National Cancer Institute

and

National Human Genome Research Institute
National Institutes of Health

The Cancer Genome Atlas (TCGA) Data Portal Use Case Workshop

January 10, 2008

Summary Report

Purpose of the Data Portal Meeting

The Cancer Genome Atlas (TCGA) Pilot Project is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). The goal of TCGA Pilot Project is to assess the feasibility of a full-scale effort to identify and catalog the genomic alterations found in all cancers. Three types of cancers: brain (glioblastoma multiforme), lung (squamous carcinoma), and ovarian (serous cystadenocarcinoma) were selected for study in the Pilot Project. The data generated by the Pilot Project are deposited on a regular basis into a Data Coordinating Center and made available to the cancer research community at-large via a TCGA Data Portal. The Data Portal was launched July 2007.

On January 10, 2008, TCGA Project Team hosted a Data Portal workshop to engage experts in gleaning suggestions for new features and future iterations of the Pilot Project Data Portal. The meeting focused on three main topics: 1) Discussing the Pilot Project and implications of cancer genomic information on cancer research, clinical practice, and health outcomes; 2) Highlighting the current external tools available for analysis of TCGA data; and 3) Generating suggestions for future iterations of the Data Portal.

The meeting was opened with a plenary talk by Dr. Bruce Johnson from Dana-Farber/Harvard Cancer Center, who highlighted the usage of genomic changes in cancer treatment. His specific examples of usage of genomic findings in treatment of lung and gastrointestinal tumors highlighted the need to make TCGA datasets available to the cancer research community via the Data Portal and set the stage for the rest of the workshop. During the meeting, TCGA leadership (also referred to as TCGA Project Team) provided an update on the current status of the Pilot Project, discussing the important role of the Data Portal in meeting the goals of TCGA and the advancement of cancer research, prevention, and treatment. TCGA Project Team described the rich TCGA data collection and analysis process from the different TCGA components:

Biospecimen Core Resource, Cancer Genome Characterization Centers, and Genome Sequencing Centers, which deposit data to the Data Coordinating Center for further processing and distribution.

The project team presented their plan for releasing multiple iterations of the portal interface throughout the course of the Pilot Project. The prototype of version 0.5 of the portal, slated for Spring 2008, was presented at the workshop to provide the cancer research community with TCGA data analysis features and capabilities including the ability to correlate data from different platforms and develop clinical outcome reports, view putative mutations, analyze pathways with anomalies, and plot gene expression profiles and Kaplan-Meier (K-M) survival curves.

The prototype presentations that followed focused on how to leverage the caBIG™ infrastructure and existing biomedical data analysis tools. For the purpose of this workshop, investigators worked with data generated from the more than 100 glioblastoma samples that were characterized and in some stage of the sequencing process. About 200 glioblastoma samples will be analyzed by the end of spring 2008. Approximately 600 genes were selected for the first round of glioblastoma multiforme (GBM) tumor sequencing; nearly 700 targets were selected for the second round of GBM tumor sequencing. The target genes selected for the first and second round of sequencing can be found at: http://cancergenome.nih.gov/dataportal/data/types/sequencing/.

Existing Data Analysis Tools

TCGA Project Team recognized that there are several existing data analytical tools that may be leveraged for TCGA. The tools presented to meeting participants included:

GenePattern provides easy access to more than 90 computations and visualization tools for the analysis of gene expression, proteomics and Single Nucleotide Polymorphism (SNP) data (http://www.broad.mit.edu/cancer/software/genepattern/).

The Integrated Genomics Visualizer is being developed at the Broad Institute to display copy number alterations (CNAs) and loss of heterozygosity (LOH) from SNP data. Other data types such as expression, mutation and methylation status are being added to the viewer. A public URL is not available at this time.

Pathway Interaction Database provides information about signaling pathways in human cells. It also provides a set of user-friendly tools to allow the pathways to be explored, visualized, and mined (http://pid.nci.nih.gov/).

Cancer Genome Workbench is a computational platform to view and analyze somatic mutation profiles from tumor samples against the reference human genome sequence to improve the accuracy of mutation identification (http://cgwb.nci.nih.gov).

Genboree is a software system for genomic research. It enables studies of genome variation, including array comparative genomic hybridization (CGH) data, PCR-based resequencing,

genome resequencing using comparative sequence assembly, genome remapping using pairedend tags and sequences, genome annotation, multi-genome comparison and pattern discovery via genome self-comparison (www.genboree.org).

UCSC Genome Browser contains the reference sequence and working draft assemblies for a large collection of genomes. The Genome Browser zooms and scrolls over chromosomes, and has the ability to display genome annotations from external groups (http://genome.ucsc.edu/cgibin/hgGateway).

caIntegrator is a data integration platform that allows researchers to correlate clinical parameters such as outcome with genomic findings (http://caintegrator-info.nci.nih.gov/).

Breakout Sessions

The workshop included concurrent breakout sessions focused on three types of usage scenarios:

- Exploratory Genomic Analysis and Visualization
- Gene- and Pathway-Centric Analysis and Visualization
- Patient-Centric / Clinical Correlative Analysis and Visualization

The groups were charged with brainstorming prospective TCGA Data Portal features, workflows, and "use cases" in each of the three sessions and then presenting them to the participants. Below are some of the major themes and suggestions presented. The feedback shared with TCGA Project Team will be reviewed and considered for future iterations of TCGA Data Portal.

<u>Exploratory Genomic Analysis and Visualization</u>
This group focused on brainstorming functionalities for people who want to perform in-depth exploratory analysis of genomic data sets. The target audience for workflows discussed in this session includes genome researchers, bench scientists, and bioinformaticians. This group developed the following suggestions:

Search capabilities:

- The user should be able to select for certain data types based upon analysis technology platform (i.e., chip-based RNA expression) and download data for a select list of samples (i.e. samples in batch 1 and 2).
- The data should be exportable in tab-delimited file formats, or another spreadsheet friendly format.
- Arbitrary sub-sets of the annotation elements should be selectable.

Visualization capabilities:

- Basic data visualization capabilities should be the first priority, including box plot, colored histograms, scatter plots, and K-M plots.
- The next set of priorities included: projections on pathways and genomes; heatmap with integrated genomic tracks; 3-D visualization of PCA analyses.

Data analysis tools and capabilities:

- Pre-computed results should be presented by the portal. In addition, certain "on-the-fly" data analysis and presentation would also be useful.
- The portal should be flexible and users should be able to integrate data with other sources.
- Output data should be in formats compatible with common analytical tools.
- The portal should link to other analysis tools.

Gene- and Pathway-Centric Analysis and Visualization

This group was charged with developing suggestions that would support users who would come into the portal with a particular gene or pathway of interest and would like to relate TCGA data to specific hypotheses that they are working on in their research. Such users would likely be bench scientists, including animal model biologists; translational researchers, pharmaceutical researchers, and bioinformaticians.

Search capabilities:

- Users would want to be able to initiate data export or analysis with a gene ID, pathway, or region. Alternatively, they may want to select a subject (gene, pathway, etc.) and then select a data type such as gene mutation or expression associated with that subject.
- The user would want to select one or more pathways of interest and browse TCGA data available on genes in those pathways.
- Users would want to group genes by pathway interactions (adjacency); use TCGA data to find 'important' pathways; find other genes with mutations when given a set of samples with a mutation in an interesting gene; browse by pathway name, ID, set of genes and gene products and molecules in pathway.
- A query result might show mutations, gene expression, copy number, and methylation.
- It would be helpful to enable the launch of batch access through an Application Programming Interface (API) published for bioinformaticians.

Visualization features:

- Chromosomal context (tracks). The following visualization capabilities were suggested as being useful: heatmaps, Kaplan-Meier outcome curves, color-coded pathway networks, and display of clinical parameters along with heatmaps of characterization data.
- Develop an entry-to-data "front page" that is a karyotypic view with thumbtacks on gene(s) user has visited. This implies the ability to save user information / sessions in a "shopping cart" type system.
- Pre-computed data for popular searches. For example, a gene page (selecting genes with known disease association) with key findings about those genes from TCGA data, such as information on mutation frequencies, methylation status, RNA expression and copy number abnormality across samples for a gene of interest.

User interface features:

- Data should be annotated to indicate whether or not they have been validated.
- Linking to other databases, such as ENCODE and COSMIC.
- Pathway diagram with overlaid molecular data (e.g. mutation, expression change, CNA,

methylation alteration).

Patient-Centric / Clinical Correlative Analysis and Visualization

This group was charged with identifying use cases and scenarios to correlate clinical information with molecular phenomena. The tools described here would be useful for clinical researchers and physicians.

The participants in this breakout session, primarily physician scientists, defined the primary goal as enabling patient-centered users to generate and test hypotheses by:

- identifying potential genetic markers for disease, which could lead to better diagnostic or prognostic tests.
- revealing possible genetic bases and pathways of cancer, leading to identification of targets for drug therapies.

The group noted that TCGA data cannot be used for identifying diagnostic tests or therapies for use in the clinic; the data can be only used for discovery, validation of research, and clinical hypothesis-generation.

Search capabilities:

- By clinical information such as cancer type, pathology, demography, treatment, progression, and outcome information. Search capability by subtypes of a tumor may also be available.
- The user should be able to search for correlations between clinical parameters and genomic anomalies.

Data analysis tools and capabilities:

- Data analysis tools should enable clustering based on anomaly. Pre-calculated clusters and links to tools that allow clustering capability were also suggested.
- The portal should provide pre-computed views to summarize the patient cohort in the study, including based on demography (age, sex, race, etc), clinical condition (diagnosis, stage, grade, etc.). Such tables would list the numbers of cases in cells predefined by clinical annotations, outcomes or molecular measurements. The cell contents could be hyperlinked to sample/patient lists.

User Interface features:

- Scatter plots to find correlations among clinical/genetic information
- Kaplan-Meier plot of survival rates for different subgroups
- Pathways
- Links for retrieving raw data files

Conclusions

The participants reached a broad consensus that TCGA Data Portal should:

- Provide ability to select for data types and patients of interest for easy bulk download of datasets along with clinical and tissue annotations.
- Enable Google-like search term-based simple user interface to query the data.
- Provide easy access to high-level summary views for pre-computed TCGA data.
- Capture versions of the data. At a minimum, dates of data updates must be shown.
- Provide a gene information page that collects TCGA data for the gene.
- Provide the capability to construct (by query) lists of samples (or patients) and lists of molecular abnormalities. Users would have the option to name and save lists.
- Simple lay person high-level summary views of TCGA data to serve the patient and patient advocate community.
- Implement security features to allow for open and controlled access to TCGA data as defined by TCGA's data release and patient protection policies.

TCGA Project Team will create an informatics tool development plan, based on the input received from this meeting. This development plan will be reviewed during regular calls with the Project Team and several TCGA investigators who contributed to this portal workshop. Over the course of the pilot and during various meetings, the software tools will be demonstrated and user feedback will be solicited. This feedback will be used to iteratively adjust the tool development plan.