National Cancer Institute
and
National Human Genome Research Institute
National Institutes of Health

TCGA Data Release Workshop

May 9-10, 2006

Wyndham Garden Hotel—LaGuardia Airport
East Elmhurst, NY

SUMMARY REPORT

Executive Summary

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) whose goal is to determine the technical feasibility of implementing a full-scale TCGA project that would develop a complete "atlas" of the genomic alterations involved in cancer. To achieve this end, the pilot project will institute a network of centers to analyze three tumor types. The Biospecimen Core Resource (BCR) is the centralized collection center for the tumor and normal tissue samples and any associated clinical annotation data. The BCR also is responsible for checking sample quality and for distributing biomolecules to the Cancer Genome Characterization Centers (CGCCs) where the samples will undergo analyses and to the Genome Sequencing Centers (GSCs). The CGCCs will perform genomic analyses to determine the presence of different genomic alterations and to identify targets that will be sequenced at the GSCs.

The data generated by the CGCCs and GSCs will be deposited into a TCGA database developed through the NCI's cancer Biomedical Informatics Grid (caBIG™). In order to develop release guidelines for these data, the NCI and the NHGRI convened a meeting to bring together experts in gene sequencing, database development, and stakeholders to discuss how to balance the need to make the pilot phase data available to a wide range of stakeholders while protecting the integrity of the data and the privacy of individuals who consent to participate in this effort.

In developing data release guidelines, TCGA is guided by earlier efforts, including the Human Genome Project. The data release principles from the January 2003 Fort Lauderdale meeting entitled *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility* defined the tripartite responsibilities of data producers, data users, and funding organizations. These Fort Lauderdale principles stressed the rapid, pre-publication release of sequence data as a tremendous benefit to the scientific community that should be continued and extended to other large datasets. The meeting also helped to define a "community resource project" as a project with the primary goal of producing a resource for the broad scientific community. TCGA will be such a community resource project.

TCGA will involve an expanded set of stakeholders that, at a minimum, will include data producers, data users, funding organizations, journals, and patients. There will be a continually increasing number of stakeholders as use of TCGA datasets intensifies. Furthermore, since the first phase of TCGA is a pilot phase, there are multiple project areas still under development, which again has the potential to increase the number of types of stakeholders who may require access to TCGA data. To address the potential impacts of an expanding stakeholder pool, further scholarship is needed in the area of bioethics as it

relates to cancer biology and clinical oncology. Most notably, further scholarship is needed in the area of ethics as it relates to cancer biology and release of comprehensive genomic datasets.

The pilot phase of TCGA will reaffirm the notion of rapid, pre-publication data release by a community resource project and will abide by a data release policy consistent with other NHGRI-sponsored medical sequencing projects. Once verified, all TCGA data that cannot be used to identify a patient by current scientific methods will be placed in a publicly accessible, open database. To help reduce the risk of patient identification, one model examined for data placed in the public domain was gene-sized batches, but the exact definition of data release unit is still under examination. All data that can determine patient identity will be placed in a controlled-access database. By establishing this two-tiered data access system (an open database and a controlled-access database), TCGA recognizes the responsibility of undertaking human subjects research and respects the bioethics concerns associated with genomic studies.

For the controlled-access database, access will be granted to users who demonstrate a legitimate scientific interest. A user requesting access to the controlled dataset will be required to receive approval from a TCGA Data Access Committee (DAC). However, in abiding by the principle that a community resource project should make data as fully accessible to the public as possible and in promoting data exchange within the scientific community, the DAC will exercise extreme care when establishing approval standards, with barriers to access set reasonably low. As the pilot project progresses, the risks and benefits of the data release policy in effect will become increasingly clearer. As a result, the data release policy will continue to evolve during the next few years.

The pilot project will use samples that have been donated to be used in research (either retrospectively or prospectively), particularly from patients enrolled in clinical trials or patients treated by a uniform protocol with rich clinical information. Surviving donors will be contacted to obtain their consent so that their samples can be used in this specific genomics project. TCGA's informed consent explains in clear terms to participants the project parameters, risks, and benefits. The form also will explain that, although the strictest measures will be put into effect to protect each patient's identity, all systems inevitably harbor the potential for abuse. However, the project will ensure that the risk for identification remains extremely low.

TCGA data, which will be placed in the public domain, represent prior art and should be unencumbered by intellectual property (IP) protections. Researchers within the scientific community should be able to build upon TCGA datasets and then apply for patents provided that substantial value is added to the data. Advice will need to be obtained from the United States Patent and Trademark Office (USPTO) affirming TCGA data release practices and procedures discouraging parasitic patents on primary data.

A marker paper from TCGA should be published to define the project to the broader scientific community. Additionally, once TCGA data have been placed in the public domain, any publications from use of the data are strongly encouraged to acknowledge TCGA data producers and funding agencies.

Over the course of the two-day workshop, participants reached a broad consensus that:

- TCGA should develop a two-tiered system of access in the spirit of allowing as much open data release as possible while still protecting the identity and privacy of participants.
- The open-access system will be available to everyone over the internet while the controlled-access system will require registration and agreement to follow certain rules.
- Data released by the project should be considered pre-competitive and should not be patentable.
- While adherence to the policies (e.g., IP and publication) of TCGA is not explicitly enforceable – the exception being rules regarding restricted data access, which should be enforceable – social forces are likely to be sufficient for encouraging scientists to behave responsibly.

- The pilot project should be used to determine the scientific and social risks and limitations of the project, especially in relation to the identification of patients and determining which patient populations are or are not willing to participate.
- Scholarship on the issue of data release is needed, and international issues should also be explored.
- The pilot project will be useful in learning about the scientific and social aspects of the larger TCGA Project, in attempting to build public trust, in determining how the broad array of stakeholders will fit into the process, and in influencing the development of policies for other projects related to cancer and other diseases.
- Genetic privacy legislation is urgently needed, but until that legislation is passed, the policies of TCGA will attempt to protect patients.

Introduction

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 ultimate goal of the TCGA Pilot project is to determine the technical feasibility of implementing a full-scale project whose aim would be to develop a complete "atlas" of the genomic alterations involved in cancer. This compendium of changes could accelerate the development of new targeted approaches to diagnose, treat, and prevent cancer that are based on the central feature of cancer, namely, that it is fundamentally a disease of the genome.

Aside from the technical challenges that TCGA Pilot project will address, this unique collaboration between the NCI and NHGRI must also solve issues relating to the storage, safeguard, and release of data generated by TCGA, both in its pilot phase and full-scale implementation. In addition to developing protocols for managing the anticipated large datasets incorporating several types of data – genomic, clinical, and pathological data, among others – TCGA will need to develop data standards for emerging data types, create portals for different communities to access data, and provide a secure network for storing and granting access to confidential data that meet the demands of stakeholders and the Health Insurance Portability and Accountability Act (HIPAA).

TCGA is not the first genome-related project to consider issues relating to data release, and thus this new initiative can build on models developed and used successfully with the genome sequencing efforts. A February 1996 gathering of the leaders of the world's genome sequencing laboratories produced a set of rules that created two important precedents for data sharing. The first of these so-called Bermuda Rules called for all sequencers to pledge to share the results of their work as soon as possible by releasing all sequences longer than 1000 base pairs. The second rule called for all sequencers to deposit these data within 24 hours in the public database GenBank with the goal of preventing "centers from establishing a privileged position in the exploitation and control of human sequence information." Part of the logic behind this second provision was that once gene sequencing data were in the public domain, it would rule out the possibility of patenting those sequences.

The Bermuda Rules were followed by the Ft. Lauderdale Rules, which were developed at a January 2003 meeting of large-scale sequencers. The data release principles from this meeting, which was titled *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility* (www.wellcome.ac.uk/assets/wtd003207.pdf), defined the tripartite responsibilities of data producers, data users, and research funders. These Fort Lauderdale principles stressed the rapid, pre-publication release of sequence data as a tremendous benefit to the scientific community that should be continued and extended to other large dataset types. The Ft. Lauderdale meeting also helped define a "community resource project" as a project with the primary goal of producing a resource for the broad scientific community.

Together, the Bermuda and Ft. Lauderdale rules established the precedent of creating sets of genomic data freely available to all researchers. However, TCGA and other initiatives aimed at merging genomic and

clinical data raise new issues. Unlike previous sequencing efforts that generated sequence data from "anonymous" human samples, TCGA uses samples from specific individuals with associated personal and health-related data. Indeed, the promise of TCGA lies in the associations that researchers will be able to draw between genomic and clinical data. But this close association raises issues of privacy and access that are not of concern when sequences come from anonymous samples. Indeed, one of the major challenges facing TCGA is the development of a policy for genomic and clinical data release that will inform the pilot project while preserving the integrity of clinical trials and preserving patient privacy.

To begin the process of addressing these data-related issues, the NCI and NHGRI held a Data Release Workshop for TCGA on May 9-10 in Queens, New York. The workshop was co-chaired by Robert Waterston, M.D., Ph.D., of the University of Washington, and Richard Schilsky, M.D., of the University of Chicago. Workshop participants included representatives from genomic data production centers, genomic sequencers, bioinformatics data users, journal editors, cancer biologists, clinical researchers, patient advocates, data users from pharmaceutical and biotechnology companies, intellectual property experts, and regulatory agency representatives. The goal of the workshop was for these parties to examine the issues and provide suggestions for development of a data release policy to cover all data types in TCGA. As a starting point, the attendees considered a draft, or "strawman," proposal as a framework for developing a data release policy that the NCI and NHGRI can use to move forward on the pilot phase of this project.

Overview of TCGA Infrastructure and Milestones

Before addressing the strawman proposal in detail, the workshop organizers provided the attendees with an overview of TCGA, including information on plans for clinical sample and data collection and data management. The ultimate goal of TCGA is to characterize the genetic changes that occur in cancer and provide a systematic understanding of the genetic basis of all cancers. Such an understanding should have a transforming effect on the study, diagnosis, treatment, and prevention of cancer by allowing for the stratification of cancers into more homogeneous subclasses and providing new molecular candidate targets for intervention within each.

Two key components of TCGA are caBIG™ (http://cabig.nci.nih.gov) and the Biospecimen Core Resource (BCR) (http://cancergenome.nih.gov/components/hcbcr.asp), both of which build on existing resources. Piloted in Cancer Centers during two years of testing, caBIG™ is a standards-based, open-source, open-access system that uses common software to enable efficient data transfer and storage. With its ability to integrate genomic, proteomic, and animal model data with clinical data, the Grid has the capability to connect all sectors of the cancer research community. The Grid infrastructure currently includes more than 800 individuals and 80 organizations. Through its capability to support electronic medical records and its integration with the U.S. Food and Drug Administration's (FDA) clinical trials reporting systems, the Grid will ultimately support personalized medicine.

The BCR is the centralized collection center for the tumor and normal tissue samples and any associated clinical annotation data. In addition, the BCR is responsible for checking sample quality and for distributing biomolecules to the Cancer Genome Characterization Centers (CGCCs) and the Genome Sequencing Centers (GSCs) where the samples will undergo analyses. The CGCCs will perform genomic analyses – expression profiling, chromosomal copy number alterations, and epigenomic changes – to determine the presence of different genomic alterations; the data will be compiled and analyzed to determine which genomic regions will be sequenced by the GSCs. The data generated by the CGCCs and the GSCs will be deposited into a single TCGA database, Data Coordinating Center (DCC), developed under the leadership of caBIG™.

The BCR will create a network capable of providing tumor samples according to strictly controlled guidelines (http://cancergenome.nih.gov/components/hcbcr_process.asp). In November 2005, the NCI issued a Request for Information (RFI) to identify existing biospecimen collections that may be suitable

for the pilot phase of the project. Responses to the RFI from biorepositories were evaluated on the basis of various criteria, such as biospecimen quality and quantity. TCGA has established a number of primary and secondary criteria for selecting the 500 samples and matching normal samples for each of the three tumor types that will be studied during the pilot project.

The data management aspect of the pilot project faces several challenges, including managing the data pipeline, developing data standards for emerging data types, creating "portals" for different communities to access the data, providing a secure network for confidential data, and encouraging new analytical methods. However, investigators addressing these challenges will be able to leverage the already substantial resources of caBIG™ and other relevant public databases, including the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO), trace archives, and SNP database (dbSNP).

Currently, TCGA is on target to have all components of the pilot project operational by the end of 2006. The three-year TCGA timeline includes the following milestones:

- Completion of genomic analysis of three tumors that will lead to the identification of new genes involved in cancer
- Ability to find and identify specific genomic alterations in genes associated with cancer
- Ability to differentiate tumor subtypes based on genomic alterations
- Sequencing of genomic targets to identify recurrent mutations associated with the cancers studied
- Establishment of a genomics database that researchers and other stakeholders can access
- Ability to translate genomic information into positive clinical outcomes

Background to the Workshop Deliberations

Before considering the strawman proposal in detail, workshop organizers discussed some of the important issues that participants needed to consider in their deliberations. The chief challenge facing the workshop participants was the need to balance the desire to maximize public benefit from enabling wide access to TCGA's database while minimize privacy concerns of patients and their advocates. Meeting participants considered open and controlled models of data access. Types of genomic data that can be considered for open access are those elements that do not identify unique patients, including genomic data such as individual sequence traces, gene expression data, chromosomal copy number alteration data (if SNPs were used for this analysis, then only the interpreted results can be freely accessible), and limited information about tumor type, stage, and cellular heterogeneity. Biospecimens are stripped of their "identity" but are not anonymous in that they are linked to an anonymous sample identifier. The informed consent must provide for this kind of open access.

In contrast, a controlled-access database may contain all clinical data and linkages to genomic data. Access to such information requires that the user specify an intended use and certify that there will be no attempt made to use the data to identify individuals or distribute data to third parties. Approval for access to these data requires consent from a Data Access Committee. Informed consent statements for studies that will release data must include mention of the fact that some data will be in the open-access database and include the attendant risks of identification. For biospecimens already collected, a reconsent is required; model consent forms have been developed for prospective and retrospective studies that will need to be approved by each Institutional Review Board (IRB).

A data release policy for a large-scale genomic study should also include a viable publication policy that recognizes the value of publishing a "marker" paper that outlines the policies and goals of the project. The publication policy must also recognize the tripartite responsibilities of funding agencies, data producers, and data users – all must adhere to the publication etiquette. Also, it should be understood that deposition of data into a public database is not the recognized equivalent of a peer-reviewed publication

for purposes of professional advancement and that data producers have a legitimate interest in publishing large-scale analyses of the data.

Another concern in developing sound data release policies relates to the fact that intellectual property (IP) issues for such a project are complex. Given that the goal is to maximize public benefit, most (if not all) TCGA data should be considered precompetitive and are best placed in the public domain. Therefore, grant applicants must propose an IP policy for approval by TCGA. Users will need to acknowledge a specific IP policy, and in the future, a Declaration of Exceptional Circumstances (DEC) may need to be considered to accommodate provisions of the Bayh-Dole Act.

Finally, for any data release policy to be successful, it must balance the needs of all stakeholders. In the case of TCGA, the list of stakeholders is extensive. For example, as long-time, actively involved stakeholders in the cancer research enterprise, oncologists, cancer patients, and cancer advocates have a strong, vested interest both in participating in TCGA and seeing that the data generated by the program have the biggest impact possible. A panel discussion with leaders of these communities raised several concerns that they wished the workshop attendees to address in their deliberations. Chief among these was the need to explain the importance of TCGA for the future of cancer research and clinical oncology and to provide these communities with frequent updates on the status of the project and the findings it generates. However, care must be taken to ensure that findings are not released prematurely so as to not raise expectations and hopes beyond what is truly warranted. In addition, great care must be taken to ensure that data collected by TCGA are not used to discriminate against those found to be at risk of developing cancer.

Potential stakeholders in an open-source research database and their perspectives are included in Table 1.

Table 1. Stakeholders in an Open-Source Research Database and their Perspectives	
Stakeholder	**Key Perspectives**
Research subject and family	• Cure • Altruism • Autonomy • Economic, social, and psychological risks
Community, population, or group	• Implications for stigmatization/discrimination • Promise of prevention/cure
Public/media	• Trust in scientific process • Safeguards to address risks • Misuse • Focus on advancing cures
Researchers	• Access to data • IP • Publications • Grants • Ensuring professional standards and ethics • Producer/user rights and interests
IRB/OHRP	• Ethical and regulatory concerns • Privacy/confidentiality • Compliance with regulations (Common Rule, HIPAA, FDA)
Information technology infrastructure	• Secure safeguards to address privacy, confidentiality, access, utilization • System QC and QA
Institutions	• Policies for ownership and control of specimens and information • Accountability • Liability
Funding agencies	• National resource • Promotion of custodianship and data sharing policies
Industry partners	• Safeguarding proprietary information • Shareholder accountability • Access and utilization
Legal community	• Ownership and property rights • Liability • Case law

Several Models to Choose From

TCGA is not operating without precedent, however. The prepublication release of worm genomic data in 1991 was the first shift away from the old paradigm of releasing data to public databases only after manuscript submission and acceptance. In 1996, the genomics community endorsed automatic, rapid release of human sequence data by the adoption of the Bermuda Principles. At a meeting in Ft. Lauderdale in 2003, the Bermuda Principles were extended to other types of genomic data. Participants at the Ft. Lauderdale meeting developed the concept of the "community resource project" and acknowledged the tripartite responsibility of data producers, end users, and funding agencies for ensuring open data release. Although rapid, open data release has worked quite well for the genome sequencing community, for other data types, the determination of the best method of data release is in evolution. It is possible that the genomic community resource project model could be applied to TCGA by automatically releasing raw sequence traces into appropriate databases upon generation and releasing other data after validation or as appropriate using the genome browser model to provide the data in an appropriate context.

The Cancer and Leukemia Group B (CALGB) project, which has been running since 1956, also served as a useful example for the ensuing discussions. CALGB is an NCI cooperative group that is run through a central office and involves many labs, university medical centers, community hospitals, and a statistical center; the components are involved in many kinds of research, including large therapeutic clinical trials, and produce many data types. The goal of CALGB is to develop better treatments for cancer, and its data release policies are driven by regulations established by the Health Insurance Portability and Accountability Act (HIPAA), which denotes 18 fields of protected health information (PHI) that cannot be released unless they are deidentified, which entails separating data from all information explicitly linked to a particular individual or that reasonably could be expected to allow individual identification. The availability of other data depends on whether the original informed consent allows for their release. CALGB data users are required to sign a confidentiality form and to obtain Institutional Review Board (IRB) approval, even to obtain deidentified data, which are not covered under HIPAA.

The National Center for Biotechnology Information (NCBI) Genetic Association Information Network (GAIN) Project database is a public-private collaboration between NIH, Pfizer, and Affymetrix. GAIN is performing whole-genome association studies on phenotyped de-identified samples collected from different disease studies. Metadata, including a catalog of studies, protocols, questionnaires, phenotype summaries, genotype summaries, and precomputed associations, will be released to public databases, and a controlled-access database will store, display, and allow the download of phenotype and genotype data. Data users who want access to the controlled-access database will be required to write a paragraph about their proposed research interest and must agree not to identify study participants and not to share data with third parties; they must also acknowledge the project's intellectual property policy. Gaining access to the controlled-access database will be mediated by the Foundation for NIH (FNIH) through a data access committee, and the authentication of data users will be done through the NIH Center for Information Technology (CIT); this process should take a week or less. Gaining access to controlled data will require authorization by the relevant NIH Institute or Center and may involve local IRB and institutional administrative approvals; data access may also be limited based on restrictions in the consent forms used for the original study. Authorized users will be allowed to download data from the NCBI.

Finally, the Sanger Institute Cancer Genome Project (CGP) has sequenced cancer genes in a number of different cancers and will start sequencing even more genes in about 800 cancer cell lines. In the future, somatic mutation data will be released to the Sanger website after curation, and raw sequence traces will be available once a potential data user has entered into a data access agreement that prohibits efforts to identify patients, use data for nonscientific purposes, or give the data to others. Limited phenotypic data will be associated with each sample. Project officials have not yet made a decision about whether IRB approval will be required for users to access the data. Data and resources related to the CGP are available online at www.sanger.ac.uk/genetics/CGP/ and include the Cancer Gene Census (a list of cancer genes),

COSMIC (a catalog of curated somatic mutations and literature mutations in cancer), CGP resequencing studies (somatic mutations from systematic large-scale resequencing of genes in human cancers), the CGP Cancer Cell Line Project (resequencing of known cancer genes and other analyses of human cancer cell lines), and copy number and LOH analyses in cancer cell lines and primary tumors.

Barriers to Open Data Release – A Discussion

Workshop participants were divided into four groups for the first breakout session on identifying barriers to open data release. Their discussions yielded a number of key concepts that would inform later discussions on specific proposals for a data release policy. For example, any policy should operate on the guiding principle of maximizing the immediate release of data while ensuring patient privacy, optimizing intellectual property, and ensuring high data quality. Data release policies may also be based on the central premise that informed consent is the key to protecting the interests of patients and their families, but that consent to participate in TCGA is a blanket consent to cover the fact that much of the future research that uses a biospecimen cannot be predicted. To help alleviate any concerns that might come with such a blanket consent, TCGA must develop creative methods for educating patients about the pilot project and the types of findings that arise from this research.

Barriers that the group identified include:

- While the many stakeholder groups all see the good that can result from TCGA, each group has its own sensitivities that raise potential barriers. Patients, for example, want to be altruistic yet are concerned about privacy and the potential for the denial of health insurance for themselves and their family members. Physician scientists and pathologists want attribution, while all investigators want to maximize their research opportunities. Journal editors, meanwhile, would like to avoid ambiguity so they can determine whether scientists who submit manuscripts have been adhering to the appropriate data release practices.
- Universities and startup companies may have issues with an IP policy that stresses the precompetitive, open-access nature of data generated by TCGA.
- An open-access policy could create significant public relations issues for the NCI and NHGRI.
- This pilot project will enroll patients who have been diagnosed and treated. Many patients will likely sign an open-access consent form without worrying about the long-term ramifications. However, family members may not be protected.
- Open access to germline DNA data could be problematic because these data may be associated with an individual's family in ways that clinical information is not.
- Because genetic information offers a level of granularity that effectively fingerprints a patient, transmissibility of this information to third parties that are not conducting research (e.g., insurance companies, law enforcement) and liability associated with this transmission are concerns.

The Strawman Proposal for Data-Access Policies

Having been thoroughly briefed on TCGA and the concerns of various stakeholder groups, the workshop participants were provided with a strawman data release plan as a starting point for indepth discussions. This strawman proposal included the following provisions:

Data management: All data from TCGA will be managed through a central Data Coordination Center (DCC) that will use caBIG™ infrastructure. For TCGA data, two levels of access are planned.

Open access: Limited genomic data generated for each properly consented sample will be deposited into available public databases with no individual identifying information and no connection to the complete phenotype data, though a limited amount of information about the tumor type, stage, and cellular heterogeneity will be included. Examples of these public databases include the DCC itself, Trace Archive at the National Center for Biotechnology Information (NCBI) for preliminary DNA sequence, data and

Gene Expression Omnibus (GEO) at NCBI for data derived from microarray analyses. If a public database currently does not exist for a particular data type, the DCC will develop it for those data that will be open to public access. All data in the public databases will be linked to an anonymous sample identifier. Selected data on individual samples can only be deposited in the open-access database if the original consent is provided for this possibility (see below).

Controlled access: Clinical data along with linkage to TCGA genomic data will be deposited in a controlled-access section of the DCC. TCGA intends that access to these combined sets of data (a TCGA Project Dataset) will be provided to all investigators who want them for legitimate research purposes (which they must specify), while at the same time ensuring that the data are used in accord with the terms of the informed consent under which the tumor samples were originally obtained. Accordingly, full access will be limited to approved users with the expectation that obtaining such access will not be a roadblock to legitimate users. Approved users are those who, along with their institutions, have agreed to the requirements and terms of access established for the TCGA project discussed below. To ensure that access to the controlled-access database is limited to legitimate uses, a Data Access Committee composed of senior NIH program staff at NCI and NHGRI will review all requests for data access from interested users involved in biomedical research.

The NCI and NHGRI intend that access to TCGA Project Datasets will be granted with the understanding that the data will be used in accord with the following parameters, which will be described in more detail on TCGA's website (http://cancergenome.nih.gov/index.asp):

- Approved users and their institutions will acknowledge responsibility for ensuring that all uses of the data are consistent with federal, state, and local laws and any relevant institutional policies.
- Users agree not to use data for other than approved purposes.
- All approved users will certify that they will not distribute individual data contained within a TCGA Project Dataset in any form to any third party, other than their own research staff who also have agreed to these terms.
- Approved users will certify that they will not attempt to identify the individual participants included within any TCGA Project Dataset.

Informed consent requirements: The rapid, prepublication release of detailed genomic analyses of patient samples along with corresponding clinical data requires that patients be adequately informed of the risk of the projects. The informed consent that will govern TCGA samples will be developed to allow for rapid, prepublication release of data. A model informed consent form has been developed by NCI and NHGRI that allows for genomic data to be released to public databases with coded sample identifiers. More detailed clinical information, with links to the genomic data, would only be available in the controlled-access database. Any retrospective samples for TCGA will require that patients be recontacted and agree to these informed consent guidelines. For prospective samples, only samples from patients who agree to these informed consent guidelines will be considered for TCGA.

Release of genomic data: At the outset, the TCGA project considers it relevant to distinguish between "verification" of data from an experiment, which is understood to mean the reproducibility of the technique used in the experiment, and "data validation," which is understood to refer to confirmation by other, independent methods. Genomic data should be released to open-access public databases as soon as the data are verified, with the recognition that each experimental platform will likely have a different set of standards for verification. The TCGA project will, in consultation with the data producers and the Project's External Consultant Panel, identify a minimal verification standard that will trigger public release of each data type once the project starts. Examples of data verification would be a minimum Phred score for DNA sequencing or multiple replicates of an experiment with good correlation in results for microarray hybridizations. TCGA project also may identify additional levels of validation that will be applied in subsequent analyses of the data or with additional experimentation where appropriate. For instance, if a base change is identified in sequence traces, validation of such data would be to identify this

change by a SNP genotyping method. When possible, estimates of the false positive and false negative rates for the particular experimental approach derived from data validation experiments will be included in the data releases as a measure of data quality.

For some platforms, primary data will be transformed by analyses performed by the data producers into derived data that would be valuable to the cancer biology community. For example, sequence trace data from patient samples can be analyzed to identify the location of specific base changes in the genome. Once verified using an appropriate data standard, this analyzed data should also be released to a public database.

Release of clinical data: Clinical data associated with the tumor samples to be used as part of TCGA pilot project will be collected by the Human Cancer Biospecimen Core Resource and tracked by the DCC. TCGA management team will identify the relevant data fields to be imported into TCGA project. A limited amount of anonymized clinical data will be abstracted from the database and released to the open-access database. Examples of the types of clinical data for which this would be done include the tumor type, stage, and cellular heterogeneity of the sample. Access to coded clinical data will be through the controlled-access database that will be limited to approved users of the data. The controlled-access database will provide links between the clinical data and the genomic data in public databases.

Publication policy: As recommended at the Ft. Lauderdale meeting for a community resource project, TCGA pilot project will publish an initial manuscript, a so-called "marker paper," describing the goals of the project, its data release practices, and the publication policies that it intends to follow. The data users should recognize that the data producers have a legitimate interest in publishing prominent peer-reviewed reports describing and analyzing the resource that they have produced (and that neither the "marker paper" nor data deposits in databases are the equivalent of such publications). Furthermore, the report of the Ft. Lauderdale meeting recognized that deposition in a public database is not equivalent to publication in a peer-reviewed journal. Thus, meeting participants suggested that, until TCGA data are published, users adhere to normal scientific etiquette for the use of unpublished data. Participants also recommended that the data users cite the source of the data (referencing the "marker paper") to acknowledge TCGA project data producers. In addition, if the specimens used for TCGA are from a clinical trial, data users also should acknowledge the trial and trial sponsors. At the end of the pilot project, TCGA plans to publish an analysis of the different TCGA datasets describing the utility of the data and a recommendation for the future of TCGA. Funding agencies, reviewers, and journal editors should help enforce these publication principles.

Intellectual property: The purpose of TCGA policy regarding intellectual policy is to maximize public benefit from data produced by TCGA. It is the view of the project organizers that this goal is achieved best if the data remain publicly accessible without any restrictions. In order to accomplish this intent, successful applicants will need to propose to NCI and NHGRI an IP policy for approval by TCGA. During the pilot project, the NCI and NHGRI will monitor the patenting activity of its grantees; if the data become restricted by IP claims, then an alternative strategy, such as the issuance of a Declaration of Exceptional Circumstances (DEC) allowing the government to retain IP rights in the data, will be pursued in the scaled-up project should it be initiated. Approved users will acknowledge TCGA IP policy, the goal of which is to sustain the public benefit of TCGA, by not pursuing intellectual property protections that would prevent or block access to, or use of, any element of TCGA data, or conclusions drawn directly from those data. If patents are pursued, grantees are expected to implement appropriate licensing policies as described in the NIH's "Best Practices for the Licensing of Genomic Inventions" (http://www.ott.nih.gov/policy/genomic_invention.html).

Strawman Proposal Discussions

Workshop participants then discussed the strawman proposal while considering the obstacles that they had raised for developing a data release policy. Meeting participants agreed with the general provisions of the strawman proposal, albeit with important modifications regarding open access and patient consent. Among the recommendations at the closing session made by the workshop participants are:

- Open access to data should be granted in such a fashion as to guard against abuse and a "test set" of project data should be made available online with open access. However, any data that could identify a patient should be subject to controlled release.
- The informed consent proposed should be modified to clarify that consent implies the permanent online presence of a participant's data. In the case of genotype information, which is shared to some extent in blood relatives, currently that data should be protected.
- The strawman informed consent focuses on acknowledgment rather than agreement, and the tone should be shifted toward an agreement-based document.
- A small group or committee, rather than extensive community input, should be employed to decide about the final specifics of the data access and informed consent policies. However, policies must be clear about biospecimen use to avoid potential requests to withdraw specimens from the study at a later date.
- The open-access data release policy should contain agreements not to identify patients, whereas the controlled-access policy would employ a model similar to that used by GAIN (e.g., agreement with a paragraph that underscores the general goals of research). The main difference between the two approaches is that the controlled-access strategy allows sponsors to identify those researchers who have access to the data.
- The data release policy should be more specific in describing potential uses of the data or biospecimens, be more explicit about the benefits to individual participants, specify the risk of identification, make explicit ownership issues, and provide a recontact policy.
- Three points on which researchers who want to have access to TCGA databases should agree are that: (1) data/biospecimens will be used for research purposes, with "research" defined broadly and not restricted to a preconceived hypothesis, (2) there will be no attempts to identify patients, and (3) data/biospecimens will not be distributed to individuals outside the approved lab and institution.

During the general discussion of the strawman proposal and these recommendations, workshop participants reached a consensus that nonlinked traces and other nonidentifying information should be available in the open-access resource but that links between traces and identifying information should be available only in the controlled-access database. They also agreed that classification of which data types will fall into each category needs further refinement. Participants also advocated being more explicit in the informed consent about the potential risks of involvement in TCGA. However, assuming that patients can fully understand the full implications of the risks of participation in TCGA shortly after diagnosis is unrealistic, and thus protections against misuse of data still need to be in place.

The participants also agreed that the use of a controlled-access database would illustrate that TCGA was making good-faith attempts to protect patients in case a patient is identified from project data. This pilot phase should be used to study the success of the informed consent process and to determine who elects not to participate in the project. The participants noted that while determining who should be allowed access to the controlled-access database will be difficult and it is unlikely that all malevolent users will be blocked, these difficulties are not an excuse to not try to set up a secure database.

Conclusions

Over the course of the two-day workshop, participants reached a broad consensus that:

- TCGA should develop a two-tiered system of access in the spirit of allowing as much open data release as possible while still protecting the identity of patient samples.
- The open-access system will be available to everyone over the internet while the controlled-access system will require registration and agreement to follow certain rules.
- Data released by the project should be considered precompetitive and should not be patentable.
- While adherence to the policies (i.e., IP and publication) of TCGA is not explicitly enforceable – the exception being rules regarding restricted data access, which should be enforceable – social forces are likely to be sufficient for encouraging scientists to behave responsibly.
- The pilot project should be used to determine the scientific and social risks and limitations of the project, especially in relation to the identification of patients and determining which patient populations are or are not willing to participate.
- Scholarship on the issue of data release is needed, and international issues should also be explored.
- The pilot project will be useful in learning about the scientific and social aspects of the larger TCGA project in attempting to build public trust, in determining how the broad array of stakeholders will fit into the process, and in influencing the development of policies for other projects related to cancer and other diseases.
- Genetic privacy protection legislation is urgently needed, but until that legislation is passed, the policies of TCGA will attempt to protect patients.