NEOPLASIA
www.neoplasia.com

Volume 13 Number 9 September 2011 pp. 771–783 771

G-DOC: A Systems Medicine Platform for Personalized Oncology[1]

Subha Madhavan*, Yuriy Gusev*,[2], Michael Harris*,[2], David M. Tanenbaum†, Robinder Gauba*, Krithika Bhuvaneshwar*, Andrew Shinohara†, Kevin Rosso†, Lavinia A. Carabet*, Lei Song‡, Rebecca B. Riggins*, Sivanesan Dakshanamurthy*, Yue Wang‡, Stephen W. Byers*, Robert Clarke* and Louis M. Weiner*

*Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA; †ESAC, Inc, Rockville, MD, USA; ‡Deparment of Electrical & Computer Engineering, Virginia Tech, Arlington, VA, USA

Abstract

Currently, cancer therapy remains limited by a "one-size-fits-all" approach, whereby treatment decisions are based mainly on the clinical stage of disease, yet fail to reference the individual's underlying biology and its role driving malignancy. Identifying better personalized therapies for cancer treatment is hindered by the lack of high-quality "omics" data of sufficient size to produce meaningful results and the ability to integrate biomedical data from disparate technologies. Resolving these issues will help translation of therapies from research to clinic by helping clinicians develop patient-specific treatments based on the unique signatures of patient's tumor. Here we describe the Georgetown Database of Cancer (G-DOC), a Web platform that enables basic and clinical research by integrating patient characteristics and clinical outcome data with a variety of high-throughput research data in a unified environment. While several rich data repositories for high-dimensional research data exist in the public domain, most focus on a single-data type and do not support integration across multiple technologies. Currently, G-DOC contains data from more than 2500 breast cancer patients and 800 gastrointestinal cancer patients, G-DOC includes a broad collection of bioinformatics and systems biology tools for analysis and visualization of four major "omics" types: DNA, mRNA, microRNA, and metabolites. We believe that G-DOC will help facilitate systems medicine by providing identification of trends and patterns in integrated data sets and hence facilitate the use of better targeted therapies for cancer. A set of representative usage scenarios is provided to highlight the technical capabilities of this resource.

Neoplasia (2011) 13, 771–783

Introduction

With the sequencing of the human genome and availability of high-power computational methods and a variety of high-throughput "omics" technologies (e.g., genomics, transcriptomics, and metabolomics), cancer research and care are poised to undergo a revolutionary change. These new technologies and approaches have fueled the rise of systems biology, which is now fully established as a discipline. The new and emerging field of systems medicine, an application of systems biology approaches to biomedical problems in the clinical setting, leverages complex computational tools and high-dimensional data to derive personalized assessments of disease risk. Systems medicine offers the potential for more

Abbreviations: CIN, chromosomal instability; CRC, colorectal cancer; dbSNP, the single nucleotide polymorphism database at the NCBI; G-DOC, Georgetown Database of Cancer; GI, gastrointestinal; miRNAs, microRNAs; OMIM, Online Mendelian Inheritance in Man; PCA, principal component analysis

Address all correspondence to: Subha Madhavan, PhD, Lombardi Comprehensive Cancer Center, 2115 Wisconsin Ave NW, Suite 110, Washington, DC 20007.
E-mail: sm696@georgetown.edu

[1]The G-DOC development effort was partly funded by the National Cancer Institute's *In Silico* Centers of Excellence Program (HHSN261200800001E) as well as the Center for Cancer Systems Biology (U54-CA149147).

[2]These authors equally contributed to this study.

Received 9 June 2011; Revised 28 July 2011; Accepted 1 August 2011

 1522-8002/11/$25.00
DOI 10.1593/neo.11806

effective individualized diagnosis, prognosis, and treatment options. Achieving this goal requires the effective use of petabytes of data, which necessitates the development of both new types of tools and a new type of physician—one with a grasp of modern computational sciences, "omics" technologies, and a systems approach to the practice of medicine. As part of this transformation, clinicians will need views of integrated biomedical data from disparate sources and will begin to use validated *in silico* methods for analysis. A critical factor in the success of systems medicine will be the ease with which high-quality, high-dimensional data can be integrated, redistributed, and analyzed both within and across functional domains. This is enabled through effective application of translational bioinformatics [1], which is defined as the development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data into proactive, predictive, preventative, and participatory health, and helps link knowledge across biologic and clinical realms [2].

To enable the practice of an integrative translational and systems-based approach to research and medicine, we at the Lombardi Comprehensive Cancer Center, Georgetown University, have developed a feature-rich, novel, and shareable research translational informatics infrastructure to allow physician scientists and translational researchers to mine and analyze a variety of "omics" data in the context of consistently defined clinical outcomes data for cancer patients. By providing a powerful but easy to use interface, Georgetown Database of Cancer (G-DOC) was designed specifically to address the activation barrier for use of biomedical informatics tools by basic, clinical, and translational researchers. G-DOC contains a wide variety of analytic tools and capabilities, including integrated viewers for genomic features and three-dimensional drug-target complex structures. To help support effective patient group comparisons, G-DOC supports flexible clinical criteria browsing to enable selection of specific patient cohorts and facilitates the generation of detailed reports and informative publication-quality plots. Internal chemical compound libraries can be screened easily using the integrated structure and detailed molecular property search functions, with the goal of identifying new therapeutic candidate molecules. G-DOC also allows researchers to securely share knowledge with others through a powerful suite of collaboration-enabling features operating within its secure environment.

The first public version of the G-DOC Web portal was launched in April 2011 for the Georgetown University research community and their collaborators, providing cancer researchers with a broad range of data reduction, visualization, and analysis tools and a large knowledge base of published "omics" data sets from previously published cancer clinical studies and a smaller set of private data sets. Currently G-DOC includes data collected from more than 2500 samples of breast cancer and nearly 800 samples of gastrointestinal (GI) cancers (liver, colon, stomach, and pancreas). Four types of "omics" data are supported: mRNA and microRNA (miRNA) expression, copy number variation, and metabolite mass spectrometry data. All are linked to de-identified patient clinical information, markedly increasing their value. G-DOC also contains a manually curated database of small molecules as potential drug candidates for key biomarkers/target proteins and a set of curated cancer findings from integrated data sets and publications.

The G-DOC data repository is also designed to store multiple types of metadata associated with individual samples and patients including demographic data, clinical outcome, and tumor-specific phenotype data that could be either quantitative or qualitative, and could be either categorical or continuous. The data in G-DOC are uniformly processed using validated algorithms within the R-based bioinformatics toolbox (Bioconductor) [3], formatted and mapped using R scripts, and then uploaded to the central database. The data and the analysis results are shared within a G-DOC collaborative group, or a set of groups, administered by the data provider to provide controlled access to data and analysis.

Specific data analysis tools in the G-DOC environment include differential expression analysis, heat maps and hierarchical clustering, principal component analysis (PCA), survival analysis (Kaplan-Meier), gene-disease, gene-compound, gene-protein interaction networks rendered in the Cytoscape environment, and a growing collection of more specialized tools such as a toolbox for copy number alteration (CNA) data analysis. The latter suite of analyses includes chromosomal instability (CIN) index calculations for DNA segments, cytobands, and whole chromosomes based on data from CGH array and single nucleotide polymorphism (SNP) array technologies. The results of omics data analysis are mapped onto an integrated human genome browser at the level of either the individual patients and/or cohorts of patients, each defined by clinical attributes for ease of viewing and analysis. G-DOC allows researchers to combine the commonly used clinical information-personal history, physical examination, laboratory studies, radiology studies, family history, and other pertinent data—with a detailed "omics" analysis of the patient's cancer to facilitate exploration of the clinical and molecular factors that determine disease outcome. The results section provides a detailed and demonstrative workflow of a prototypical analysis of a breast carcinoma data set that highlights many of the value-adding capabilities of the G-DOC resource, including cohort selection, group comparison, PCA, and target exploration.

Key Features of G-DOC

G-DOC has easy-to-use search capabilities for clinical data, studies, biospecimen, omics data, small molecules, and key published findings. The G-DOC Web portal was designed to provide powerful bioinformatics capabilities to users with a variety of backgrounds and skill levels with computational tools. One of the most effective ways for a new user to begin using G-DOC's capabilities is through the "Quick Start" page. The left-hand side of this page consists of a series of selection options, and it allows users to filter the data; thus, the user can focus on only the most appropriate data sets for their needs. The right-hand side of the page displays a graphical summary of the available data that fits the specified criteria. An illustrative example can be seen in Figure 1, which shows the number of breast cancer studies available in the database that have data on recurrence; a tool-tip feature helps display the various "omics" and clinical data elements available for each of these studies. Data from patient cohorts can be further analyzed from this page in a variety of ways using a right-click on the study name.

Cancer research findings can be searched by entering names of genes, proteins, cancer type, investigators, or authors in the simple textbox search on the G-DOC home page. Findings are not meant to be comprehensive or cover all known cancer biology but instead provide a quick search and retrieval mechanism for key discoveries in the disease areas of interest to investigators who provide data to G-DOC. It is anticipated that this collection will be significantly expanded over time, and may be augmented by other cancer summary data compilations.

G-DOC Systems Biology Analysis and Visualization

G-DOC supports compute-intensive, high-memory tasks such as class comparison, hierarchical clustering, principal component analysis,

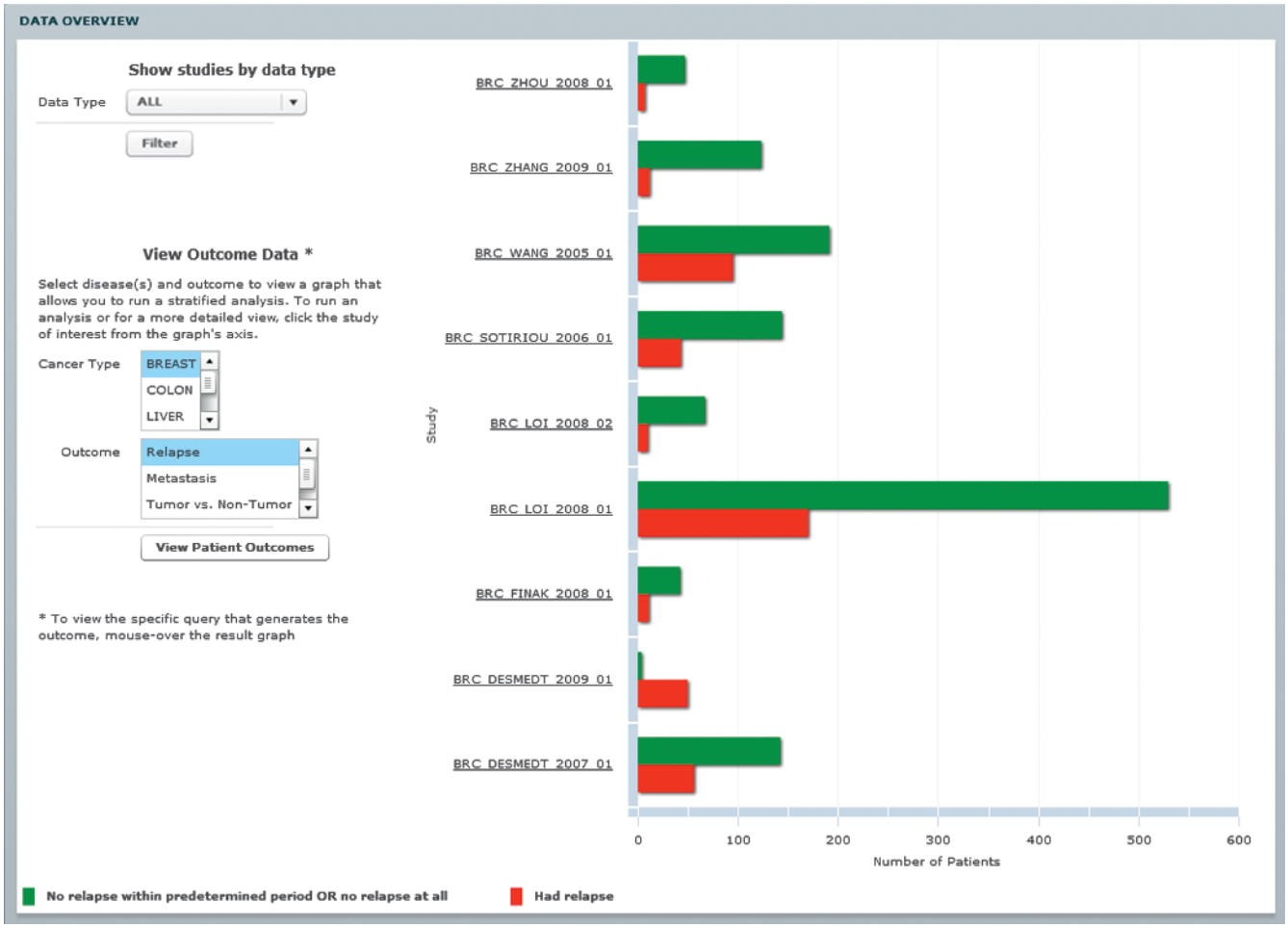

Figure 1. G-DOC quick search showing the number of breast cancer studies available in the database patient annotations for neoplastic relapse.

and network analysis for transcriptomic, genomic, and metabolomic data. Because these data sets could be as large as 4 GB, the development of an analytic cluster to allow for several simultaneous analytic jobs was needed to support community-level use of these services. Data in G-DOC can also be easily used to perform advanced systems biology analysis of regulatory pathways and interaction networks of genes, miRNAs, and metabolites that are both perturbed and most relevant to the available phenotypic changes.

miRNA has recently emerged as a new and important class of cellular regulators. There is strong evidence that aberrant expression of miRNA is associated with a broad spectrum of human diseases including cancer, diabetes, cardiovascular, and psychologic disorders [4–7]. The relatively small number of miRNAs discovered in humans to date (~1733 miRNAs, miRBase17.0 [8]) are involved in regulation of a large number of human genes, perhaps as many 80% of known genes [9]. miRNAs have exceptional potential as biomarkers because of their relative abundance, highly specific expression, and stable presence in serum and plasma [10]. In fact, circulating miRNAs are sensitive biomarkers for colorectal cancer (CRC) detection and compare favorably with the fecal occult blood test [11]. Circulating miRNAs in both urine and serum hold tremendous potential as biomarkers for both early detection of GI cancers and prognostic assessment. Processed miRNA data (see Methods section) can be visualized using heat maps and PCA plots to identify signatures that distinguish patient groups of interest (e.g., cancer *vs* normal, relapse *vs* nonrelapse). As an example, Figure 2 shows miRNAs differentially expressed between colorectal cancer and normal samples stored in G-DOC. A meta-analysis of differentially expressed miRNAs from stage II, III, and IV CRC samples can be derived using the Venn diagram feature and then exported for further pathway analysis.

Metabolomics is a rapidly evolving field that aims to identify and quantify the concentration changes of all the metabolites in a given biofluid, or tissue extract, from a patient. The anticipated contribution of metabolomics to the field of biomedical science is highlighted by its presence in the current National Institutes of Health roadmap [12]. The application of metabolomics to understand the manifestation and progression of complex diseases such as GI cancers represents a powerful means to identify the earliest markers associated with attributes such as recurrence and treatment response. G-DOC includes a sophisticated data analysis pipeline (see Methods section) to enable detection of potential prognostic and diagnostic molecular markers in (noninvasive) serum and urine using metabolomics. Figure 3 shows a PCA plot using 42 metabolite peaks that differentiate between recurrent and nonrecurrent cases in a currently private G-DOC GI cancer cohort (unpublished observations; public access to these data will be available through G-DOC upon acceptance for publication). Individual samples,

represented as points on the PCA plot, can be selected to view further clinical details of that specific patient.

DNA copy number changes are common in cancers, often driving underlying biology and affecting clinical outcomes. G-DOC provides the genomic "map" of patient copy number profile and the significant consensus regions derived from the CIN index. The utility of this novel technique has been shown in the identification of a correlation between CIN index and the grades of ovarian cancer subtypes [13].

Differences in a plethora of genomic features—including SNPs, miRNAs, and gene copy number—can be visualized in the G-DOC genome browser by selecting the chromosomal position or by using a gene identifier. Several genomic features are available as data tiers including phenotype information for mendelian disorders from Online Mendelian Inheritance in Man (OMIM) and annotated SNPs from dbSNP. The integration of information from OMIM, dbSNP, and other sources helps to focus the investigator on genomic features that are most likely to be functionally significant. Figure 4 shows copy number changes in a region of chromosome 12 from a pilot GI cancer study. Clear differences in copy number changes can be seen between two patients, a case that recurred and one that did not, viewed alongside additional genomic features such as miRNAs and SNPs in the region. Additional genomics features will be made available over time to supplement this type of view.

My G-DOC

The G-DOC application allows for configurable security levels for studies, and data can be made public (accessible to all users), restricted to one or more collaboration groups, or available only to the data owner, as dictated by the provider of the data. Previously published data are always made public on loading within G-DOC. Collaboration groups are limited groups of users, specified by the group manager, who can share information among the members without exposing it to the entire user population. The collaboration group manager has full control

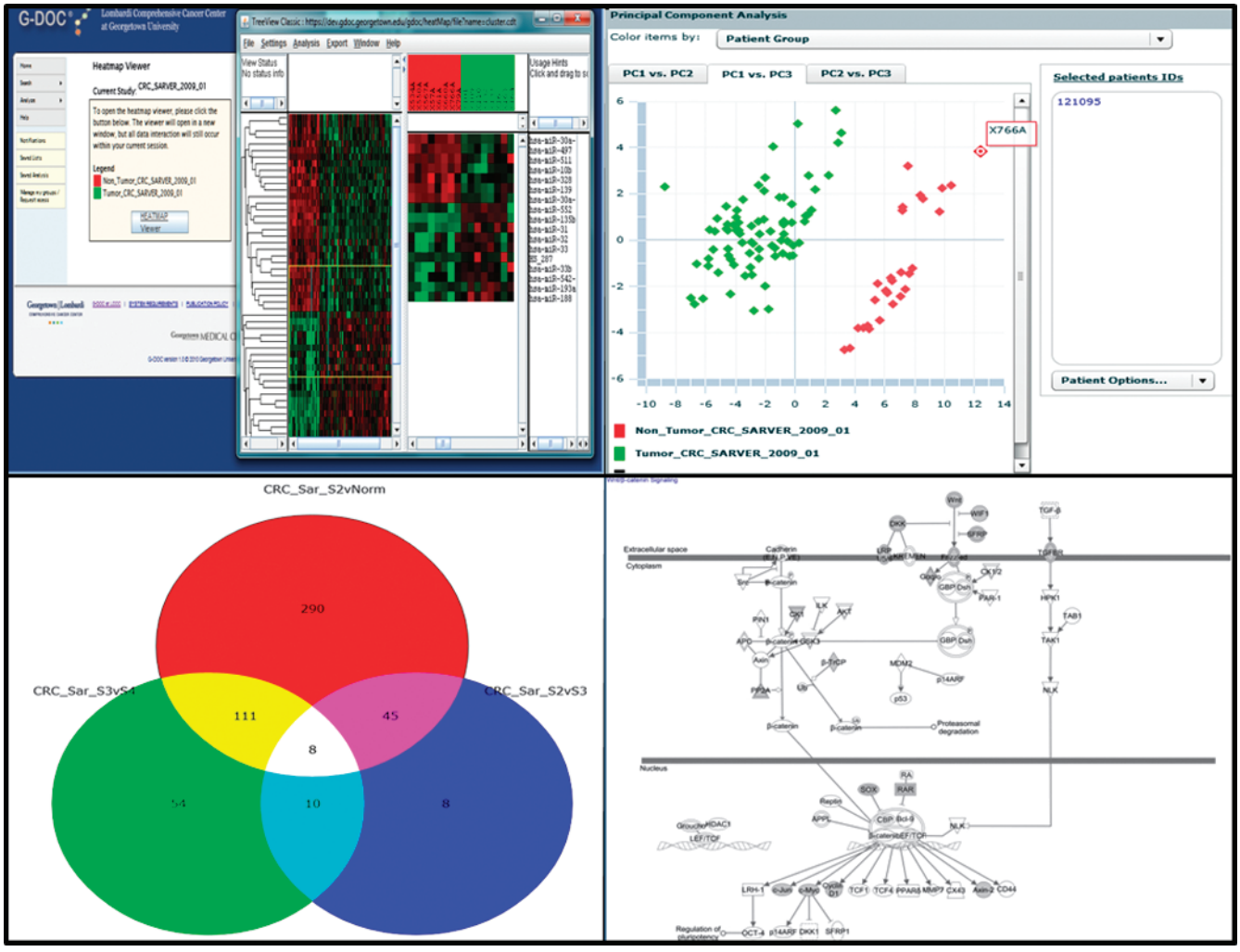

Figure 2. Analysis of miRNA expression data in G-DOC: Analysis and visualization of differentially expressed miRNA in CRC samples *versus* normal samples. Left, Heat map viewer showing clusters of coexpressed miRNAs. Middle, PCA scatter plot of tumor *versus* normal samples based on expression data for 61 miRNA showing well-separated clusters of tumors and normal samples. Right, Venn diagram showing only partial overlap between miRNAs differentially expressed in CRC stage II, III, and IV, with only eight miRNAs found to be in common for all three sets of miRNAs. Far right, WNT signaling pathways with predicted targets of the eight miRNAs shown in gray. Analysis of predicted targets has shown that this small group of miRNA regulates WNT signaling pathway known to be affected in colorectal cancer.

Figure 3. PCA using 42 differentially expressed metabolite peaks between relapse and nonrelapse cases with colorectal cancer; fold change of 1.5 or higher; $P \leq .01$.

over access to restricted data and to either approve or reject requests for access. Collaboration groups allow users to share lists of patients, genes, reporters, and analysis results in a secure, collaborative environment that fosters communication and team science. Additional administrative functionality provided in G-DOC includes the ability to manage user accounts and collaboration groups. The sum of these features provides a workspace for groups that are either working on a data analysis project or writing a grant as a team and can share analysis results, tools, and biomarker lists within their collaboration group—all within a secure and managed environment.

G-DOC System Architecture

G-DOC is a Web-based application that provides users with a comprehensive set of analysis routines and visualizations for a rich user experience. The application is written in Groovy & Grails, an open-source development framework that runs on the Java Virtual Machine. The jQuery JavaScript library is used to provide users with a cohesive and interactive interface. For more complex data visualization, the Adobe Flex framework provides users with integrated visualization components that can handle complex charts and graphs and allow these displays to interact with other functions within the application. Besides the components developed in-house, G-DOC also incorporates many third-party tools that provide data visualization capabilities. For example, Java TreeView [14] is used to display heat maps, Cytoscape [15] to display interaction networks, and JBrowse [16] provides a genome browser with multiple annotation tracks.

G-DOC is engineered and architected with future scalability as a top priority. As such, the application and architecture were designed to scale horizontally. The analysis server, Web application, and database are each deployed on different virtual machines and sit behind a load balancer. As the load on the application server increases, more virtual machines can be added behind the F5 load balancer to keep up with demand.

The G-DOC infrastructure consists of services and domain objects, using the common open-source frameworks Spring and Hibernate (an industry-standard object relational mapping technology). G-DOC has a set of RESTful services that use JSON as a transfer medium, allowing the different components to communicate with G-DOC in a simple manner. A third-party tool, Lucene (cross-platform text search engine), is used to index the database and provide users with a global search capability.

The G-DOC analysis server provides an extensible framework for analysis of study data. Analysis functionality includes the ability to perform group comparisons (*t* test, Wilcoxon) and clustering (PCA and hierarchical clustering). The analysis capability is implemented

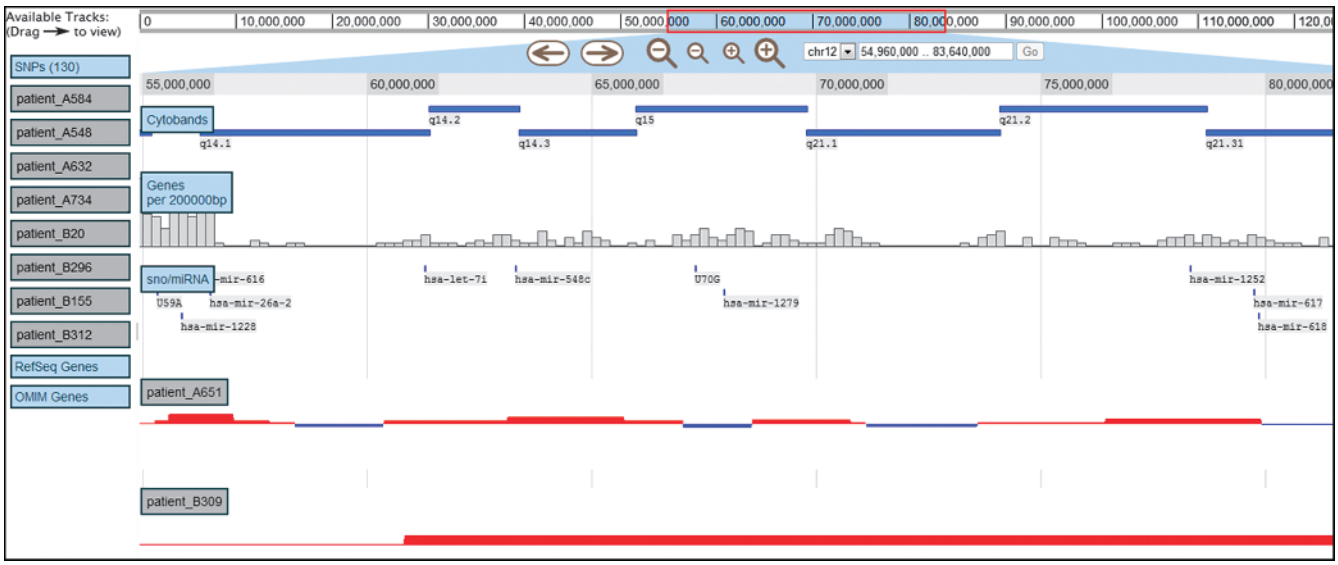

Figure 4. Analysis of processed copy number data in conjunction with clinical information within the G-DOC genome browser. Patient tracks can be dragged to the workspace to view genomic and clinical details. The "omics" tracks can be dragged in to see features that map to various locations on the genome.

using Java technologies including Java Messaging Service and the Java Executor class library for multithreaded processing. The analysis infrastructure is hosted on a virtual machine with 16 GB of dedicated memory. The analysis compute resources are easily scalable to a larger memory or compute capacity as needed.

G-DOC provides functionality to save user-created lists of genes, reporters, or patients for future reference and downstream analysis. A robust set of list operations and visualizations is available including the ability to perform intersections, unions, differences and to create and export publication-quality Venn diagrams. User lists are persisted in the database so that they can be referenced in future sessions and shared with other users.

The G-DOC security infrastructure provides for secure login as well as project- and role-based data access. User authorization in G-DOC is implemented using the Spring security module which allows for user provisioning. All communication between the G-DOC browser session and the middle-tier is encrypted using the https protocol.

Methods

Currently, data collection within G-DOC is focused primarily on areas of notable research and clinical strength at Georgetown University and, as such, contains data from predominantly breast and GI cancers. However, over time, this collection will broaden considerably and in a manner consistent with feedback received by the user community. Because facilitating translational research is the overarching goal of G-DOC, high-throughput "omics" data sets that have corresponding clinical information from human subjects are preferentially entered and will be so for the foreseeable future, although a few "tumor-*versus*-normal" data sets are also present. A list describing the current data collection, including types of data available, is available at all times on the G-DOC Web portal front page.

The data collection of G-DOC is composed of two divergent sets—public data that are available to all registrants and private data that are supplied by, and accessible to, individual investigators and their collaboration groups. Public data sets are typically obtained from repositories such as National Center for Biotechnology Information Gene Expression Omnibus [17] and EBI Array Express [18], whereas private data sets are uploaded to a secure SSH file transfer protocol server for handling by an analysis team. A set of standard operating procedures are followed before data are accepted from collaborators to ensure that all data are deidentified in accordance with Health Insurance Portability and Accountability Act regulations; G-DOC neither stores nor distributes patient identifiable information.

Once data sets have been obtained, either from public repositories or through private transfer, the "omics" data are preprocessed and the clinical information is mapped to the existing data structures as a precursor to loading within the G-DOC database. Uniform preprocessing and normalization ensures maximum comparability between analyses and studies and ensures that the data within G-DOC provides the greatest scientific leverage to the user community. Specific pipelines have been established, tailored to each data type, and emphasize standard and uniform data preprocessing to ensure utmost quality, a key factor in minimizing noise and false-positives. All data and accompanying query features are subjected to rigorous QC procedures before being made available in the production environment.

Several files are created describing the clinical attributes with respect to their type and vocabulary, outlining specifics such as format and range. Special files are also created for each data type in the study describing the mapping between the clinical and corresponding high-throughput data samples (after preprocessing). The summary, study characteristics, and contact information are captured in a separate format. This set of files is stored in a version-controlled data repository using a consistent naming convention to describe each study: Cancer-abbreviation_Principal-investigator_Publication-year_iteration (e.g., BRC_WANG_2005_01 for Wang et al., 2005; PMID: 15721472). Special attention is paid to capturing and persisting the disease outcomes and end point information because these serve to enable a series of value-added features of G-DOC (e.g., Quick Start, interactive Kaplan-Meier plots) that better support translational research activities. Studies are stored within G-DOC in an Oracle 11g relational database, which consists of 44 common tables. For each new study loaded, a separate schema is created consisting of a set of 12 study-specific tables. All processed data files pertaining to a particular study are loaded separately onto a computation-centric server designed to handle high-throughput data analysis. Analyses that are run against study data reference their respective processed (binary) file to complete a variety of statistical routines. All analysis routines that run in the G-DOC environment are written in the R language to ensure modularity, ease of deployment, and high performance on the computational server nodes that provide analytic services to the user community. A high-level overview of data, annotations, and analysis available in G-DOC are illustrated in Figure 5. The processing that occurs before data entry is detailed below for the four major types of "omics" data presently available in G-DOC (mRNA, miRNA, metabolomics, copy number). Whereas G-DOC does not currently contain any high-throughput sequencing information (exon sequencing, CHIP-Seq, RNA-Seq), these and other related data types will be supported in a future version of G-DOC.

mRNA Expression Data

Much of the data currently within G-DOC are mRNA expression data produced by array hybridization experiments, including both two-channel ratio data and single-channel intensity data. These are retrieved in a raw .CEL file format (Affymetrix, Santa Clara, CA) or a tab-delimited text file format (Agilent, Santa Clara, CA), as appropriate, from the public archives or from the laboratory that generated these data. Other formats are being considered for future versions of the tool. Preprocessing of microarray data primarily involves normalization with either Robust Multichip Average [19] or Quantile Normalization [20] followed by log transformation of the data. More information on these standard normalization strategies is available at http://www.bioconductor.org. Significant postprocessing effort is expended to ensure data quality and retention of the biologic information provided. Transcripts (mRNAs) are mapped on the genome in the JBrowse genome browser interface based on Build 36 of the National Center for Biotechnology Information genome.

miRNA Expression Data

MicroRNA is a subject of growing interest for the clinical, translational, and basic science cancer communities, and G-DOC supports this data type. A data preprocessing pipeline was developed for miRNA expression data that supports the major highthroughput platform formats: oligonucleotide microarrays (Agilent and Illumina) and real-time quantitative polymerase chain reaction arrays (Life Technologies, Inc, Carlsbad, CA). Microarray-generated data sets are processed from raw data files using global median normalization [21,22], whereas real-time quantitative polymerase chain reaction data are processed using comparative C_T method [23] and normalized to the average signal of

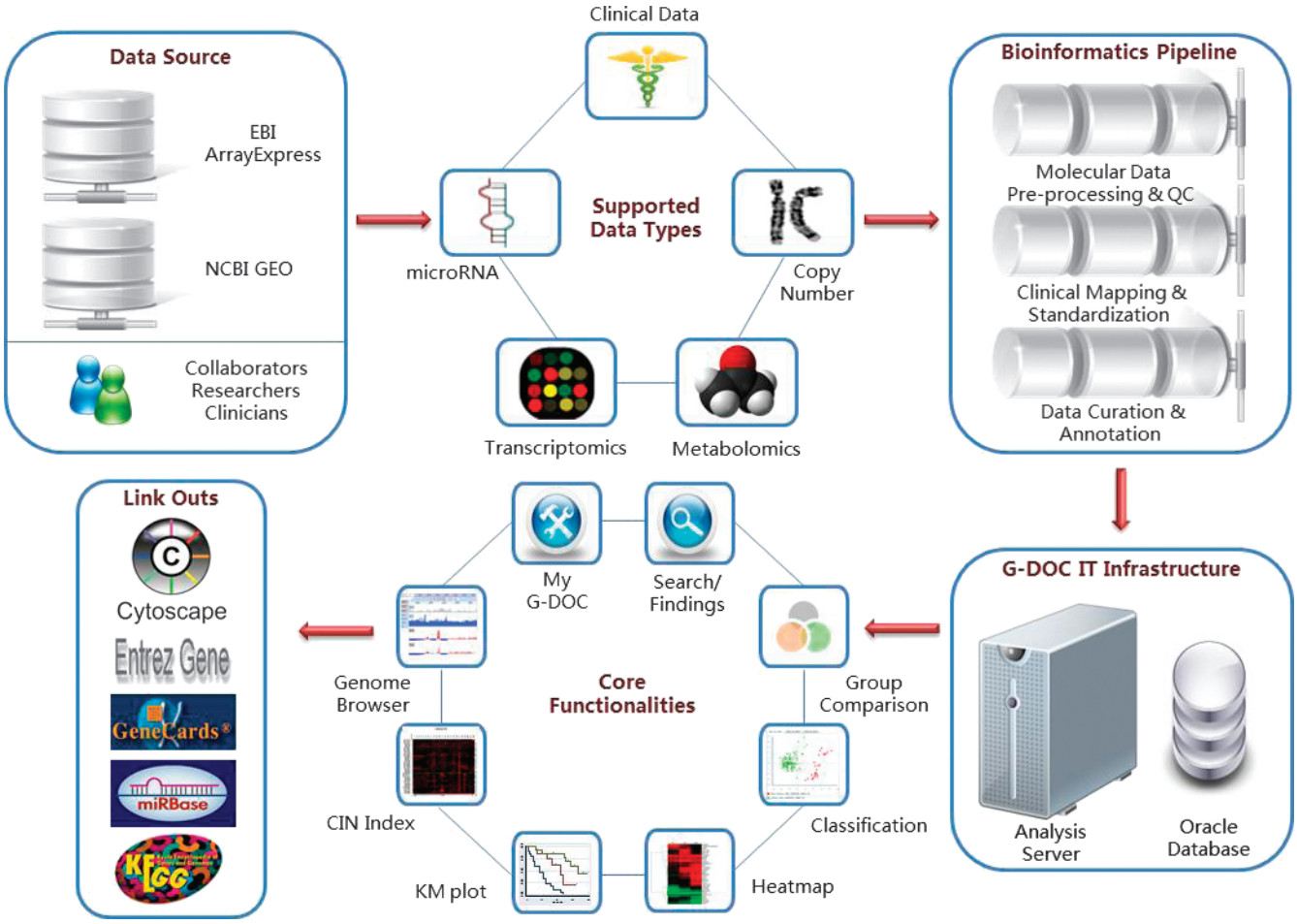

Figure 5. Overview of data and analysis features in G-DOC. Data (public or private) are uniformly processed through standard bioinformatics pipelines and made available to various analysis tools through a clinician- and researcher-friendly Web interface.

endogenous controls [24]. These miRNA reporter IDs are mapped to mature miRNA accession numbers in the miRBase [25] and hyperlinked to online public databases (miRBase, Entrez, and iHOP), providing instant access to comprehensive miRNA genomic and deep sequencing information as well as predicted targets. miRNAs are also mapped on the genome using the JBrowse genome browser interface.

Metabolomics Data

Metabolomics is one of the newer "omics" sciences and aims to study global profiles of small-molecule metabolites within a biologic system under a given set of conditions. Typically, these experiments are performed on biofluids such as urine, saliva, or blood plasma, but isolated cells and tissues may also be used. The G-DOC metabolomics data collection is exclusively mass spectrometric data, but the data structures are sufficiently generic to support other typical metabolomics data types (e.g., nuclear magnetic resonance, gas chromatography) in the future. For mass spectrometry data, a number of vendor-specific software programs, tailored to the specifics of the acquisition hardware, are available to convert spectral data into universal data exchange formats such as network common data form (http://www.unidata.ucar.edu/software/netcdf/), mzXML [26], and mzDATA. To ensure maximal future flexibility, the G-DOC preprocessing pipeline is built to work with all these formats. The metabolomics data contained within G-DOC were processed into a data matrix format with samples as columns and peaks/metabolites as rows and were normalized row-wise or column-wise in a sequential manner to minimize systematic variance and improve the performance for downstream statistical analysis.

DNA Copy Number Data

Considerable attention has been paid to understanding the gross chromosomal modification events that are common within many types of cancer. Although the technologies used have progressed (e.g., SNP and cDNA array hybridization replacing loss of heterozygosity), significant interest remains in identifying the aberrations that occur within the development and progression of neoplasias. To ensure that G-DOC can enable investigators to best use this type of data, a data processing pipeline was developed using R (http://www.r-project.org/) for analysis and visualization of DNA copy number data obtained from a variety of platforms. Raw data from the most common platforms, Affymetrix SNPchip and Agilent CGH arrays, are preprocessed using D-Chip [27] to extract a signal for individual probes. Piecewise constant segments of copy number profiles are estimated based on the fused margin regression method [28]. Probe-level data are further processed to calculate copy number segments and CIN index [13], one

of the value-added analyses that come pregenerated within G-DOC. Segment data are used for calculation of CIN index at the level of whole chromosomes and individual cytobands [13].

Results

A G-DOC Storyboard

To exemplify the powerful integration that G-DOC provides to analyze large molecular and clinical data sets, we demonstrate here how a user could generate and validate a scientific hypothesis using this system. In this example, we will test the hypothesis that there are reproducible gene expression differences that can be identified between recurrent and nonrecurrent estrogen receptor–positive (ER+) tumors in tamoxifen-treated node-negative breast tumors. The G-DOC Web tool permits us to perform this analysis quickly and easily using nothing but a collection of publicly available data sets obtained from the biomedical literature. This exercise will include identification of a molecular profile in one public study [29] and its validation in another [30].

1. Can we identify, within each data set, two tamoxifen-only–treated patient cohorts that are ER+, irrespective of nodal status, have uniform gene expression array data available in G-DOC, but differ only by whether they recurred within 5 years or did not? Using G-DOC, users can specify these criteria in a clinical data search form to create two (or more) lists of patients that meet these criteria (Figure 6). These two cohorts frame the question posed above; other clinical considerations that are part of the published data could also have been added to the stratification. Upon saving, both sets of patient lists will be immediately available in the "Saved Lists" section, and they can be revisited at any time.
2. Are there clear molecular signatures that are distinct between these two patient cohorts (recurrent; nonrecurrent)? Selecting the cohorts to compare, the optimal statistical parameters to use, and the experimental data set to be used are all needed to fully configure the analysis that will be run (Figure 7*A*). Output of this analysis is a list of annotated probes, filtered by the input specifications, which differentiate between the patient cohorts in the first (training) data set (Figure 7*B*); output comes in a sortable table. Visualization through an expression heat map generated by a modified Java TreeView [14] is supported, permitting the investigator to easily view the results of his/her search to ensure scientific validity of the separation and, if desired, select a subset of the probes to examine in more detail. The saved list of probes identified in this group comparison analysis can be used as the input variables in a PCA classification test [31]. PCA can be used to determine whether the data are linearly separable in the two-dimensional data space defined by the top two principal components. The list of reporters generated can also be used to probe the validation data set in Sotiriou et al. [30] to explore reproducibility of the results from the training data set.
3. Explore the reporter list to identify genes that are transcription factors and that potentially regulate the effects of tamoxifen treatment. After identification of a probe list in G-DOC that shows efficacy in separating tamoxifen-treated, ER+, node-negative breast cancer patients who had recurrence from those who did not, it is expected that some additional examination of these genes would be undertaken to probe the biologic mechanism

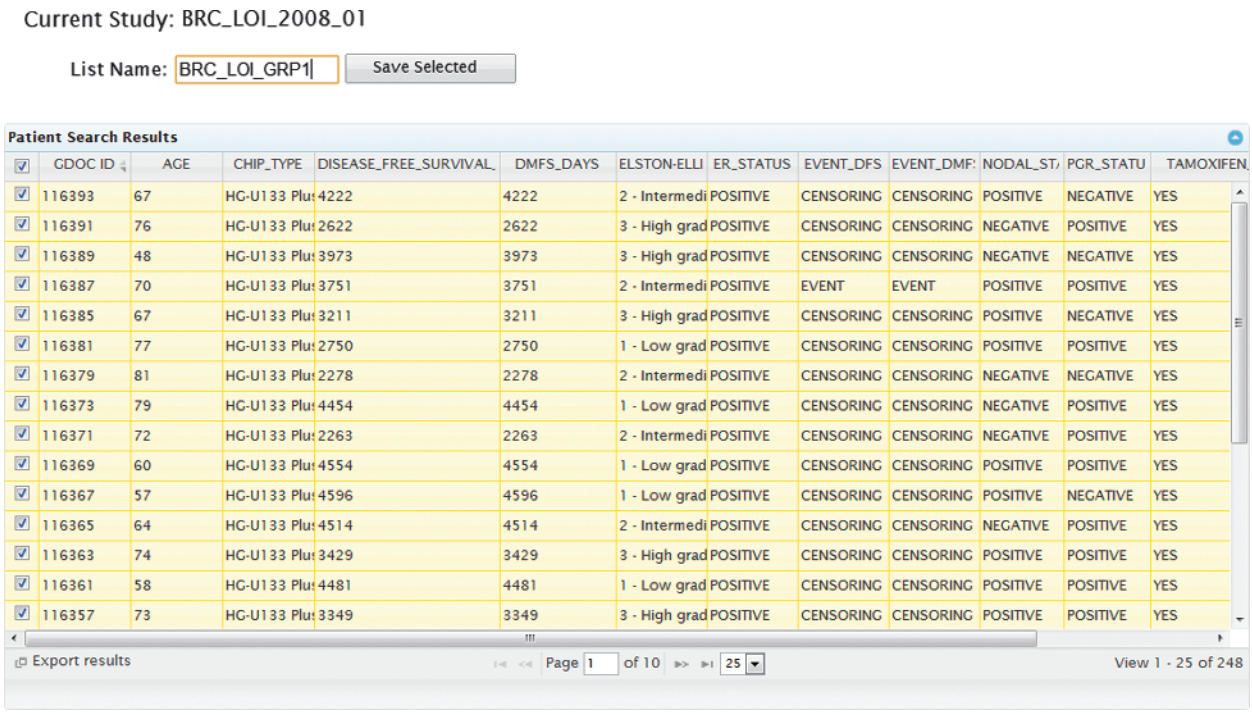

Figure 6. The results from a clinical search are shown in a sortable table in G-DOC. A variety of options for saving and exporting these results are supported.

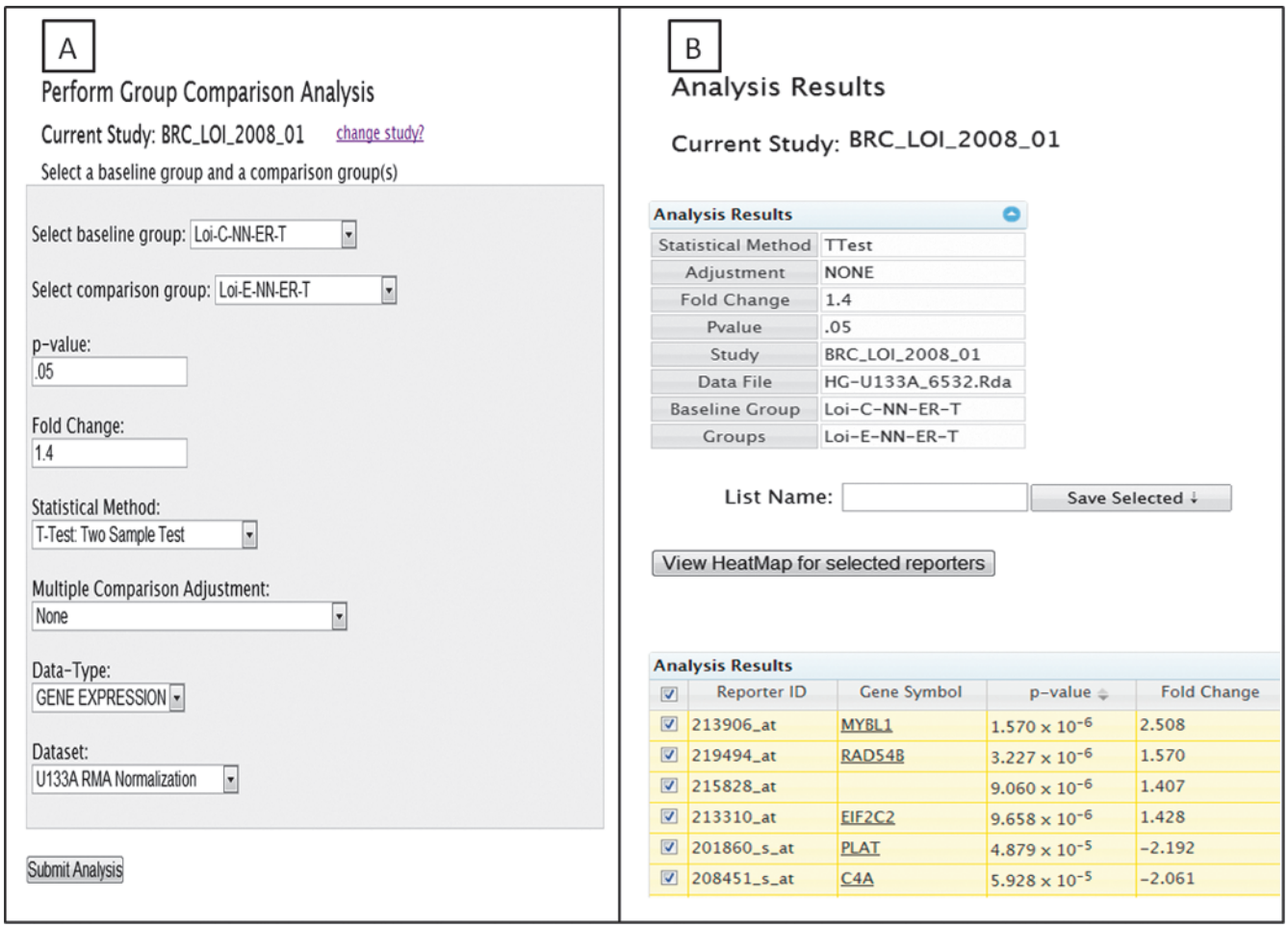

Figure 7. A. It is possible to very specifically configure analyses within G-DOC to generate results to enable scientific discovery or validation of hypotheses. Both basic (e.g., fold change cutoff) and advanced (e.g., multiple hypothesis correction) options are available. B. Results of an analysis in G-DOC are displayed clearly, and a variety of options for saving or exporting are supported within the tool. This is the results page from the comparison of the two cohorts identified from the study of Loi et al. [29].

and validity of any new discovery. For example, the top-scoring gene on the group comparison of Loi et al. [29] (Figure 7*B*), *MYBL1* is a known regulator of transcription [32].

4. Does up-regulation of *MYBL1* impact survival in these patient cohorts? From the reporter list, this gene can be selected to generate a Kaplan-Meier [33] plot in G-DOC based on disease-free survival for these patients. As can be seen in Figure 8, there is a statistically significant ($P < .05$; log-rank test) separation between samples with 1.4-fold upregulated *versus* downregulated *MYBL1*.
5. In the face of this information, can even more detailed analysis of the role of this gene in cancer biology be undertaken through links to external data resources? The entire list of genes identified in this analysis, or similar searches, can be visualized as a whole using the Cytoscape viewer [15] and integrated within G-DOC to retrieve all the Cancer Gene Index data (https://wiki.nci.nih.gov/display/ICR/Cancer+Gene+Index+End+User+Documentation) available for the set of genes, including their degree of interconnectedness. Within the flexible and powerful Cytoscape viewer, it is possible to explore groups of genes and biologic concepts or, conversely, to focus in detail on one gene of interest, such as *MYBL1* (Figure 9). Users can access a number of public annotation databases through G-DOC including the STRING database [34]. The data set in STRING shows affiliations between this top-scoring gene (*MYBL1*) and some others of note, including *BCL2*, a known regulator of apoptosis, whose over expression is caused by *MYBL1*, which is in turn negatively regulated by nucleolin (not shown). Entrez, GeneCards [35], and the Protein Information Resource [36] are a small subset of resources that can be easily and directly accessed from this view (Figure 9).
6. Finally, is it possible to gain a more complete understanding of this gene of interest by exploring its genomic context using the integrated genome browser within G-DOC? In this case, we wish to understand not only the structure of the gene but also the features of adjacent genome regions. A number of these features are available as existing JBrowse [16] (see also Figure 4) data tiers, including OMIM [37] data and SNPs from dbSNP [38]. In this case, we wish to take the analysis further by examining the copy number changes at the *MYBL1* locus for cancer patients. Users can explore the chromosomal region around this gene of interest, which is noted to be a region of chromosomal loss for patients in a third study [39], which has copy number

alteration data already processed and loaded into G-DOC. Another possibility would be to use the CIN index [13] to view the degree of instability between cohorts of patients at the chromosomal or cytoband level, shown here for two cohorts (in this case, metastatic and nonmetastatic, not the clinical parameters used in the study of Loi et al. [29]) from the data set of Sircoulomb et al. [39], most quickly accessed using the Quick Start feature. In this case, *MYBL1*, which resides on 6q13.1, is in a region that does not show a marked difference of instability between metastatic and nonmetastatic patients from the study of Sircoulomb et al. [39].

The storyboard presented here shows that G-DOC can effectively be used to develop and test a scientific hypothesis entirely *in silico*, allowing more resources to be spent on further downstream validation or hypothesis refinement in the laboratory or clinic. The G-DOC Web portal was designed to provide powerful integrated bioinformatics capabilities to the user community, with the hope of advancing biomedical and translational research in oncology.

Discussion

G-DOC and Translational Research

The G-DOC portal was developed as a resource for basic and translational research, and it can greatly speed the process of discovery and validation by providing a powerful platform that supports a wide variety of data analyses and diverse exploration of results. By integrating a

Figure 8. G-DOC supports the generation of Kaplan-Meier plots to help visualize the effect of gene overexpression on patient survival. Note here that the effect of strong overexpression (red) *versus* strong underexpression (blue) of MYBL1, as seen in the data set of Loi et al. [29], shows a statistically significant impact on patient survival. Differences in survival between either of these groups and patients with intermediate expression (yellow line) are not statistically significant.

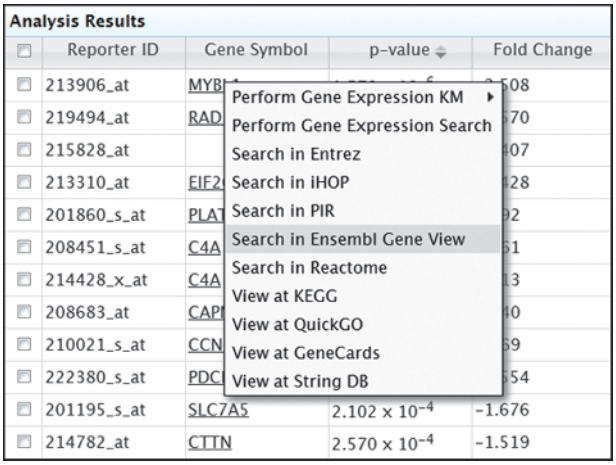

Figure 9. Many links from G-DOC to external resources are supported, enabling investigators to use G-DOC as a central resource for their scientific explorations of public or private data sets.

variety of clinical and "omics" data types, researchers can use the resources and capabilities of G-DOC to more effectively pursue their translational research agenda. As an illustration of the value of systems approaches used in G-DOC, the predictive power and robustness of biomarkers can be significantly increased by integrating transcriptome profiles with interactome data to reveal more relevant functional subnetwork modules [40]. Clearly, transcriptome and proteome analyses of collections of cancer samples combined with functional annotation and modeling of perturbations in molecular pathways and networks have revealed useful biomarkers for the classification and diagnosis of cancer subtypes, the prognosis of patient outcomes, the prediction of treatment responses, and the identification of putative targets for drug discovery [41,42]. G-DOC not only provides a platform to interrogate individual data types but also allows for combination of data from various platforms, such as transciptomics and metabolomics, helping to identify more robust signatures of disease. By supporting easy access to valuable outside resources such as pathway networks and protein-protein interactions, it is hoped that G-DOC can be used as a central hub for discovery and hypothesis generation, as well as validation, in cancer research. Secure exchange of data and analyses within this multi-institutional project team will facilitate closer interactions among researchers and rapid exchange and testing of working hypotheses.

G-DOC and Systems Medicine

More than conventional medicine, systems medicine attracts increasing research interest in the cancer community because it offers a true paradigm shift that may efficiently lead to large, rather than incremental, advances in clinical practice [43]. Importantly, preliminary data show that inexpensive high-throughput "omics" analyses of blood and urine can predict clinical course as well as or better than traditional genomic analyses of tissues (unpublished observations). Integrative and systems medicine platforms such as G-DOC are critical to facilitating the eventual use of "omics" data to drive innovative advances in personalized clinical care and improve the quality and quantity of life for cancer patients. As part of the exploration of this long-term trend, global "omics" profiling studies from a variety of high-throughput technologies are providing comprehensive surveys of molecular changes that are involved in the occurrence and recurrence of many cancers [44]. Combined with an expected concurrent increase in the availability of clinical, pathology, and outcome information from hospital medical center electronic health records systems, data from omics studies are expected to provide an unprecedented opportunity for the advancement of clinical practice. In the near future, physicians will be able to integrate and explore these data sets to understand the heterogeneity of cancers and more efficiently identify diagnostic and prognostic markers. As this paradigm shift becomes more accepted, demand from physician researchers to navigate seamlessly between the phenotypic and genotypic characteristics of a patient, to better tailor their treatment plans, will likewise markedly increase.

Finally, as these large data sets become available for research and to inform clinical practice, it is anticipated that an even bigger challenge will arise—the desire to explore and understand how the cancer genome functions as a complex biologic system in individual patients in relationship to environment, lifestyle, and genetic heritability. New tools will be needed to support these requirements, although many are anticipated to be addressed in future versions of G-DOC.

G-DOC and the Clinical Practice of Systems Medicine

While systems biology will provide the foundation for a practice of systems medicine in the future that will be predictive, personalized, preventive, and participatory [45], it needs to be optimally integrated with health care management systems, imaging centers, and biobanks, as well as subjected to updated ethical regulations, review, and oversight to produce an effective regimen [46]. This will require dedicated efforts of interdisciplinary experts and special attention to clinical practice and education. Platforms such as G-DOC are one part of the systems medicine puzzle to extract knowledge from various types of data and present them to multidisciplinary teams to provide a medium of communication among them.

G-DOC—Comparison with Other Resources

G-DOC was designed and engineered to be a unique resource for translational cancer research that fills critical gaps in the existing research space. It integrates clinical, transcriptomic, metabolomic, and systems-level analysis into a single platform. While Oncomine [47] provides biologist-friendly data mining, the focus of this resource is primarily on cancer transcriptome data, and unlike G-DOC, many of the cancer data sets and features of Oncomine require an annual subscription fee, rather than being freely provided. ArrayExpress [18], the Stanford Microarray Database [48], and the Gene Expression Omnibus [17] repositories have proven to be highly valuable in standardizing and distributing cancer microarray data, but these resources do not well support data analysis, data mining, or systems-level analysis. They also focus primarily on microarray data rather than the range of nonarray "omics" data types, such as mass spectrometry metabolomics data, present in G-DOC. In summary, G-DOC is a unique resource for cancer data and integrative analysis and is currently available freely to the cancer research community.

Future Advances and Needs

The long-term vision for G-DOC involves establishing a robust and comprehensive systems medicine platform that can directly impact health care delivery in the clinical setting by providing more effective clinical decision support. An additional challenge for the future will

be to use decision support tools to improve quality of life even when improved health outcomes are not possible, such as in palliative care [49]. In its current form, G-DOC already provides the means by which an array of existing and emerging "omics" data can be marshaled to improve the outcome for individual patients, and subsequent integration with electronic health records is prioritized for implementation in G-DOC based on its expected impact on both clinical research and clinical practice.

We anticipate that G-DOC and tools like it will soon have a direct impact on human health, although the pace of adoption and utilization within the clinical community is, and will likely remain, the major rate-limiting factor. To effectively use these tools in clinical care, a new type of physician, one with a grasp of modern computational sciences, "omics" technologies and a systems approach to medicine, is a critical component. We expect these subjects, and the tools that underlie them, to enter the medical curriculum to better enable future physicians to effectively use systems based approaches in the clinic.

Availability

G-DOC is freely available to all users at https://gdoc.georgetown.edu. Registration and acceptance of terms of use are required before first login.

Acknowledgments

The authors thank Howard Federoff, Peter Shields, John Marshall, Minetta Liu, Hartmut Juhl, Erin Hedlund, and Michael Vander Hoek for their invaluable guidance throughout the G-DOC development process.

References

[1] Butte AJ (2008). Translational bioinformatics: coming of age. *J Am Med Inform Assoc* **15**(6), 709–714.

[2] Sarkar IN, Butte AJ, Lussier YA, Tarczy-Hornoch P, and Ohno-Machado L (2011). Translational bioinformatics: linking knowledge across biological and clinical realms. *J Am Med Inform Assoc* **18**(4), 354–357.

[3] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**(10), R80.

[4] Calin GA and Croce CM (2006). MicroRNA signatures in human cancers. *Nat Rev Cancer* **6**, 857–866.

[5] Hennessy E and O'Driscoll L (2008). Molecular medicine of microRNAs: structure, function and implications for diabetes. *Expert Rev Mol Med* **10**, e24.

[6] Callis TEE and Wang D-ZZ (2008). Taking microRNAs to heart. *Trends Mol Med* **14**, 254–260.

[7] Uchida S, Nishida A, Hara K, Kamemoto T, Suetsugi M, Fujimoto M, Watanuki T, Wakabayashi Y, Otsuki K, McEwen BS, et al. (2008). Characterization of the vulnerability to repeated stress in Fischer 344 rats: possible involvement of microRNA-mediated down-regulation of the glucocorticoid receptor. *Eur J Neurosci* **27**, 2250–2261.

[8] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, and Enright AJ (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140–D144.

[9] Friedman RC, Farh KK, Burge CB, and Bartel DP (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92–105.

[10] Brase J, Wuttig D, Kuner R, and Sultmann H (2010). Serum microRNAs as non-invasive biomarkers for cancer. *Mol Cancer* **9**, 306.

[11] Huang Z, Huang D, Ni S, Peng Z, Sheng W, and Du X (2010). Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer. *Int J Cancer* **127**(1), 118–126.

[12] Availablet at: http://www.docstoc.com/docs/10956998/The-NIH-Roadmap-to-Understanding-Biological-Pathways-and-Networks-with-Metabolomics. Accessed June 1, 2011.

[13] Kuo K-TT, Guan B, Feng Y, Mao T-LL, Chen X, Jinawath N, Wang Y, Kurman RJ, Shih IeM, and Wang T-LL (2009). Analysis of DNA copy number alterations in ovarian serous tumors identifies new molecular genetic changes in low-grade and high-grade carcinomas. *Cancer Res* **69**(12), 5267.

[14] Saldanha AJ (2004). Java TreeView—extensible visualization of microarray data. *Bioinformatics* **20**(17), 3246–3248.

[15] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, and Ideker T (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**(11), 2498–2504.

[16] Skinner ME, Uzilov AV, Stein LD, Mungall CJ, and Holmes IH (2009). JBrowse: a next-generation genome browser. *Genome Res* **19**(9), 1630–1638.

[17] Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau W-CC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, and Edgar R (2005). NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res* **33**, D562–D566.

[18] Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, et al. (2005). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res* **33**, D553–D555.

[19] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**(4), e15.

[20] Bolstad BM, Irizarry RA, Astrand M, and Speed TP (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**(2), 185–193.

[21] Wright GW and Simon RM (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448–2455.

[22] Iorio MV, Visone R, Di Leva G, Donati V, Petrocca F, Casalini P, Taccioli C, Volinia S, Liu CG, Alder H, et al. (2007). MicroRNA signatures in human ovarian cancer. *Cancer Res* **67**, 8699–8707.

[23] Livak KJ and Schmittgen TD (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(–Delta Delta *C*(T)) method. *Methods* **25**(4), 402–408.

[24] Schmittgen TD, Lee EJJ, Jiang J, Sarkar A, Yang L, Elton TS, and Chen C (2008). Real-time PCR quantification of precursor and mature microRNA. *Methods* **44**(1), 31–38.

[25] Kozomara A and Griffiths-Jones S (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, D152–D157.

[26] Pedrioli PGA, Eng JK, Hubley R, Vogelzang M, Deutsch EW, Raught B, Pratt B, Nilsson E, Angeletti RH, Apweiler R, et al. (2004). A common open representation of mass spectrometry data and its application to proteomics research. *Nat Biotech* **22**, 1459–1466.

[27] Li C and Wong WH (2003). DNA-chip analyzer (dChip). In *The Analysis of Gene Expression Data: Methods and Software.* G Parmigiani, ES Garrett, R Irizarry, and SL Zeger (Eds). Springer, New York, NY. pp. 120–141.

[28] Feng Y, Yu G, Wang T-L, Shih I-M, and Wang Y (2010). Analyzing DNA copy number changes using fused margin regression. *Intl J Funct Inform Pers Med* **3**(1), 3–15.

[29] Loi S, Haibe-Kains B, Desmedt C, Wirapati P, Lallemand F, Tutt AM, Gillet C, Ellis P, Ryder K, Reid JF, et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor–positive breast cancer treated with tamoxifen. *BMC Genomics* **9**, 239.

[30] Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, Nordgren H, Farmer P, Praz V, Haibe-Kains B, et al. (2006). Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* **98**(4), 262–272.

[31] Raychaudhuri S, Stuart JM, and Altman RB (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*, 455–466.

[32] Golay J, Loffarelli L, Luppi M, Castellano M, and Introna M (1994). The human A-myb protein is a strong activator of transcription. *Oncogene* **9**(9), 2469–2479.

[33] Kaplan EL and Meier P (1958). Nonparametric estimation from incomplete observations. *J Am Stat Assoc* **53**, 457–481.

[34] Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**, D561–D568.

[35] Rebhan M, Chalifa-Caspi V, Prilusky J, and Lancet D (1997). GeneCards: integrating information about genes, proteins and diseases. *Trends Genet* **13**(4), 163.

[36] Wu CH, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu ZZ, Ledley RS, Lewis KC, Mewes HW, Orcutt BC, et al. (2002). The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res* **30**(1), 35–37.

[37] Online Mendelian Inheritance in Man (OMIM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). Availablet at: http://www.ncbi.nlm.nih.gov/omim/. Accessed June 1, 2011.

[38] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**(1), 308–311.

[39] Sircoulomb F, Bekhouche I, Finetti P, Adelaide J, Hamida AB, Bonansea J, Raynaud S, Innocenti C, Charafe-Jauffret E, Tarpin C, et al. (2010). Genome profiling of ERBB2-amplified breast cancers. *BMC Cancer* **10**, 539.

[40] Chuang H-YY, Lee E, Liu Y-TT, Lee D, and Ideker T (2007). Network-based classification of breast cancer metastasis. *Mol Syst Biol* **3**, 140.

[41] Reymond MA and Schlegel W (2007). Proteomics in cancer. *Adv Clin Chem* **44**, 103–142.

[42] Chin L and Gray JW (2008). Translating insights from the cancer genome into clinical practice. *Nature* **452**, 553–563.

[43] Roukos DH (2010). Systems medicine: a real approach for future personalized oncology? *Pharmacogenomics* **11**(3), 283–287.

[44] Madhavan S, Zenklusen J-C, Kotliarov Y, Sahni H, Fine HA, and Buetow K (2009). Rembrandt: helping personalized medicine become a reality through integrative translational research. *Mol Cancer Res* **7**(2), 157–167.

[45] Weston AD and Hood L (2004). Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* **3**(2), 179–196.

[46] Auffray C, Chen Z, and Hood L (2009). Systems medicine: the future of medical genomics and healthcare. *Genome Med* **1**(1), 2.

[47] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, and Chinnaiyan AM (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**(1), 1–6.

[48] Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, et al. (2001). The Stanford Microarray Database. *Nucleic Acids Res* **29**(1), 152–155.

[49] Madhavan S, Sander A, Chou W-Y, Shuster A, Boone K, Dente M, Shad A, and Hesse B (2011). Pediatric palliative care in the age of eHealth: opportunities for advances in HIT to improve patient-centered communication. *Am J Prev Med* **40**(5 suppl 2), S208–S216.