features for transcriptomics applications include:

- Probe-level or gene-level expression analysis on virtually all microarray platforms, including Agilent, Affymetrix, and Illumina
- microRNA analysis and identification of gene targets using integrated TargetScan information
- Exon splicing analysis using multi-variate splicing ANOVA and filtering for transcripts on splicing index
- Real-time PCR data analysis
- NCBI Gene Expression Omnibus importer tool for expression data sets

Comprehensive Analytical and Visualization Tool Kit

At the core of GeneSpring GX is a set of statistical tests, algorithms, and visualization tools that allow researchers to analyze a broad range of experimental designs.

Statistical Tools for Testing Differential Expression

Using GeneSpring GX's comprehensive suite of statistical tests, scientists can apply differential analysis on a variety of experimental designs. The GeneSpring GX statistical toolkit includes parametric and non-parametric tests that can be applied to paired and unpaired experimental designs, permutative and asymptotic p-value computation, and multiple testing correction methods, including permutation based options. Post-hoc statistical tests can also pinpoint pairs of experimental conditions where differential expression is detected. Multivariate analysis is available to test the effects of each factor, and their interaction, on changes in expression across experimental conditions.

Pattern Discovery

GeneSpring GX provides a broad choice of tools to identify unique patterns in data. Clustering algorithms can be employed to group entities and samples based on the similarity of their expression profiles, revealing information regarding the biological function or the co-regulation of genes. GeneSpring GX also provides robust classification algorithms that use training data sets to find clinically predictive expression patterns. By offering multiple classifiers including Decision Tree, Support Vector Machine, Naive Bayes, Neural Network, and Partial Least Squares Discriminant, GeneSpring GX enables biomarker discovery for a variety of experimental designs.

Extensible Functionality with Jython and R Languages

GeneSpring GX enables scientists to write, execute, and save their own scripts to combine operations in GeneSpring GX with a more general Jython (Python with Java-class import capabilities) programming framework. Users can develop their own data transformation operations, automatically pull up data views, and run external algorithms within GeneSpring GX. An embedded R scripting editor also allows R scripts to be written and run from within GeneSpring GX. Any R function can be given access to GeneSpring GX data, with results being automatically incorporated back into the GeneSpring GX environment.

Intuitive Graphical Displays

GeneSpring GX displays data in ways that help researchers conceptualize and communicate information through various types of plots, graphs and diagrams that highlight different aspects of the data, and allow visual information to be

extracted in multiple ways. Virtually any graphical image can be exported as HTML or as a tiff, jpeg, png or bmp image compatible with publishing software applications. A powerful Genome Browser aids in visualizing genomic structural data and integrating results from multi-omics studies. Results from different experiments can be dragged and dropped into individual data tracks and viewed simultaneously. For example, genomic copy number data can be displayed as a data track along with gene expression data from a different experiment. The Image Overlay feature permits the user to overlay any data or annotation tracks, thus allowing researchers to qualitatively assess correlation between different data types. Other visualization tools in GeneSpring GX include Scatter Plot, MvA, Profile Plot, Histogram, and many more.

Biological Contextualization

Placing statistically significant findings into a biological context is a critical step in data analysis. GeneSpring GX facilitates this process by providing a single analysis environment enriched with tools fitted for both parts of the workflow. Instead of looking at results at the individual gene level, scientists can explore what biological process, molecular function, and biological pathways are involved in a disease process using provided GO analysis, gene set enrichment analysis (GSEA), gene set analysis (GSA), and pathway and network analysis tools. Scientists can then go quickly from statistical analysis identifying relevant entities, to modeling interactions that elucidate the underlying mechanism of the disease. By providing an integrated toolbox for statistical analysis and biological contextualization of multiple genomics data types, GeneSpring GX streamlines the process of gaining biological insights from microarray data.

Figure 3: The new interactive Genome Browser allows users to visually integrate heterogeneous data by simultaneously importing data tracks from multiple experiments, and permitting overlay of data and annotation tracks.

Pathway and Network Analysis

Genes and proteins interact in a biochemical network to orchestrate the biological processes involved in disease. GeneSpring GX's modeling capabilities allow researchers to quickly generate and dynamically explore these networks. Using a set of algorithms and provided organism-specific interaction databases, researchers can build a range of network types, including targets and regulators, transcription regulators, biological processes, and shortest connect. Networks can also be created based on user-specified MeSH terms. Networks generated in GeneSpring GX are dynamic and interactive, allowing overlay of expression values and analysis results onto the network, interrogation of nodes and edges, and expansion of network from selected nodes. The natural language processing (NLP) based algorithm can be applied to a body of text, HTML, pdf, and Medline XML to extract and add interactions to an existing interaction database. This allows researchers to customize network analysis to their experimental models.

In addition to network building, GeneSpring GX allows researchers to import and view any pathways in the BioPAX exchange format. These pathways can be used in the Find Similar Pathway tool to determine if there is an enrichment of the genes of interest in any pathways. Beyond its built-in pathway and network analysis tools, GeneSpring GX extends biological contextualization capabilities through its integration with Ingenuity Pathway Analysis (IPA), where lists of genes and experimental data can be seamlessly transferred between the two applications for iterative analysis.

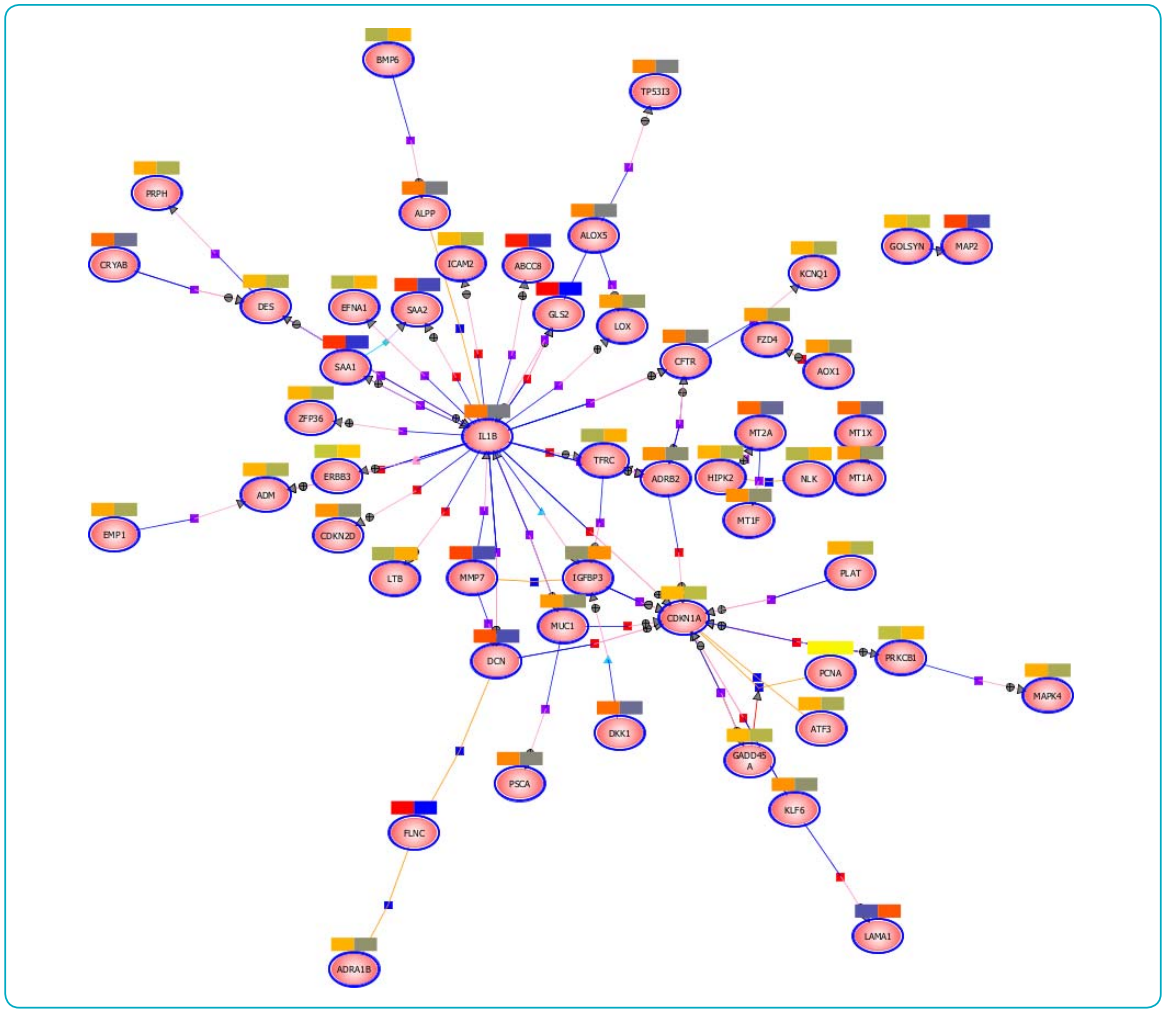

Figure 4: Pathway and network diagrams help place statistical results in a biological context. Direct navigation between biological pathways and their associated genes provides a rich user experience and systems-level insight.

Scalable Architecture

GeneSpring Workgroup

In addition to a desktop computing environment, the GeneSpring platform includes GeneSpring Workgroup, a scalable client-server product that addresses the need for higher computation and throughput and provides a secure enterprise solution for computing, storing, and sharing. GeneSpring Workgroup consists of a central server for secure content management of multi-omics data that is connected to various components and clients. The Data Browser and Web Client interfaces allow scientists to search, visualize, and share analysis results. Importing data into the Workgroup server for shared access and analysis is done efficiently through automatic batch sample loading and experiment creation. Computationally intensive analyses can be off-loaded to a remote server farm. Web services of the Workgroup server provide secure programmatic access to any data in the system.

Conclusion

GeneSpring GX and GeneSpring Workgroup—both part of the GeneSpring Analysis Platform—are powerful tools that enable scientists to perform complex, integrated analysis of various kinds of microarray data for a systems-level view into genomic function. Across any scale of project, GeneSpring provides a robust set of statistical analysis and visualization tools for use with a range of biological data types, helping scientists extract more meaning from their data.

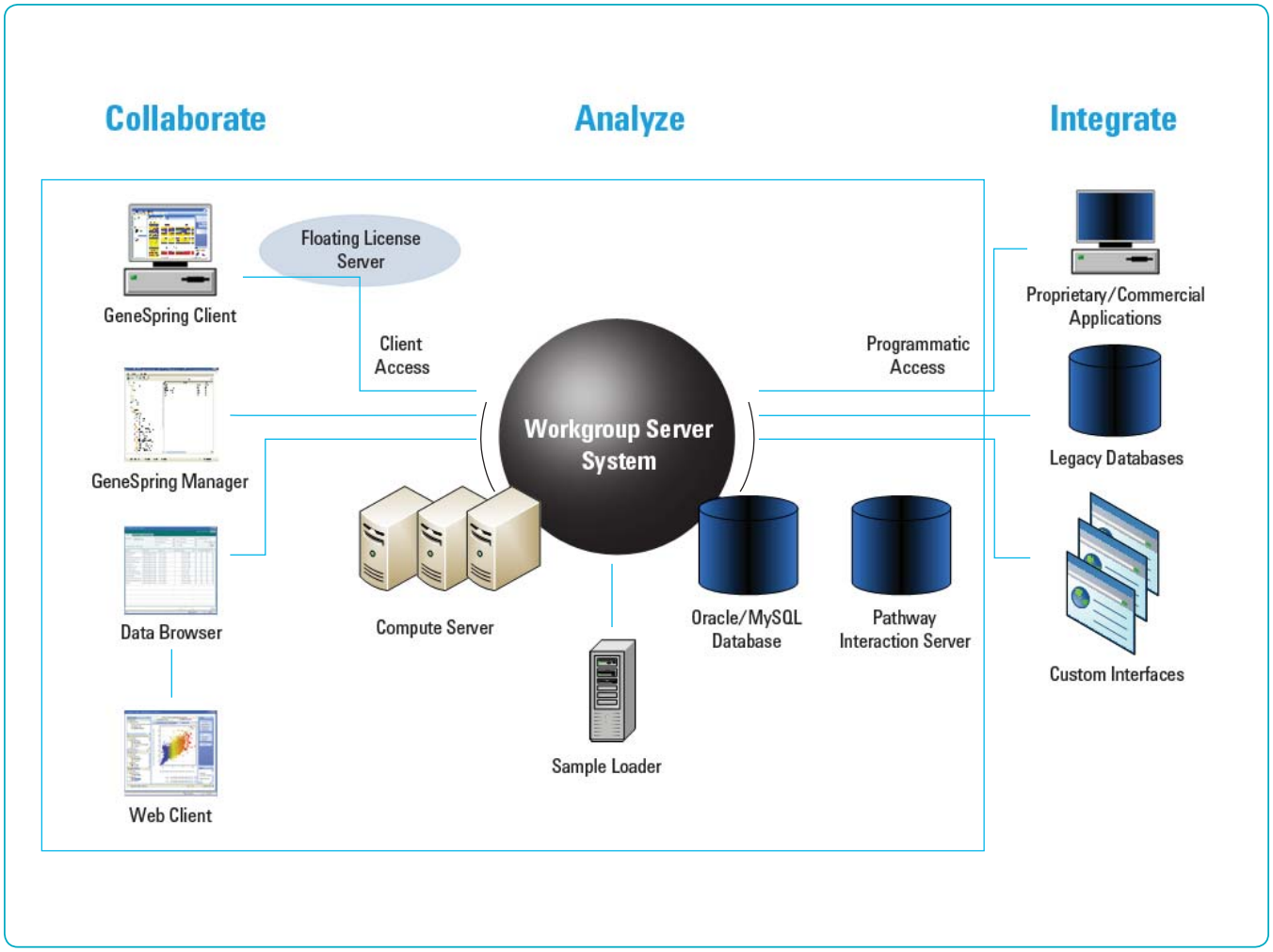

Figure 5: GeneSpring Workgroup provides enterprise-level central storage and analysis capabilities for multi-omics data.

For More Information

Learn more:
www.genespring.com

Buy online:
www.agilent.com/chem/store

Find an Agilent customer center in your country:
www.agilent.com/chem/contactus

U.S. and Canada
1-800-227-9770 (Option 2, Option 3)
sig_sales@agilent.com

Asia Pacific
inquiry_lsca@agilent.com

Europe
info_agilent@agilent.com

Research use only. Information, descriptions and specifications in this publication are subject to change without notice.

Agilent Technologies shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance or use of this material.

© Agilent Technologies, Inc. 2009
Printed in the USA, November 16, 2009
5990-5005EN

Agilent Technologies