The Cancer Genome Atlas Pilot Project

NATIONAL CANCER INSTITUTE & NATIONAL HUMAN GENOME RESEARCH INSTITUTE

Human Subjects Protection and Data Access Policies

Summary

The Cancer Genome Atlas (TCGA) Pilot Project is designed to catalog at unprecedented scale genomic variations associated with cancer. The Pilot Project will generate large volumes of detailed genomic data derived from human tumor specimens collected from patient populations also granting access to significant associated clinical information. The information generated will be sufficiently specific to be unique to each individual and, despite de-identification of data, there is an attendant risk of individual re-identification by comparison of TCGA data with other databases. Consequently, careful attention must be paid to ensure the privacy of the donors and the confidentiality of their data.

This document describes a set of policies that the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) have adopted to ensure privacy of patients donating specimens and associated data to the Pilot Project. Three key human subjects protection and data access policies have been adopted by TCGA. The first describes donor protection considerations and conclusions in the context of the "Common Rule" (45-CFR-46, governing human subjects protection in federally funded research) and TCGA policies on informed consent. While leaving the decision about whether research in TCGA-funded institutions or access to TCGA datasets constitutes "human subjects research" to local IRBs, TCGA's policy is that all living donors must be approached for informed consent, during which important concepts pertinent to TCGA will be conveyed, prior to participation. The second summarizes TCGA data content and data flows and how those data are deposited into the Pilot Project's online databases as a background to explaining TCGA data access polices. A key feature is that project databases only house data absent of direct patient identifiers. Two tiers of access are available: a public tier of anonymized aggregated data not attributable to a single subject, and a controlled-access tier with significant associated de-identified clinical data and individually unique molecular information. The third section describes the policies adopted to ensure compliance with the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA).

The policies will be subject to review during the course of TCGA Pilot Project, and may be modified and/or augmented as necessary as lessons are learned and the project receives feedback from the many involved communities.

Introduction

In 2005, the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) initiated a collaboration to pursue a 3-year Pilot Project to determine the feasibility of comprehensively cataloging the genomic alterations associated with a set of human cancers. This Pilot Project is designed to assess the technical feasibility and clinical relevance of conducting such a comprehensive analysis, which includes characterization of DNA copy number changes, rearrangements, transcription profiling, epigenetic modifications, and sequence variation. A suite of genomic analysis platforms will be applied to a common set of molecular analytes obtained from clinically annotated high quality tumor specimens and normal tissue (control). The characterization data, along with recommendations from an expert panel, will be used to identify targets for DNA sequencing in tumor and normal tissues for detection of mutations. Because a common set of donor samples will be used in all analysis platforms, the Pilot Project will verify whether cancer-associated genes and/or genomic regions can be identified by combining diverse information from large-scale genome analyses with tumor biology and clinical data, and whether the genomic analysis and sequencing of materials from notoriously heterogeneous tumor tissues can be achieved in an efficient and cost-effective manner. Furthermore, combining these genomic analyses with tumor biology and clinical data may provide new insights for tumor characterization and identify potential diagnostic markers and therapeutic targets.

Successful completion of the Pilot Project, which will focus on three tumor types (glioblastoma multiforme, squamous cell carcinoma of the lung, and serous cystadenocarcinoma of the ovary), is a prerequisite for an expanded phase that would rapidly and efficiently generate analogous genomic data for all major cancer types. Collectively, genomic and clinical data generated by all the components of the Pilot Project will provide the initial contributions to a comprehensive web-based resource describing the genomic alterations of specific cancer types. This resource will be known as The Cancer Genome Atlas (TCGA).

TCGA comprises a network of institutions, including a Biospecimen Core Resource (BCR) acting as a central collection and processing site for tissues and data donated by patients at numerous clinical sites, NCI-funded Cancer Genome Characterization Centers (CGCCS) and NHGRI-funded Genome Sequencing Centers (GSCS) (collectively: Centers), and a Data Coordinating Center (DCC), which together perform all the tasks necessary to generate high-quality, integrated genomic and clinical information that will be deposited into biomedical research databases. The full extent of TCGA components, from clinical sites, to sample processing, molecular data generation, and data management, integration and access can be reviewed at the Pilot Project website (http://cancergenome.nih.gov).

The nature of TCGA data, including linked cohorts of cancer patients, their comprehensive clinical annotation and extensive genomic data, raises novel human subjects protection issues. TCGA management has, therefore, sought broad input in understanding these concerns and establishing a policy for managing TCGA data in light of scientific, ethical and legal concerns. The NCI and NHGRI convened a workshop in 2006 to examine both general and TCGA-specific issues of broadly releasing large quantities of clinical and genomic data. A summary can be found on TCGA website (http://cancergenome.nih.gov/components/TCGA_101706.pdf).

Background

The information below summarizes considerations and key conclusions in three areas related to the protection of the patients who contributed tissues and data to TCGA: human subjects research and informed consent, data access policies, and HIPAA Regulations. This information is specifically intended to promulgate TCGA policies to participating entities and investigators along with a detailed background as to how the NCI and NHGRI made decisions regarding these important policy issues. The NCI and NHGRI believe this background information is particularly important because: 1) the conclusions were the result of a deliberative process that revealed a range of well-considered opinions rather than absolute consensus and 2) the conclusions and policies are open to change due to the nature of the program and the state of the science.

A difficult aspect to establishing sound TCGA policies was balancing the requirement to protect patients donating tissues and data with the importance to biomedical research of making TCGA clinical and molecular data available to the research community. During this process, the NCI and NHGRI received input from many sources. As may be expected, there was not unanimity of views with regard to many of the specific issues involved.

TCGA policies on patient protection take into account all the input accrued by the NCI and NHGRI, including:

- A large number of national and international subject matter experts.
- Policies established for related programs at those Institutes, such as for the Cancer Genetic Markers of Susceptibility project (CGEMS; http://ocg.cancer.gov/programs/cgems.asp) and the NHGRI Medical Sequencing Program (http://www.genome.gov/15014882), and across NIH such as for the Genetic Association Information Network (GAIN; http://www.fnih.org/GAIN/GAIN_home.shtml).
- Ongoing development of policy regarding Genome-Wide Association studies (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-06-071.html).
- Principal Investigators and managers of large, networked clinical trial groups that include a molecular and translational research component, including several cooperative groups and Specialized Programs of Research Excellence (SPOREs; http://spores.nci.nih.gov/).

TCGA has attempted to harmonize this information to present a consolidated set of policies to the research community. Nevertheless, these policies should not be considered static; the policies are expected to evolve over the course of the Pilot Project. TCGA is designed to learn from all aspects of the project, including not only the complex workflow to generate biological data from clinically annotated tissues, but also from the ethical and legal environment in which it operates. It is expected that this knowledge will inform other similar genome-scale biomedical research efforts in all disease areas as well as TCGA if it moves forward into a full-scale after the Pilot Project.

Part 1: Donor Protections: "Human Subjects" Considerations and Informed Consent

All of the institutions involved in TCGA, whether they are clinical sites that are enrolling donors, collecting and contributing their tissue and data; processing and distributing these materials; or involved in generating genomic data, are recipients of federal funds. Consequently, these sites are subject to 45-CFR-46 (the "Common Rule") governing protection of human research subjects. TCGA review of applicability of these regulations to the project, and its decision on the implementation of an informed consent policy are described below.

"Human subjects" or not

Most interpretations of guidance from the NIH Office for Human Research Protections (OHRP) conclude that research using de-identified coded datasets by an investigator accessing TCGA data does not involve human subjects when certain strictures on the flow of "identifiable private information" are put in place. This conclusion is based on OHRP "Guidance on Research Involving Coded Private Information or Biological Specimens" published on August 10, 2004 which can be found at http://www.hhs.gov/ohrp/humansubjects/guidance/cdebiol.pdf and attached in Appendix 1.

This guidance states:

> *Under the definition of human subject at 45 CFR 46.102(f), obtaining identifiable private information or identifiable specimens for research purposes constitutes human subjects research. Obtaining means receiving or accessing identifiable private information or identifiable specimens for research purposes. OHRP interprets obtaining to include an investigator's use, study, or analysis for research purposes of identifiable private information or identifiable specimens already in the possession of the investigator.*
>
> *In general, OHRP considers private information or specimens to be individually identifiable as defined at 45 CFR 46.102(f) when they can be linked to specific individuals by the investigator(s) either directly or indirectly through coding systems.*
>
> *Conversely, OHRP considers private information or specimens not to be individually identifiable when they cannot be linked to specific individuals by the investigator(s) either directly or indirectly through coding systems. For example, OHRP does not consider research involving only coded private information or specimens to involve human subjects as defined under 45CFR46.102(f) if the following conditions are both met:*
>
> *(1) the private information or specimens were not collected specifically for the currently proposed research project through an interaction or intervention with living individuals; and*
> *(2) the investigator(s) cannot readily ascertain the identity of the individual(s) to whom the coded private information or specimens pertain because, for example:*
>
> > *(a) the key to decipher the code is destroyed before the research begins;*
> > *(b) the investigators and the holder of the key enter into an agreement prohibiting the release of the key to the investigators under any circumstances, until the*

> *individuals are deceased (note that the HHS regulations do not require the IRB to review and approve this agreement);*
> *(c) there are IRB-approved written policies and operating procedures for a repository or data management center that prohibit the release of the key to the investigators under any circumstances, until the individuals are deceased; or*
> *(d) there are other legal requirements prohibiting the release of the key to the investigators, until the individuals are deceased.*

The flow of data in TCGA and the contractual obligations put into place between the clinical sites and downstream components of TCGA meet the tests specified above to prevent "identifiable private information" from being passed to researchers. There are both protocols in place to prevent the transmission of such information from clinical sites, and there are contractual obligations upon entities and affiliated investigators to not attempt to contact or identify subjects or their relatives. Consequently, a strict interpretation of the regulations and OHRP guidance would indicate that TCGA does not constitute human subjects research for investigators generating data as part of TCGA research network nor for those investigators accessing and analyzing TCGA datasets, with the exception of the contributing investigators from the clinical sites where the donors are enrolled.

Nevertheless, a number of subject matter experts thought that TCGA should adhere to a more stringent policy for protection of patients and their relatives than called for in the OHRP guidance. Several reasons were commonly cited, including:

- A belief that patients should be specifically consented for this type of project with largely unspecified future use of the generated data.
- The long-standing precedent that human subjects are involved even when there is de-identified, but linked, clinical information being made broadly available to the research community.
- A hypothetical, but not technically challenging, risk that de-identified high density genotyping or sequence data can be matched against a third party database to effectively re-identify an individual. In such an event, de-identified clinical data could be linked back to a participant risking their privacy and the confidentiality of their information.

These expert conclusions were also based on interpretation of other OHRP guidance, including: "Issues to Consider in the Research Use of Stored Data or Tissues" published November 7, 1997 which can be found at: http://www.hhs.gov/ohrp/humansubjects/guidance/reposit.htm and OHRP Decision Charts of September 24, 2004, which can be found at http://www.hhs.gov/ohrp/humansubjects/guidance/decisioncharts.htm.

Because there is no consensus on this issue, the NCI and NHGRI decided not to adopt a project-level policy but to work with all participating institutions and their IRBs after providing them information on TCGA processes and guidance derived from consultations with numerous experts and stakeholders. TCGA expects investigators and their institutions to consider, based on their own standards of research practice, whether or not research involving the de-identified, coded and potentially re-identifiable information in TCGA datasets meets the definition of "human subjects" or not. The NCI and NHGRI presume that this determination will be made consistent with the institutional policies and in consultation with the local Institutional Review Board.

Even if the local conclusion is that TCGA involves "human subjects," institutions and their review boards should consider whether the proposed research qualifies for exemption #4, quoted below:

> *45 CFR 46.101(b)(4)Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.*

Informed Consent

During the process of establishing TCGA human subjects protection policies, the NCI and NHGRI staff and subject matter experts sought input from diverse constituencies and reviewed dozens of protocols and informed consent documents used by investigators and other groups implementing tissue and clinical data collections for genomic studies. This review led to the critical observation that many current protocols and consent processes for existing studies did not adequately describe modern, high-throughput genetic and genomic studies. Specifically, the reviewed consents did not convey the unprecedented scale of data generated from such genomic studies, nor the risks to privacy and confidentiality when such data can be quickly and widely shared on the internet. These findings led NCI and NHGRI management to initially decide that living donors of tissue specimens and data to TCGA would be consented specifically for TCGA and be provided with specific information about the project, the types of data being generated, and the potential risks to them. It was understood that this policy would necessitate that still-living donors who had in the past contributed samples to existing collections (i.e. retrospective collections) would need to be recontacted. Over the course of the first 20 months of the pilot, NCI/NHGRI received considerable feedback on this policy and also identified several informed consent documents that, while not TCGA specific, did address many of the concerns about a project with this scope of data generation and distribution. During March and April 2008, project management revisited this consent policy and modified the policy as described below.

Revised Informed Consent Policy

Under the revised TCGA consent policy, the NCI/NHGRI will review the informed consent document that was used to collect the specimens and issue a non-binding opinion memo to the contributing PI that describes the degree to which the existing consent document is consistent with the goals and activities of TCGA. Specifically, the memo will review if the informed consent document includes key concepts related to TCGA such as genetic research, broad sharing of biospecimens and clinical data, the possibility of future research use, the use of electronic database with partial public access and the risk of loss of privacy. Principal Investigators (PI) may choose to use this memo as supporting documentation in their application to their local IRB. Ultimately, the local IRB will determine if the existing consent document is sufficient for submission of specimens and data to TCGA.

Samples from deceased individuals

A significant number of the samples and data entering TCGA will be from individuals now deceased. TCGA policies are in accordance with the "common rule" that use of these samples does not constitute human subjects research and that they may be used in the Pilot Project without IRB approval. Participating institutions in TCGA are, of course, subject to their own policies, and TCGA will make available any requested documentation to provide investigators

and their IRBs with sufficient information to make their own determination. The NCI and NHGRI staff will review original consent forms to ensure that there are no specific exclusions, for example, a direct commitment not to use specimens after the patient's death.

Documentation of IRB approval

TCGA policies require that all PIs contributing annotated biospecimens provide documentation to the NCI/NHGRI that their IRBs have either a) approved their participation in the project, or b) do not consider participation to constitute "human subjects research," and therefore do not have purview. In the latter case, this is usually because all the participants are deceased. Contributing site PIs are required to provide a copy of the informed consent text under which the participant was enrolled for the original biospecimen collection. This text is considered documentation of IRB approval for the original collection protocol.

Part 2: TCGA Data Access Policies

The patient protection and data access policies developed for TCGA are designed to balance two important goals: to facilitate investigations of genomic changes related to cancer and, at the same time, to respect and protect the patients whose data and materials have been contributed to TCGA.

The Pilot Project's ultimate goal is to create a database of genomic and phenotypic (i.e. clinical) data that can be used in correlative analyses to support research to alleviate suffering and death from cancer. Thus, TCGA policy is to promote wide dissemination of these data for use by the biomedical research community and to assure their maximum utility. TCGA data will be considered a community resource. To achieve this, the NCI and the NHGRI are committed to the rapid and complete release of TCGA datasets for use by all investigators throughout the global scientific community who, along with their institutions, certify their agreement with TCGA policies. All investigators in TCGA's research network are required to adopt the project's policies on data access, publication, and intellectual property, many of which are specifically designed to address patient protection. TCGA data release goals include full recognition that patients donating to this project expect to have their privacy protected and their data safe-guarded according to the law and to best ethical practice.

Background

Because the data collected and generated by TCGA derive from a complex research network, the following background section of this document lays the groundwork for understanding TCGA data policies by explaining <u>how</u> the data are collected, generated, and stored. Key characteristics of the data that can potentially impact the privacy of patients will be highlighted. After the background explanation, TCGA's data access policies are described.

TCGA data comprise information imported and generated along a multi-step workflow, culminating in clinical and varied molecular datasets housed in multiple databases. Integration of, and access to, the complete datasets occur at The Cancer Genome Atlas Data Coordination Center (DCC) with certain types of molecular data being housed at the National Center for Biotechnology Information (NCBI). Those data which could be aggregated to potentially identify a participant will be managed with additional levels of restriction, including both technical security and a requirement that investigators and institutions accede to the terms of data use and participant protection obligations stated in this policy document.

This section describes the steps by which donors, their clinical data and tissue samples, and the molecular data from those samples will be collected, generated, and deposited into TCGA databases. The steps are outlined because the involvement of multiple institutions and data exchanges between those institutions impact the policies and legal requirements attached to the data. The following figure and workflow summary describe, in general alignment with the sequence of activities, TCGA data generation process.

1. Tissue samples and associated clinical data that meet the requirements of TCGA are identified by a NCI and NHGRI team after extensive evaluations of candidate collections from NCI funded centers. In addition to the technical quality of the materials, a key requirement of the biospecimens is that the ethical and legal stringency of the human subjects protocols under which the collections were established enable clear access to the resource by TCGA. TCGA management makes final decisions about what cancers are chosen for the pilot and which tumor collections would transfer samples and data into TCGA research network. See the project website (http://cancergenome.nih.gov) for an up-to-date list of cancers and contributing sites.

   The following are important characteristics of the material transfer from a contributing site to TCGA network that impact the data access policies:
   - Clinical data associated with specimens are stripped of direct patient identifiers before distribution by the custodian. Specifically, the data received by TCGA's research network is compliant with HIPAA defined "Limited Data Set" and does not include the designated identifiers (see HIPAA compliance section for details).
   - The institution contributing tissue and data will maintain a link between donor IDs and materials transferred to TCGA, so that longitudinal and outcomes data can be associated with the genomic data. This link will not be made available to TCGA project, but will be used to enable the flow of additional patient data, accumulated over time, into TCGA.

2. A Biospecimen Core Resource (BCR) has been established by the Pilot Project, which is a central site using uniform protocols to receive and process all tissues and data. The BCR is the sole TCGA interface to all contributing sites, from which it collects tissue samples and clinical data. Details about the BCR can be viewed at: http://cancergenome.nih.gov/components/hcbcr.asp. Samples from contributing sites are processed into molecular analytes with strict quality controls, and then distributed to the Centers. Clinical data associated with the samples are transformed into standardized structures. The BCR operations include the following data generating, formatting and distribution functions:
   - Pathology review of each received tissue specimen, during which typical surgical pathology data (e.g. tumor stage and grade) are collected and compared to the patient's diagnostic surgical pathology report submitted by the contributing institution. Additionally, information on cellular composition and digital images are captured. These data are captured in a structured electronic format to support inclusion in the project database.
   - Isolation of nucleic acids from the samples with concomitant quality control (QC) to ensure suitability for use by the Centers.
   - Distribution of analyte aliquots to the Centers.

- Formatting of incoming clinical data and locally generated pathology and molecular QC data into data structures compliant with standards from the NCI's cancer Biomedical Informatics Grid (caBIG) project. All data associated with TCGA will use terminologies and Common Data Element structures as maintained by the NCI Center for Bioinformatics in their centralized Enterprise Vocabulary Service (EVS) and cancer Data Standards Registry (caDSR) servers. Visit the caBIG™ website for more information: http://cabig.nci.nih.gov.

The following are important characteristics of the sample and data transfer to and from the BCR that impact the patient protection and data access policies:

- BCR will establish contractual relationships, including Material Transfer Agreements, Data Use Agreements, and warrants of IRB approvals and informed consent to enshrine the ethical, legal and technical requirements established by TCGA policies relating to access and transfer of patient information. This requirement applies to relationships with contributing sites transferring samples and data to the project, and all entities receiving samples and/or data from the BCR.
- BCR will generate a secondary donor/sample (TCGA ID), and maintain a link between this TCGA ID and the ID received from each contributing site. TCGA ID will be the one distributed to the Centers along with the analytes. To be clear, the IDs being discussed here are random numerical IDs that, while linked to donors, are not identifiers themselves. The linkage to actual patient identifiers is only maintained at the contributing site.
- The BCR will transmit only minimal clinical data (histopathology diagnosis, tissue, gender, and approximate age), sample histopathology and molecular QC results, and sample logistical information to TCGA Centers as required to support their data generating operations. Transmitted clinical information will meet the definition of de-identified per HIPAA.
- The BCR will transmit all data to TCGA Data Coordinating Center (DCC) (See below). These data will include clinical data compliant with the HIPAA Limited Data Set specification, and therefore, the BCR will enter into a Data Use Agreement with the DCC.

3. Cancer Genome Characterization Centers (CGCCSs) and Genome Sequencing Centers (GSCS) conduct the DNA- and RNA-based molecular characterizations. The Centers will receive samples and log them into local securely managed material management / LIMS databases. The Centers also will have access to sample logistics and QC data from the BCR, as necessary, and may store local copies of such data for operational support. Center databases will maintain the link between the sample IDs provided by the BCR and the derived data.

   The Centers will conduct a variety of high-throughput comprehensive genome-wide analyses using established technologies. CGCCSs will use gene expression profiling, detection of chromosomal segment copy numbers alterations, and epigenetic changes to reveal the spectrum of genomic changes that exist in human tumors and to identify genomic regions for further study by the Genome Sequencing Centers. The GSCSs will re-sequence loci to identify new variant sites and identify DNA changes present in the tumor samples.

The following are key aspects of Center operations that relate to data access policies developed by TCGA:

- Centers will not receive from the BCR any clinical data covered by the project's HIPAA compliant Data Use Agreements as part of operations for TCGA data generation. To access such data for research purposes, Centers will have to apply, as any other member of the research community, to the DCC according to TCGA data access policies as described below.
- As data are generated by the Centers, they will be deposited in the DCC and/or the National Center for Biotechnology Information (NCBI) according to the rapid data release policies of TCGA. Centers will only distribute data to the DCC and NCBI.

4. TCGA has established a Data Coordinating Center (DCC), managed by the NCI Center for Bioinformatics (NCICB), which link together all data generated by the project into a single integrated resource, including clinical information extracted from medical records and all results from the CGCCSs and GSCSs. The DCC is operated using resources, principles and technology developed in the cancer Biomedical Informatics Grid (caBIG) program.

   To help ensure the protection of patients in a manner consistent with the policies of TCGA, the designers of the DCC database have taken steps to ensure that the database cannot readily be used to identify donors. In no case will the DCC database include direct identifiers such as name, medical record number, address, social security numbers, contact information, or any other HIPAA identifiers excluded under the definition of a Limited Data Set – as noted above, such data are not even collected by TCGA.

   The following are key aspects of the DCC that relate to data access policies developed by TCGA:

   - All access to TCGA data that can be used to identify patient participants by the biomedical research community will be via the DCC from a controlled access database, in accordance with TCGA data access policies. The DCC will implement the database and software applications with security capabilities that apply the policies established by TCGA Data Access Committee.

Policies

As described in Part 1 above, it is technically possible that genomic information (DNA sequence, genotype, etc.) generated in TCGA could lead to identification of an individual if similar specimen data from that person (or blood relative) were obtained from a third-party database and correlated (as could happen in a forensic analysis). There is also a risk of individual identification by computer-based analysis of the clinical data in conjunction with, for example, third-party demographic and healthcare management databases. This potential identification would then publicly link the individual to their clinical information collected by TCGA, and could lead to social risks such as loss of privacy or discrimination.

Although the risk of this occurring is judged to be small at present, the NCI and NHGRI have decided to apply stricter requirements than are currently required by the NIH Office for Human Research Protections (OHRP). (See Part 1 on Human Subjects considerations for more

discussion and policies related to recruitment of donors and the informed consent process.) The data access policies described below encapsulate these requirements. The first set of policies describes limitations to data content and requirements to access that content resident at the DCC. The second set of policies covers a key issue regarding data access across TCGA pipeline, i.e. reaching back from the DCC to link tissue sample IDs at the BCR.

Policy on Access to TCGA Data Managed by DCC

To minimize the risk of patient identification, the NCI and NHGRI established a policy that TCGA data be available from a two-tiered database. The first tier will be publicly accessible and contain only completely anonymized data that cannot be aggregated to generate a dataset unique to an individual. A second tier will contain composite genomic and clinical data that are associated to a unique, but not directly identified, person. Access to this tier will require researchers and their institutions to ascribe to the Data User Certification described below. See Table 1 for a detailed list of data elements and the data access tier within which they will be available.

Center/Dataset	Content	Potential for donor identification	Access Policy
BCR complete set	Detailed Phenotype and Outcome data in the following domains: • TCGA case ID • TCGA Sample ID, slide ID, image ID • Patient demography (DOB, DOD, date of last follow-up) • Clinical history and outcome • Examination (Dx results and dates) • Surgery (Procedure and date) • Tumor Pathology • Treatments (agents, radiation, dose, duration, dates) • Sample data (tissue portions, analyte aliquots, size/volume, QC results)	Yes	DAC approval required
BCR minimal clinical dataset	• TCGA case ID, sample ID, and image ID • Clinical Diagnosis • Histologic Type • Tissue Anatomic Site or Tissue Type • Pathologic Status	No	Open/Public
CGCCS expression	• Gene Expression (raw and normalized)	No	Open/Public
CGCCS methylation	• DNA methylation	No	Open/Public
CGCCS raw SNP	• Raw genotype calls	Yes	DAC approval required
CGCCS raw CGH	• Raw signal of hybridizing oligonucleotides	No	Open/Public
CGCCS summary SNP and CGH	• Genotype frequencies • Computed Copy number • Loss of Heterozygosity	No	Open/Public
GSC mutation data	• Newly discovered somatic variants	No	Open/Public
GSC linking table	• Information that links released data to the BCR complete set of clinical annotations	Yes	DAC approval required

GSC sequence traces	• Trace files with NCBI-required annotations. Traces from the same amplicon (forward-reverse reads) will be identified. Ability to aggregate all traces from a single sample across amplicons, however, will not be supported in the open/public data set.	No	Open/Public

Table 1. Listing of data categories being collected and generated by TCGA project, organized by network source, description of content, consideration that the data type is potentially identifying, and level of access restriction on that data.

Open Access Data Tier

Open access data will be available in public databases, *e.g*. the NCBI Trace repository, the DCC, the caBIG™ web portal, dbSNP, etc. These data types may include:

- TCGA Case identifier, individual sample identifiers (barcodes of analyte aliquots sent to Characterization and Sequencing centers), and image identifiers (pointers to anonymous pathology images used to confirm histology).
- Tumor Anatomic Site or Tissue Type, for example to identify blood as the source of germline DNA.
- Tissue sample histopathology.
- Pathologic Status, to describe the status of a sample as being malignant, metastatic, normal, etc.
- Clinical Diagnosis.
- Gene expression profiles.
- Copy-number aberrations, as long is the experimental approach did not utilize single nucleotide polymorphisms (SNPs) analysis.
- Data summaries such as copy number alternations and loss of heterozygosity by SNP analysis, genotype frequencies for each locus.
- Sequence traces (data output of sequencing) without identifiers to an individual subject.
- Newly discovered germ-line variants at each locus.
- Summaries or aggregations of somatic mutations.

Controlled Access Data Tier

The controlled access data tier will not be freely available to the public, but will be made available to any *bona fide* researcher for the purpose of biomedical research, once the investigator, along with his/her institution, has certified agreement to the statements within TCGA Data Use Certification (DUC). The data types in the controlled access tier will include:

- All the data available in the Open-Access Data tier, and
- Demographic and clinical data up to the level of detail permitted by the contributing site IRB-approved protocol, TCGA data access policies as codified in the DUC, and compliant with a HIPAA Limited Data Set.
- Genome-wide genotypes.
- The information linking all sequence traces to a single (de-identified) patient.

Process for DCC Data Access

Investigators seeking access to TCGA data in the controlled-access database will be asked to complete a Data Access Request (DAR). The submission of the DAR ensures that investigators, along with their institutions, understand the broad goals and policies of TCGA patient protection and have specifically agreed to the requirements and terms of access. Such terms include assurance that the data will be used for "appropriate research" in accord with the definition on TCGA, including any limitations on such use. Specific terms and conditions for access to and use of TCGA datasets by Approved Users can be found in TCGA Data Use Certification (DUC) document.

DARs will be evaluated by a Data Access Committee established by the NCI and NHGRI. It is anticipated that most DARs will be evaluated within two weeks of receipt. Applicants that are approved will become Approved Users, subject to adherence to TCGA policies.

All Approved Users will certify through the DAR process that they will not distribute TCGA controlled-access data in any form to any third parties, other than those of their own research staff who have agreed to the terms of the DAR. Approved User's execution of the DUC, which also incorporates the terms of a HIPAA Data Use Agreement, also obliges them not to attempt to identify or contact individual patients or their relatives. For collaborative projects, any independent investigator from a separate institution involved in the use of TCGA data is required to submit a separate DAR. All Approved Users and their institutions will be required to acknowledge responsibility for ensuring that all uses of the data are consistent with federal, state, and local laws and regulations, and any relevant institutional policies.

Policy on Access to Sample IDs

TCGA has a linked protocol, in that the samples and molecular data generated from them are not anonymized. (The term "anonymized" is used in the technical "human subjects research" sense as defined to mean that all links between the sample and data back to the patient have been irretrievably broken.) Many patients who have contributed samples and data to TCGA are still living, and the medical center at which they were enrolled for tissue banking continues to collect clinical, longitudinal, and outcomes data that can be transmitted to TCGA under this linked protocol.

Contributing sites may be able to leverage TCGA generated data to a great extent to further understand the cancer for which they enrolled donors and contributed tissues to the project. This is possible for two reasons. First, it is not currently possible to transmit the full breadth of donor clinical information that may exist at a contributing site or with a contributing investigator to TCGA. For example, many TCGA donors, after resection of tissues, are placed onto therapeutic trials at these institutions and extensive data are collected. Second, most contributing biorepositories contain "sister" samples, from the same tumor as the one donated to TCGA. The potential scientific value of such additional data collection and focused tissue studies is very high.

To enable this, however, contributing sites would need to access the link between their contributing site sample ID and TCGA ID generated at the BCR in order to link their research biorepository records to the genomic characterization data generated by the Centers. To ensure that the best possible cancer research is supported by TCGA, the project is not categorically

opposed to such linkage, but has established a policy that such data access abides by the following additional requirements:

- Sample ID links between contributing site IDs and TCGA IDs will only be revealed to a contributing site investigator documenting a separate IRB approved protocol to use this information.
- In the event that such IRB approval is presented to TCGA DAC, the protocol will require that the purpose is research only and make clear that the data are not transmitted back to the patient or the patient's medical record.

TCGA Data Access Policy review

All TCGA policies are subject to change as deemed necessary to sustain program principles and priorities, to ensure the highest standards for responsible research conduct, and to be consistent with comparable policies established by the NIH, NCI and NHGRI for other programs.

To that end, the NCI, NHGRI, their advisory boards, and their subject matter experts will continually evaluate the risks and benefits associated with collection, generation and deposition of all TCGA data and will consider modification of these policies accordingly when appropriate.

The NCI and the NHGRI, in consultation with their respective Advisory Boards, will make all final decisions concerning TCGA policies. All changes to policies or procedures will be posted to TCGA website (http://cancergenome.nih.gov).

Part 3: TCGA HIPAA Privacy Rule Compliance

Clinical information collected about TCGA tissue donors that resides at contributing medical centers is Protected Health Information (PHI) as defined by the Health Insurance Portability and Accountability Act (HIPAA). Consequently, the transfer of such data into TCGA is covered by the HIPAA "Privacy Rule" and the project must implement policies to ensure compliance with the regulation.

Background

The contributing clinical sites at which patients are enrolled and clinical data are collected to annotate TCGA biospecimens are "covered entities" under HIPAA. Therefore, that clinical data is subject to the HIPAA privacy rule set in place to protect the confidentiality of patient information. The purpose of this rule is to minimize social risks to patients resulting from non-permitted distribution of their health information. This purpose is achieved by regulating the conditions under which clinical data may be disclosed, and includes various mechanisms ranging from obtaining patient authorization, waiver from an IRB, or limiting the data content so that the data do not specifically identify an individual.

Scientifically, however, TCGA goals are best supported if the project can maximize the breadth of clinical information associated with tumor samples, as TCGA is primarily creating datasets for the purpose of hypothesis generation. Consequently, it is not predictable what clinical data elements will correlate with molecular characteristics and therefore it is preferable to collect the greatest possible amount of clinical information per donor. Nevertheless, the transfer of patient data from contributing sites to TCGA must be fully compliant with HIPAA. HIPAA only applies to clinical data being disclosed by contributing sites and not to samples or molecular data generated from those samples within TCGA.

Therefore the goal of TCGA HIPAA policy is to set up a fully compliant clinical data pipeline that enables the maximal amount of potentially relevant clinical data annotating biospecimens to be transmitted to the project. It is noted that the HIPAA Privacy Rule is a U.S. Federal regulation that can be superseded to greater levels of restriction by state and local laws. TCGA will address this eventuality in the context of sample and data procurement relationships with each contributing site, and necessary additional restrictions will be embedded in the material transfer agreements and data transmission operations with that site.

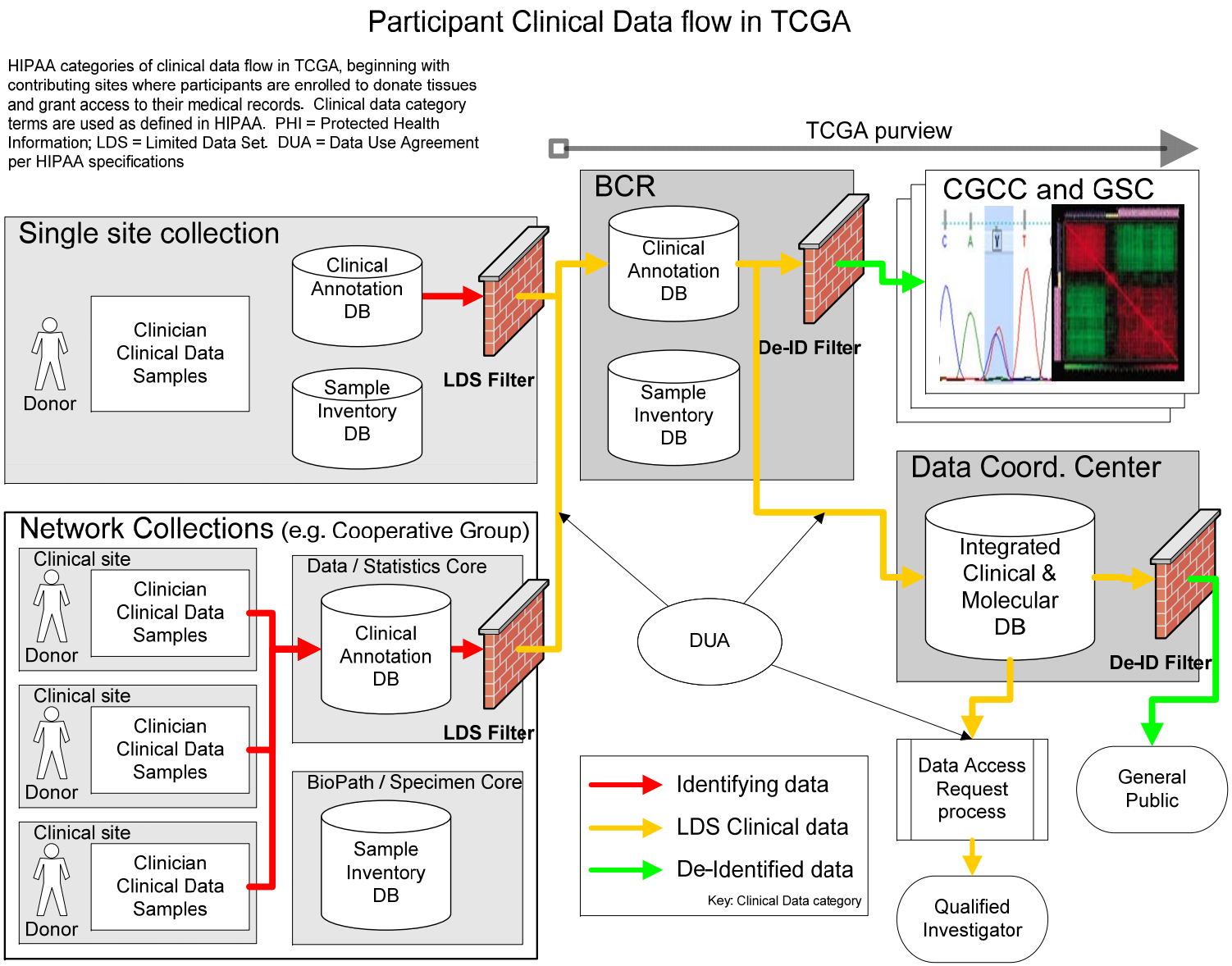

Figure 1: Schematic diagram of permitted clinical data flow in TCGA, in terms of HIPAA-defined categories.

Specific Implementation under Limited Data Set regulations

Of the mechanisms permitted for clinical data disclosure under HIPAA, two were considered for contributing sites' collaboration with TCGA. First, complete HIPAA-defined de-identification, per 164.514 (b)(2)(i), that defines a safe harbor for clinical data that have been stripped of 18 specified data types considered identifying. Clinical data, devoid of these identifiers, are no longer considered "individually identifying" and are therefore not subject to the regulation. Second, distribution of clinical data compliant with the Limited Data Set (LDS) definition at 165.514 (e)(2). The permissible content of LDS compliant clinical data is very similar to HIPAA-defined de-identified clinical data except that more precise date/time and geographic information may be included. The LDS option for permitted disclosures was added to the privacy rule in late 2002, resulting in the so-called "modified privacy rule," after a comment period indicated that the original rule would significantly hamper research. (Specific excerpts from the HIPAA regulations for the two types of data described above are attached in Appendix 1.)

The LDS option was chosen because it achieves the goal of enabling the greatest breadth of phenotypic data to be associated with samples while both protecting patients' identities and requiring minimal addition bureaucratic activity to achieve compliance. TCGA is implementing the necessary policies, contracts, and operations to be compliant with LDS-compliant disclosure and transmission of clinical data from covered entities to the project. See Figure 1 for a data flow diagram illustrating what HIPAA category of data is permitted to be transmitted between TCGA collaborating entities

Key features of LDS Disclosures for Research

The Limited Data Set option for permissible disclosures of clinical data was put in place specifically to support research projects like TCGA. Under HIPAA, key features of the LDS option include:

- LDS-compliant data are still considered Protected Health Information (PHI), so disclosure can be regulated.
- The regulations specify data elements that must be stripped from clinical data to become compliant with the LDS definition. In comparison to the HIPAA definition of de-identification, the list is exactly the same except: (a) date / time information, such as birthdays or procedure dates, may be included; (b) geographical information, at the level of town or city, state, and zipcode, may be included; and (c) the catch-all 18th identifier ("Any other unique identifying number, characteristic, or code") is <u>not</u> included.
- LDS disclosure must be for the purpose of research, public health, or health care.
- LDS may be disclosed <u>without</u> an authorization or an IRB or privacy board waiver of authorization or alteration of authorization. (For documentation, see Appendix 1 for (a) HHS Office of Civil Rights (OCR) guidance; and (b) Page 53231 of HHS Federal Register commentary and response on privacy rule.)
- LDS may be disclosed for research <u>without</u> requirement of HIPAA accounting regulations, under which covered entities must maintain a tracking database of all disclosures. (See Appendix 1 for HHS OCR guidance specific to this subject.)

- The disclosure of LDS imposes a contractual requirement between the discloser and recipients to enter into a Data Use Agreement (DUA). In the DUA, the recipient obliges to:
    - use the data only for intended purposes
    - not attempt to identify or contact individuals
    - further disclose the information only as permitted in the DUA
    - ensure the data are safeguarded
    - impose the DUA restrictions upon any of its agents or contractors

  The requirements under this section begin a "chain" within which all further disclosures by a recipient must be done under additional DUAs already envisioned by the originating DUA.

LDS implementation within TCGA

Operationally, clinical data in TCGA flows unidirectionally from contributing sites (typically HIPAA "covered entities") to the BCR, then from the BCR to the DCC, and finally from the DCC to researchers who have been approved by a TCGA Data Access Committee (DAC). This flow from entity to entity enumerates the set of Data Use Agreements put in place as part of TCGA, and includes the following components:

- A DUA is in place between the contributing entity and BCR contractor. Note, that if the contributing site is itself a "hub" of a network (e.g. as in a cooperative group setting), the DUA in place between the contributing site and the BCR may not be the originating DUA. The DUA with the BCR is in the form of an enhanced Material Transfer Agreement (MTA), since such a contract was already going to be required between any contributing site and the BCR to cover the transfer of tissue samples. NCI worked with several university technology transfer officers to develop an MTA that included a HIPAA compliant DUA. Thus an MTA was drafted that addressed both the physical material and data transfer from a contributing site to the BCR. The MTA includes a warrant by the discloser that all Protected Health Information associated with TCGA tissue donors will be compliant with the Limited Data Set specification. The DUA also pre-authorizes additional data disclosure by the recipient to the DCC for purposes of the project.
- A DUA is in place between BCR contractor and DCC (which is, in fact, the NCI) to permit transfer of LDS compliant data. The BCR's functional role with clinical data is to reformat them into data formats compliant with the NCI's cancer biomedical informatics grid (caBIG) standard.
- A DUA, executed as part of the TCGA Data User Certification (see the section above on Process for DCC Data Access), between the DCC (NCI) and any researcher's employer will be required as one component of the process being implemented by TCGA Data Access Committee to enable access to the combined clinical / molecular data resulting from this project.
- No DUA will be put in place between the BCR and CGCCSs or GSCSs, as the Centers will only receive minimal data associated with the molecular analytes being received. Those data will comply with being completely de-identified per the HIPAA definition.

Important Note

It must be emphatically noted that, while TCGA HIPAA compliance policy is designed to enable the legal maximum amount of data to flow into TCGA, other policies can reduce that data content. Specifically, contributing site IRBs are still free to place greater restrictions on the breadth of patient data distributed with tissue samples. Furthermore, TCGA policies, as promulgated by the project's management, may also place higher restrictions on patient data collected by the project or on what data resident in the DCC are eventually accessible to researchers. Such policies, possibly based upon other bioethics considerations and future risk-to-patient assessments, are described elsewhere in this document.

Appendix 1 – Attachments

TCGA Data Use Certification	
TCGA Data Release Policy	
TCGA Publication Policy	
45 CFR 164.514 (b): HIPAA De-identification Safe Harbor	HIPAA de-id 164.514-a+b.pdf
45 CFR 164.514(e): HIPAA Limited Data Set Specification	HIPAA LDS 164.514-e.pdf
OCR Guidance on HIPAA about Limited Data Set Disclosures	HIPAA OCR guidance 2003-04-03.pdf
HHS Commentary and Response in Federal Register on LDS disclosure as part of privacy rule modifications	HIPAA HHS privacy rule.pdf
OHRP Guidance on Research Involving Coded Private Information or Biological Specimens	cdebiol.pdf