# The Cancer Genome Atlas Program

**NATIONAL CANCER INSTITUTE & NATIONAL HUMAN GENOME RESEARCH INSTITUTE**

## Human Subjects Protection and Data Access Policies

## Summary

The Cancer Genome Atlas (TCGA) Program is designed to catalog, at an unprecedented scale, genomic variations associated with cancer. TCGA is generating large volumes of detailed genomic data derived from human tumor specimens. The genomic information is combined with newly collected and/or existing clinical information gathered from many different patient populations. The genomic and clinical information is organized into two categories: one that is openly accessible to the public and one that has controlled access, available only to qualified researchers obligated to secure the data. The open access data set contains information that does not pose a risk of patient re-identification. The controlled access data set, in which data have been stripped of names, addresses, birth dates and other traditional identifiers, nevertheless contains information that could carry a small risk for re-identification by comparing TCGA data with information in other databases. Ensuring, to the extent possible, the privacy of specimen donors and confidentiality of their data, while promoting and encouraging impactful scientific discovery, has been a paramount concern to TCGA.

This document describes a set of policies that the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) have adopted to address the protection of privacy of participants donating specimens and associated data to TCGA. Three key human subjects protection and data access policies have been developed and implemented by TCGA. The first policy describes participant protection considerations and conclusions in the context of the "Common Rule" (45-CFR-46, governing human subjects protections in federally funded research) and TCGA policies on informed consent. TCGA values a thorough and understandable informed consent process, which includes a comprehensive discussion about the protocol, benefits, risks, etc., with each participant. TCGA management staff (Project Team) has developed documents that can serve as guides to assist investigators in developing their own protocols, and consent forms and processes. This policy leaves the responsibility for the ultimate decision about whether the research conducted under TCGA involves "human subjects" or not to local Institutional Review Boards. The second policy summarizes what information is included in TCGA datasets, how those data are deposited into the program's online databases and TCGA data access mechanisms that are in place. The third policy describes what is being done to ensure compliance with the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA).

These policies have been reviewed throughout the course of the TCGA Program and have been modified as experience is gained and lessons are learned, when the Project Team receives suggestions for clarification or improvements from the many involved researcher or participant communities and as underlying regulations or policies are modified.

# Introduction

In 2005, the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI) initiated a collaboration to pursue a Pilot Project to determine the feasibility of comprehensively cataloging the genomic alterations associated with human cancer. Three cancers, involving the brain (glioblastoma multiforme), the lung (squamous cell carcinoma of the lung), and the ovaries (serous cystadenocarcinoma of the ovary), were selected for study. The Pilot Project assessed the technical feasibility and potential value of conducting a comprehensive genomic analysis of selected tumors, including the characterization of DNA copy number changes, rearrangements, transcriptional profiling, epigenetic modifications, and sequence variation. A number of genomic analysis platforms were applied to a common set of molecular analytes obtained from clinically annotated, high quality, tumor tissues case-matched with a source of germline DNA for comparison. The genomic characterization data, along with recommendations from an expert panel, were used to identify targets for DNA sequencing in tumor and normal tissues for detection of variations. Because a common set of donor samples was used in all platforms, the Pilot Project was able to verify that cancer-associated genes and/or genomic regions can be identified by combining research results from large-scale genomic analyses with tumor biology and clinical data; and that the genomic characterization and DNA sequencing of tissue samples isolated from heterogeneous tumor specimens can be achieved in an efficient and cost-effective manner (TCGA Research Network, 2008). Furthermore, combining genomic analyses with tumor biology and clinical data provided new insights into the biology of tumors and resulted in the identification of potential diagnostic markers and therapeutic targets for selected cancers.

The Pilot Project was judged to be successful and, as such, TCGA was expanded in 2009 to study over 20 additional cancers. The goal for the expansion of TCGA is to rapidly and efficiently generate analogous genomic and clinical data for many major cancer types. As in the TCGA Pilot, genomic and clinical data generated by all the components of TCGA Program are deposited into web-based databases (with both open and controlled access). More information on TCGA data sets can be found at the TCGA Data Coordinating Center (DCC) (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) and the Cancer Genomics Hub (CGHub) (https://cghub.ucsc.edu/).

TCGA data are generated by a wide network of researchers from many institutions that provide comprehensive integrated data and analysis. There are five core functions performed by members of the TCGA Network: 1) clinical site patient enrollment, data collection, and tissue collection; 2) centralized data standardization and sample processing; 3) genomic characterization and DNA sequencing; 4) data collection, management and distribution and 5) data analysis. The specific names of these components for TCGA are described in the table below. Note that the table describes generic institutional roles in TCGA; many institutions participate in more than one role.

| Entity | Function | Name in Program |
|---|---|---|
| Clinical sites | Patient consent and enrollment.<br><br>Sample collection and QC. | Tissue Source Sites (TSS) |

| | | |
|---|---|---|
| | Clinical data entry | |
| Centralized biospecimen and clinical data processing center | Sample pathology and QC.<br><br>RNA and DNA isolation.<br><br>Clinical data standardization.<br><br>Analyte distribution. | Biospecimen Core Resource (BCR) |
| Molecular characterization centers | Genomic characterization of single nucleotide variants, DNA copy number changes and rearrangements.<br><br>mRNA and miRNA transcription profiling.<br><br>Epigenetic modifications.<br><br>DNA sequencing. | Genome Characterization Centers (GCC)<br><br>Genome Sequencing Centers (GSC) |
| Centralized data storage and redistribution centers | Primary sequence data repository.<br><br>Genomic characterization data and analysis results repository. | Cancer Genomics Hub (CGHub)<br><br>Data Coordinating Center (DCC) |
| Bioinformatic analysis centers | Cross-platform and cross-cancer integrated dataset analyses | Genome Data Analysis Centers (GDAC) |

*For further information about TCGA components, please visit the TCGA Program website (http://cancergenome.nih.gov).

The nature of TCGA data, including linked cohorts of cancer patients, their clinical annotation and extensive genomic data published on networked databases, raised novel human subjects protection issues when the program was initiated in 2005. These issues are now more broadly discussed throughout the genomics community, but many of TCGA's policies serve as potential precedent and models. TCGA management has continually sought broad input in understanding these emerging issues and establishing policies for managing TCGA data in light of complex scientific, ethical, legal and societal concerns. As a first step, the NCI and NHGRI convened a workshop in 2006 to examine both general and TCGA-specific issues of broadly releasing large quantities of coded, but linked clinical and genomic data. A summary of this initial workshop can be found on the TCGA website: (http://cancergenome.nih.gov/PublishedContent/Files/pdfs/6.5.1.2_TCGA_DataRelease.Workshop_101706.pdf). Since this workshop, NIH Institutes have held numerous forums on participant protection in genomic studies and the knowledge gained from these forums have continued to inform TCGA policy.

# Background

A difficult aspect to establishing sound TCGA policies has been balancing the requirement to protect participants donating tissues and data with the importance to biomedical research of making TCGA clinical and molecular data available to a broad research community. The information below summarizes the issues discussed and key conclusions in three areas related to the protection of the research participants who contributed tissues and data to TCGA: informed consent, data access policies, and HIPAA regulations. This information is specifically intended to communicate TCGA policies to the research investigators, their institutional officials, and the public; and describe the rationale behind decisions of NCI and NHGRI staff regarding these important policy issues. The NCI and NHGRI are providing this background information to convey that the conclusions were the result of an extensive deliberative process that revealed a range of well-considered opinions; and that the conclusions and policies are open to modification due to the exploratory nature of the program and the transforming state of the science.

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 participant 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/resources/genome-wide-association-studies-cancer-genetic-markers-susceptibility-cgems-initiative) 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/work/past-programs/genetic-association-information-network-gain).
- Ongoing development of policy regarding Genome-Wide Association Studies (http://grants.nih.gov/grants/gwas/).
- 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 (SPORE) and their successors (http://trp.cancer.gov/).

  (The above links are active as of January 16, 2014, but may change.)

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 program. TCGA is designed to learn from all aspects of the program, 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 because of this thorough process the TCGA policies can serve as models for other genome-scale biomedical research efforts in all disease areas.

# Part 1: Donor Protections: "Human Subjects" Considerations and the Policy on Informed Consent

Almost 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. (There are several collaborators at foreign sites, including tissues source sites who abide by HIPAA requirement even if not legally obligated to do so.) Consequently, these sites are subject to 45-CFR-46 (the "Common Rule") governing protection of human research subjects. TCGA Project Team review of applicability of these regulations to the program, 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 October 16, 2008 which can be found at http://www.hhs.gov/ohrp/policy/cdebiol.html and attached in the Appendix. 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 identifiable private information or identifiable specimens includes, but is not limited to:*

1. *using, studying, or analyzing for research purposes identifiable private information or identifiable specimens that have been provided to investigators from any source; and*
2. *using, studying, or analyzing for research purposes identifiable private information or identifiable specimens that were 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 45 CFR 46.102(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:*

1. *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);*
2. *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*
3. *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 contractual obligations with sites contributing specimens and data 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 contractual obligations upon entities and affiliated investigators to not attempt to contact or identify subjects or their relatives (see the section on HIPAA-compliant Data Use Agreements, below). 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 participants and their relatives than called for in the OHRP guidance. Several reasons were commonly cited, including:

- A belief that participants 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 technically possible, risk that de-identified high density genotyping, sequence or clinical 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/policy/reposit.html and OHRP Decision Charts of September 24, 2004, which can be found at http://www.hhs.gov/ohrp/policy/checklists/decisioncharts.html.

The lack of consensus led the NCI and NHGRI not to adopt a program-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 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 IRB.

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 Project Team 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 existing protocols and consent processes in 2005 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. It is important to note, however, that over the course of TCGA, protocols and consents cognizant of these issues have become more globally adopted.

### Initial Informed Consent Policy

These findings led the Project Team 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 program, 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 re-contacted and re-consented. Over the course of the Pilot Project, the Project Team received considerable feedback on this policy and began to review relevant informed consent permissions that, while not TCGA specific, did address many of the concerns about a project with this scope of data generation and distribution. The Project Team revisited this consent policy and modified the policy as described below.

### Revised Informed Consent Policy

Under the revised TCGA consent policy, re-consent of still-living participants is no longer a program-imposed requirement. The Project Team has developed a memo describing best practices for informed consent for participating in TCGA. This document is on the TCGA web site at http://cancergenome.nih.gov/abouttcga/policies/informedconsent. Upon request, TCGA staff will review informed consents from interested investigators, and issue a non-binding opinion memo to the contributing Principal Investigator (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 via the internet, the possibility of future research use, the use of electronic database with partial public access and the risk of loss of privacy. The memo will also note if a

component of the reviewed consent is specifically incompatible with TCGA. A PI may choose to use this memo as supporting documentation in an application to the 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 are 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 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.

### Documentation of IRB approval

TCGA policies require that all PIs contributing annotated biospecimens provide documentation to the Biospecimen Core Resource and the Project Team that their IRBs have either a) approved the use of data and specimens for TCGA studies, or b) do not consider participation to constitute "human subjects research," and therefore do not have purview. Specifically, NCI policy requires contributing site PIs funded by NCI to provide a copy of the protocol, or letter from their IRB, that specifically mentions TCGA in the approval for use of the samples and data. NHGRI policy differs slightly in that a PI's TCGA tissue provision protocol, or approval letter from the IRB, need not specifically mention TCGA and will be reviewed and approved by an NHGRI program staff member. NGHRI approval is contingent upon sample and data use in genetic/genomic studies with wide data sharing and without data use limitations.

## Part 2: TCGA Data Access Policies

The participant 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 participants whose data and materials have been contributed to TCGA.

The Program'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 are 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 program's policies on data access, publication, and intellectual property, many of which are specifically designed to address participant protection. TCGA data release goals include full recognition that participants donating to this program 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 lays the groundwork for understanding TCGA data policies by explaining how the data are collected, generated, and stored. Key characteristics of the data that can potentially affect the privacy of participants will be highlighted. After the background explanation, TCGA's data access policies are described.

TCGA data sets are comprised of information imported and generated along a multi-step workflow, culminating in clinical and molecular datasets housed in two central databases developed and maintained by TCGA: the TCGA Data Coordination Center (DCC) and the primary sequence data repository at the Cancer Genomics Hub (CGHub). (TCGA data may also be housed at other databases developed by collaborators). Those data which could be analyzed to theoretically 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 participants are enrolled, and their clinical data and tissue samples, and the molecular data from those samples are 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 TCGA data generation process.

1. Retrospective collections and potential prospective collections that can provide tissue samples and associated clinical data that meet the requirements of TCGA are identified by NCI. In addition to the biological 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 staff makes final decisions about which cancers are chosen for the program and which institutions are engaged to participate in the program and transfer samples and data into the TCGA research network. See the DCC website (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp ) for an up-to-date list of cancers and contributing sites.

   The following are important characteristics of the material transfer from contributing sites (Tissue Source Sites) to TCGA network that are imparted to the data access policies:

   - Clinical data associated with specimens are stripped of direct patient identifiers before distribution by the contributing site. Specifically, the data received by the TCGA research network are compliant with HIPAA defined "Limited Data Set" and do not include the designated identifiers (see HIPAA compliance section for details).

   - The tissue source site 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 the TCGA program, but will be used to enable the flow of additional participant data, accumulated over time, into TCGA.

2. A Biospecimen Core Resource (BCR) was established by the Program. A BCR is a central site using uniform protocols to receive and process all tissues and clinical data. The BCR is the TCGA interface to all Tissue Source Sites (TSS), from which it collects tissue samples and clinical data. Details about the BCR can be viewed at: http://cancergenome.nih.gov/abouttcga/overview/howitworks/bcr.

   The BCR operations include the following biospecimen processing and 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 participant'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 program database.
   - Isolation of nucleic acid analytes from the samples with concomitant quality control (QC) to ensure suitability for use by the Centers. Nucleic acids are isolated from the samples using strict quality controls, and then distributed to the Centers.
   - Distribution of analyte aliquots to the Centers.
   - Formatting and standardization of incoming clinical data and locally generated pathology and molecular QC data into data structures compliant with standards from the NCI's data. 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.

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

   - BCR establishes contractual relationships, including Material Transfer Agreements, Data Use Agreements, and warrants of compliance with relevant Common Rule and HIPAA regulations to protect the ethical, legal and technical requirements established by TCGA policies relating to access and transfer of participant information. These requirements apply to relationships with contributing sites transferring samples and data to the program, and all entities receiving samples and/or data from the BCR.
   - BCR generates a secondary donor/sample identifier (TCGA ID), and maintains a link between this TCGA ID and the ID received from each contributing site. It is important to note that TCGA is a second link in the ID reference, thus the TCGA ID is effectively twice-removed from the patient's primary ID. The TCGA ID will be the one distributed to the Centers along with the analytes.

- The BCR will transmit only minimal clinical data (for example, diagnosis, tissue, gender, and approximate age), sample histopathology and molecular QC results, and sample logistical information to TCGA molecular characterization Centers as necessary to support their molecular characterization operations. Such clinical information transmitted to Centers will meet the definition of De-Identified per HIPAA.
- The BCR will transmit clinical data to TCGA Data Coordinating Center (DCC) (See below). These data are compliant with HIPAA Limited Data Set specification. The BCR has executed Data Use Agreements with the DCC. (See the section on HIPAA compliance, below.)

3. Genome Characterization Centers (GCC) and Genome Sequencing Centers (GSC) conduct the DNA, RNA, and protein-based molecular characterizations. The Centers will receive samples and log them into locally 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 TCGA 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. The following are key aspects of Center operations that relate to data access policies developed by TCGA:

   - Centers will <u>not</u> receive directly from the BCR any clinical data covered by the program's HIPAA compliant Data Use Agreements as part of operations for TCGA data generation. Center investigators who wish to retrieve the clinical data must do so directly from the DCC.
   - As data are generated by the Centers, they will be deposited in the DCC and/or CGHub according to the rapid data release policies of TCGA. Centers will only distribute data to the DCC and CGHub.

4. TCGA has established a Data Coordinating Center (DCC) which links together all data generated by the program into a single integrated resource, with links made to the primary sequence data managed by the University of California, Santa Cruz (UCSC) Cancer Genomics Hub (CGHub). UCSC has obtained trusted partner status with the National Institutes of Health for providing storage, integration and dissemination of protected TCGA data and other cancer genomics data.

   To help ensure the protection of participants in a manner consistent with the policies of TCGA, the DCC and CGHub database staff has taken steps to ensure that the database cannot readily be used to identify donors. The DCC or CGHub database do not receive any 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 collected by TCGA.

   Furthermore, all access to TCGA data that are individually genetically unique and pose a theoretical risk of participant re-identification may only be accessed via the DCC or CGHub,

in accordance with TCGA restricted-tier data access policies. These data resources will implement the database and software applications with security capabilities that apply the policies established by the TCGA Project Team and Data Access Committee, requiring user authentication against the NIH database of approved investigators.

## 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 data from that person (or blood relative) were obtained from a third-party database and correlated (as could happen in a forensic analysis). There also is 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 link the individual to their clinical information collected by TCGA, and could lead to loss of privacy.

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 encapsulates these requirements. The first set of policies describes limitations to data content and requirements to access that content resident at the DCC and CGHub. The second set of policies covers a key issue regarding data access across TCGA pipeline, i.e. reaching back from the DCC and CGHub to link tissue sample IDs at the BCR.

### Policy on Access to TCGA Data Managed by DCC

To minimize the risk of participant identification, the TCGA Project Team established a policy that TCGA data be made available from a two-tiered data access system. The first tier will be publicly accessible and contain only data that cannot be analyzed 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.

Open-Access Data Tier: Open-access data will be available in public databases. These data types include:

- TCGA Case identifier, individual sample identifiers (barcodes of analyte aliquots sent to Characterization and Sequencing centers), and image identifiers (pointers to pathology images used to confirm histology).
- De-Identified clinical and demographic data.
- Tissue sample histopathology, including de-identified pathology reports.
- Gene expression profiles.
- Copy-number aberrations, as long as 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.
- Summaries or aggregations of germ-line variants.
- Somatic mutations.
- Logistics and QC data.
- Microsatellite instability (MSI) locus summary calls.

Controlled-Access Data Tier: The controlled-access data tier will not be freely available to the public, but will be made available to any *qualified* 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 include:

- Individual-level germline variant data (e.g. SNP6 .cel files).
- Whole exome and whole genome sequence data (.bam files residing at CGHub).
- RNA and miRNA sequence data (.bam files residing at CGhub).
- Raw MSI data.

## 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) at the NCBI dbGaP Authorized Access webpage (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login). The submission of the DAR ensures that investigators, along with their institutions and collaborators, understand the broad goals and policies of TCGA participant 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 obliges them not to attempt to identify or contact individual participants 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 is organized as a series of "linked" protocols, 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 to mean that all links between the sample and data back to the patient have been irretrievably broken.) Many participants who have contributed samples and data to TCGA are still living, and the medical centers at which they were enrolled for tissue banking continue to collect clinical, longitudinal and outcomes data that can be transmitted to TCGA under this linked protocol.

Tissue source 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 program. 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 tissue source sites 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 program is not categorically opposed to such linkage, but has established a policy that sample ID links between contributing site IDs and TCGA IDs will only be revealed to a contributing site investigator documenting an IRB approved protocol to use this information.

### 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. Accordingly, TCGA policies are reviewed periodically to ensure they are consistent with any changes in overarching regulations (e.g. HIPAA) and reflect lessons learned throughout the program. 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.

## Part 3: TCGA HIPAA Privacy Rule Compliance

### Background

US-based clinical sites (TCGA Tissue Source Sites (TSS)) at which participants are enrolled and clinical data are collected to annotate biospecimens are "covered entities" under HIPAA. Therefore, those clinical data are Protected Health Information (PHI), as defined by the Health Insurance Portability and

Accountability Act (HIPAA), and subject to the HIPAA "Privacy Rule" set in place to protect the confidentiality of patient information. (Note that some data collected from non-US sites is not technically subject to HIPAA, but the program does not treat such data differently.) 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 program 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 compliant with HIPAA. Currently, HIPAA only applies to patient data being disclosed by covered contributing sites and not to molecular data generated from samples within TCGA.

Therefore the goal of TCGA HIPAA policy is to set up a fully HIPAA-compliant clinical data pipeline that enables the maximal amount of potentially relevant clinical data annotating biospecimens to be transmitted to the program. 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.

## HIPAA Implementation under Limited Data Set Regulations and TCGA Policy

Of the mechanisms permitted for clinical data disclosure by covered entities under HIPAA, two were considered for contributing sites collaborating with TCGA. First, 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 in the Appendix.)

The LDS option was chosen after consultation with TCGA advisors because it was suggested that some investigators would benefit from the additional data (specifically, accurate dates of clinical events) allowed under the rule. TCGA implemented the necessary policies, contracts (i.e. the Data Use Agreement), information technology, and operations to be compliant with HIPAA disclosure and transmission of clinical data from covered entities to the program.

## 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 zip code, may be included; and (c) the catch-all 18th identifier ("Any other unique identifying number, characteristic, or code") is not included.
- LDS disclosure must be for the purpose of research, public health, or health care.
- LDS may be disclosed without an authorization or an IRB or privacy board waiver of authorization or alteration of authorization. (For documentation, see the Appendix 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 without requirement of HIPAA accounting regulations, under which covered entities must maintain a tracking database of all disclosures. (See the Appendix for HHS OCR guidance specific to this subject.)
- Disclosure of LDS requires that the discloser and recipients to enter into a Data Use Agreement (DUA) a contractual requirement under which 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.

## HIPAA implementation within TCGA

Operationally, clinical data in TCGA flow 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. This data flow is graphically described in Figure 1.

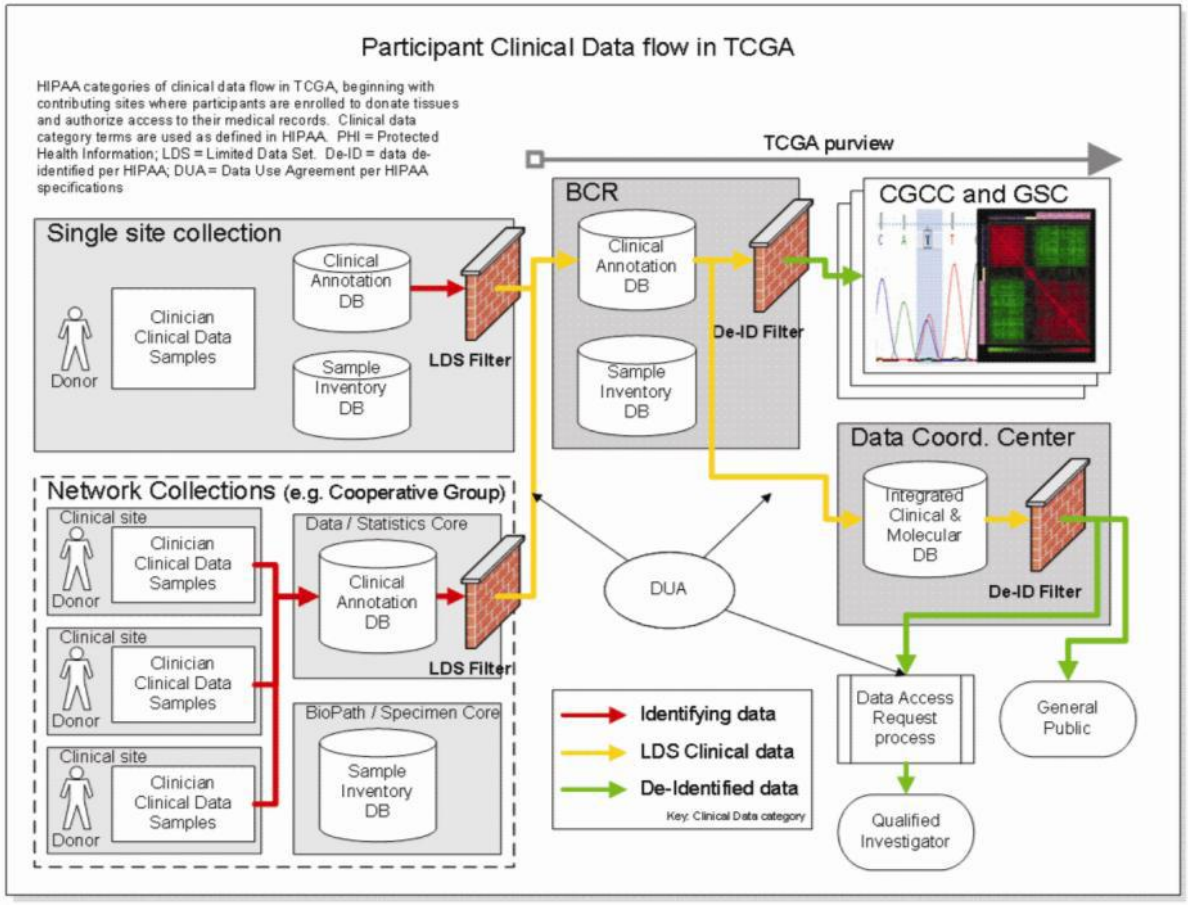


**Figure 1.** Data flow diagram illustrating what HIPAA category of data is permitted to be transmitted between TCGA collaborating entities.

## HIPAA and PHI Data flow <u>into</u> the TCGA Program

The flow of data into TCGA is regulated by a set of Data Use Agreements (DUA) between several entities in the consortium, as illustrated in Figure 1. Those DUA include the following components:

- An originating DUA is in place between the HIPAA "Covered Entity" contributing site (TSS) and BCR contractor(s). (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, which thus addresses both the physical material and LDS data transfer from a contributing site to the BCR. The MTA includes a warrant by the discloser that all PHI associated with TCGA tissue

donors will be compliant with the Limited Data Set specification. The DUA also pre-authorizes additional LDS-compliant data disclosure from the BCR to the DCC for purposes of the program.
- A DUA is in place between BCR contractor(s) and DCC (which is, in fact, the NCI) to permit transfer of LDS compliant data to the government. (The BCR's functional role with clinical data is to reformat them to be comparable between clinical sites and compliant with NCI data standards.)
- No DUA is in place between the BCR and GCC or GSC as these Centers will only receive minimal data associated with the molecular analytes being received. Those data will be De-Identified per the HIPAA definition.

### HIPAA and Data distribution <u>from</u> the TCGA Program to Investigators

The HIPAA policy covering data distribution from TCGA to investigators has evolved since the program's inception in 2006. In general, as information systems have become more sophisticated, the program has been able to incrementally restrict the distribution of LDC by making more of the data pipeline contain data compliant with HIPAA's definition of De-Identified.

Originally, during the Pilot Project, data distributed from the TCGA DCC to the broader investigator community included specific dates of participant clinical events. Thus, these data were technically PHI, compliant with a HIPAA LDS. (No geographical information is included in TCGA data.) A DUA was included as part of the TCGA Data User Certification (DUC) (see the section above on Process for DCC Data Access) executed between the DCC (NCI) and any researcher's employer.

At the end of Pilot phase, a decision was made to modify the above policy and make the distributed clinical data fully compliant with the HIPAA De-Identified definition. To be clear, the TCGA research network still receives HIPAA LDS, but no longer distributes it. Specifically, clinical event dates are no longer part of the data sets distributed from the DCC to investigators. All dates have been converted to negative or positive intervals referenced from the day of diagnosis. Any dates remaining are rounded to full years, thus meeting the HIPAA De-Identified test. It should be noted that the DCC retains full dates in its internal archives, but access to these dates requires specific permission from the Project Team.

Beginning in 2014, the process of converting full dates to intervals will move to an even earlier step in the TCGA data pipeline, to further limit such data's exposure. Under the new process, the BCR is responsible for this conversion, and the BCR will only distribute HIPAA fully De-Identified data to the DCC for inclusion in the TCGA database. The BCR will separately, regularly transfer LDS data (i.e. with dates) to the NCI for archival purposes.

### *Important Notes*

Regardless of the detailed point in the data pipeline where the data are converted from LDS to De-Identified and that all investigators only receive De-Identified data, all investigators and their institutions are still required to enter into a DUA. This requirement is in place for two reasons: a) some legacy data sets still potentially contain LDS; and, b) the terms of a DUA also include restrictions on redistribution of molecular data that are individually genotypic. Such genotypic data did not derive from a covered entity

and are not covered by HIPAA, but the same restrictions emplaced by the DUA on HIPAA data also provide important participant protections on the genetic data.

TCGA's HIPAA compliance policy is designed to enable the maximum amount of data to flow into TCGA databases, there may be other conflicting policies that may reduce o restrict the data submitted to the TCGA database. For example, IRBs from contributing sites may place greater restrictions on the amount of participant data submitted with tissue samples to TCGA. It is also possible that TCGA policies, as promulgated by the Project Team, may place greater restrictions on participant data accepted by the program or on what data residing in TCGA databases are made available to researchers. Such policies, often based on other sometimes conflicting bioethical considerations and future assessments of risks to participants, are described elsewhere in this document.

# Appendix – Attachments

45 CFR 164.514 (b): HIPAA De-Identification Safe Harbor

45 CFR 164.514 (e): HIPAA Limited Data Set Specification

OCR Guidance on HIPAA About Limited Data Set Disclosures

HHS Commentary and Response in Federal Register on LDS Disclosure as Part of Privacy Rule Modifications

OHRP Guidance on Research Involving Coded Private Information or Biological Specimens