



Issue 12 : July, 2014

OCG PERSPECTIVE



Collaborative Research to Advance Precision Medicine in the Post-Genomic World

My name is Subhashini Jagu, and I am the Scientific Program Manager for the Cancer Target Discovery and Development (CTD2) Network at the Office of Cancer Genomics (OCG).

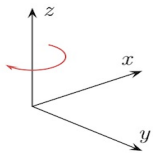
FEATURED RESEARCHERS



Frank McCormick Steps up His Efforts in Tackling a Lifelong Adversary

Dr. Frank McCormick, Professor Emeritus at the University of California-San Francisco (UCSF), is a dedicated leader in cancer research and an active member of OCG's Cancer Target Discovery and Development (CTD2) Network.

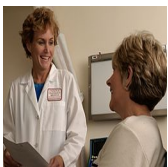
GUEST EDITORIAL



Functional Validation of Novel Cancer Driver Loci Using Three-Dimensional Organoid Cultures

Initiatives such as The Cancer Genome Atlas (TCGA), Cancer Genome Characterization Initiative (CGCI), and Therapeutically Active Research to Generate Effective Tr

NCI GENOMIC PROGRAM HIGHLIGHTS



From Patients to Data: A Glimpse into Tissue Processing and Clinical Data Collection

When genomics data from a project supported by the Office of Cancer Genomics (OCG) appear in the corresponding data matrix, the data are organized, labeled, and matched to patient information.

ANNOUNCEMENTS

OCG's New 'Guide to Accessing TARGET Data' Points Users in the Right Direction

Whether you're a researcher wanting to use genomics data for the first time, or a seasoned user in need of updating your data access login password, OCG's new Guide to Accessing TARGET Data is designed to

OCG PERSPECTIVE

Collaborative Research to Advance Precision Medicine in the Post-Genomic World

Subhashini Jagu, Ph.D.



My name is Subhashini Jagu, and I am the Scientific Program Manager for the Cancer Target Discovery and Development ^[2] (CTD²) Network at the Office of Cancer Genomics (OCG). In my new role, I help CTD² work toward its mission, which is to develop new scientific approaches to accelerate the translation of genomic discoveries into new treatments. Collaborative efforts that bring together a variety of expertise and

infrastructure are needed to understand and successfully treat cancer, a highly complex disease. The 13 Network Centers actively collaborate on a diverse range of projects and share the research findings (open access) with the broader research community. Through these activities, CTD² will contribute to understanding the mechanisms of cancer and potentially accelerate development of clinically useful markers, targets, and therapeutics.

I started working as the CTD² Program Manager in November 2013. While I am new to OCG, I have worked in cancer research for the past 15 years. During that time, I collaborated with private, public, national, and international scientific groups on identifying and developing potential anti-cancer agents. I helped manage those complex projects, which generated newfound interest in project management and led me to complete formal training in Leadership and Management in Life Sciences at the Johns Hopkins Carey Business School. Being the Scientific Program Manager for CTD² marries my deeply rooted passion for cancer research with my interest in project management.

As I have learned through my research experience, the field of cancer research is constantly evolving. Genomic and molecular characterization studies show that cancers are comprised of a patchwork of genetic abnormalities, some of which drive tumor formation, growth, and survival. These abnormalities can vary within a specific tumor type or overlap across cancer types (e.g., KRAS mutations in pancreas, colon, lung, and ovarian). Precision medicine, or identifying molecular features in individual tumors and tailoring therapies and clinical practices to those features, is a promising

new approach to treating cancers.

A major challenge in advancing precision medicine is translating the abundance of molecular data from large-scale initiatives like [The Cancer Genome Atlas](#) [3] (TCGA), [Therapeutically Applicable Research to Generate Effective Treatments](#) [4] (TARGET), and [Cancer Genome Characterization Initiative](#) [5] (CGCI) into clinically testable hypotheses. To overcome this challenge, the National Cancer Institute (NCI) created the CTD² Network initiative. This initiative undertakes highly collaborative and integrated approaches that use a variety of *in vitro* and *in vivo* biological assays to functionally validate discoveries from genomic initiatives. It brings the molecular characterization of tumors one step closer to the development of targeted cancer therapies and clinical biomarkers. Because it serves as a bridge between genomics and the clinic, I see this initiative as poised to make a huge impact in advancing cancer research.

As a community research project, CTD² releases all data it generates. Through this open access data-sharing model, CTD² strives to accelerate the development of cancer therapeutics by providing researchers access to identified and validated cancer dependencies. CTD² is the first initiative of its kind to generate and share such a diverse range of datasets. Data are deposited at the CTD² [Data Portal](#) [6], [6] the open access data portal maintained by OCG's [Data Coordinating Center](#) [7] (DCC). The Data Portal is unique in that it includes data types generated from different experimental approaches. Because it is such a useful resource, I want to make CTD² datasets as accessible as possible to cancer researchers that may have a range of backgrounds and expertise. As a result, I am working with the DCC to make the Data Portal user-friendly, for example, by reformatting the layout and standardizing data formats.

To organize and curate the growing number of unique datasets, analytical tools, and resources, representatives from NCI, including myself and the 13 CTD² Centers, formed the Data-Harmonization Informatics Portal (D-HIP) group. D-HIP created a common submission framework to: (1) exchange data within the Network and across the research community and (2) facilitate the process of downloading it from the open-access repository. As a lead representative from OCG, I advise the D-HIP team on best practices for data submissions. Additionally, I work with bioinformaticians within and outside NIH to develop and implement new data storage policies.

Since the CTD² pilot project was first initiated in 2009, collaborative efforts from the Network have made significant progress. The Network has generated many bioinformatics tools and reagents (cDNA clones & small molecule libraries), shared many datasets through the Data Portal, and published 30 joint/integrated manuscripts. The tools, for example, help researchers with little bioinformatics expertise to analyze and extract potentially meaningful information from large-scale genomic initiatives like TCGA, TARGET, CGCI, [Big Data to Knowledge](#) [8] (BD2K), and other initiatives. Providing these resources to the broader research community will accelerate the discovery of biologically relevant targets from large-scale genomic datasets.

I am excited to be part of this unique initiative. I look forward to working with the

Centers to achieve our shared goal of promoting translation of genomic data into clinically relevant cancer treatments through cutting-edge collaborative research.

[Back to Top](#)

FEATURED RESEARCHERS

Frank McCormick Steps up His Efforts in Tackling a Lifelong Adversary

Shannon Behrman, Ph.D.



Dr. Frank McCormick, Professor Emeritus at the University of California-San Francisco (UCSF), is a dedicated leader in cancer research and an active member of OCG's [Cancer Target Discovery and Development](#) ^[2] (CTD²) Network. Since the early 1980s, he has devoted much of his career to understanding the molecular mechanisms of the oncogenic RAS pathway and working toward developing novel therapies that treat RAS-driven cancers. His efforts won him numerous awards and honors, including his most recent election to the National Academy of Science on April 29, 2014.

RAS oncoproteins, initially discovered in “**rat sarcomas**,” can transform normal-growing cells into tumor-like cells. Because they are found in over 30% of all human cancers, targeting RAS oncoproteins (or components of the RAS signaling pathway) could make a huge impact on clinical outcome. Despite having studied RAS for decades, scientists have not yet found a way to successfully treat RAS-mutated cancers.

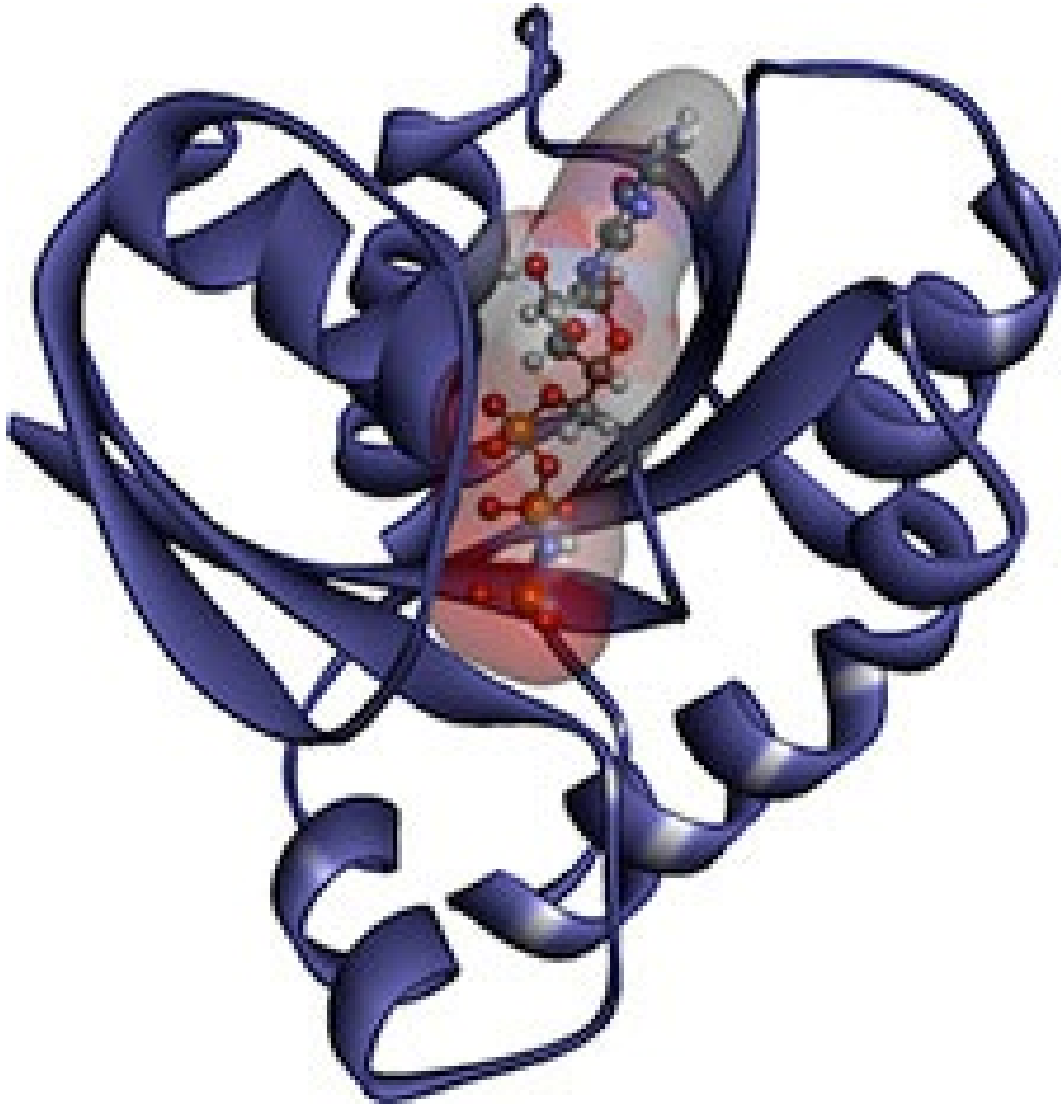
Recently, Dr. McCormick stepped up his efforts in tackling RAS by becoming the Scientific Project Leader of the National Cancer Institute's (NCI) [RAS Program](#) ^[9], which is headquartered at NCI's [Frederick National Laboratory for Cancer Research](#) ^[10] (FNLCR) in Frederick, Maryland. Now he splits his time between his UCSF lab on the west coast and the FNLCR on the east coast. “I am doing what I love to do best,” said McCormick, “which is to solve the ‘undruggable’ RAS problem.”

OCG recently interviewed Dr. McCormick to talk about his early studies of RAS, the initiation and goals of the RAS Program, and the significance of collaborative initiatives in cancer research.

What is currently known about RAS and its role in human cancers?

RAS is a family of G proteins, or guanosine nucleotide-binding proteins, which regulate cell growth and survival. RAS proteins localize to the cell membrane and mediate signals between extracellular growth factor receptors and intracellular circuitry. They function as binary on and off switches, and their activity depends on which signals are being transmitted through the cell. In cancer cells with RAS mutations, the on state is locked and the cells continue to proliferate even in the absence of growth factors. This uncontrolled growth ultimately leads to cancer.

RAS oncoproteins play a major role in driving cancers such as pancreatic, colon, and lung cancers, which frequently contain the constitutively active form of RAS. The oncoproteins are activated by a single mutation, which changes one amino acid to lock the protein in the on state. Of all the RAS family proteins, KRAS is the most commonly mutated in human cancers. Targeting the mutant KRAS proteins could improve outcomes in patients with tumors that have those mutations.



Structure of human KRAS protein (Image by Sarangan Ravichandran, Ph.D., Frederick National Laboratory for Cancer Research)

Can you briefly explain some of your earliest efforts in studying RAS?

In the early 1980s, we learned that RAS is a GTPase that binds to guanosine triphosphate (GTP) and hydrolyzes it into guanosine diphosphate (GDP). We performed biochemical analysis on various RAS mutants to further understand the molecular basis of RAS function and how it cycles from the active to inactive state. We discovered a class of enzymes, called GTPase Activating Proteins (GAP), which regulate RAS activity by keeping it turned off under most conditions. The mutant proteins are resistant to GAP and, therefore, accumulate in the active state.

Following this discovery, we screened for compounds that would fix the broken

switch of RAS. We didn't succeed for reasons that are now apparent. To get around this hurdle, we started an effort to block signaling immediately downstream of RAS. We developed a drug called sorafenib, which inhibits rapidly accelerated fibrosarcoma (RAF) kinase, a protein that RAS directly activates. The drug was successful in treating many cancers, such as liver and kidney cancers, but it didn't work on treating RAS-mutant cancers. That was 20 years ago, and now, with a better understanding of the biochemistry surrounding the RAS signaling pathway, we know that RAF kinase inhibition actually causes RAF hyperactivation.

Why have previous efforts in targeting RAS been unsuccessful?

The RAS protein is like a tennis ball. It lacks pockets or grooves in which a small molecule could easily bind and turn off signaling. GTP and GDP bind RAS, but with extremely high affinity, which makes finding a drug that competes with GTP/GDP binding technically out of reach. This is in contrast to protein kinases, in which small molecules can more easily compete with relatively low-affinity binding of the nucleotide, adenosine triphosphate (ATP), to successfully inhibit kinase activity.

What recent advances and insights are making a concerted effort to targeting RAS possible?

Despite the challenges we encountered in trying to target RAS in the past, technologies have changed a lot in recent years. Now there are clever methods, such as fragment-based screening [11], which may reveal crevices in the proteins that could be utilized for therapeutic intervention.

More information about RAS has also generated new ideas for inhibition. One example is that new evidence suggests RAS proteins may actually function as dimers or higher-order structures, rather than as monomers. If this is true, then we could disrupt RAS function by interfering with the dimer interface. There is also other evidence suggesting that, when RAS inserts into the plasma membrane, it creates a pocket between the protein and the membrane. Finding small molecules that fit into this pocket could also disrupt RAS function.

Describe the RAS program, including its goals and how it will achieve those goals.

The RAS program is based at the Frederick National Lab for Cancer Research (FNLCR). The goal of the program is to find new ways to target RAS directly or indirectly by filling in the knowledge gaps of RAS biology and by facilitating drug discovery. Our projects primarily focus on *KRAS*, because it is the most frequently mutated *RAS* gene in human cancers.

Many people in industry want to develop drugs that target RAS, but starting drug discovery programs is overwhelming. There are too many unknowns. For example, we don't know how RAS activates RAF. By providing the groundwork for understanding RAS biology, FNLCR will make it easier for the external research community to identify new targeting strategies and evaluate them through pre-clinical and clinical studies.

The RAS program operates through a "hub and spoke" model. The FNLCR is the

“hub” that exchanges insights and reagents with the “spokes” of the biopharmaceutical industry, academia, and NCI intramural labs. We all work together to achieve the ultimate goal of getting to market clinically approved drugs that treat RAS-mutated cancers.

How is the RAS Program collaborating with CTD²?

Several CTD² investigators are now focusing part of their efforts on validating KRAS as a cancer target and identifying direct or indirect ways of potentially targeting it. These collaborations are examples of “spokes” in the “hub and spoke” model of the RAS Program.

To name a few of the many CTD² collaborations:

- 1) Martin McIntosh from the Fred Hutchinson Cancer Research Center will map *KRAS*-mutated cancer cells to find surface proteins for immunotherapy attacks.
- 2) Andrea Califano from Columbia University is using bioinformatics approaches to identify mutant KRAS vulnerabilities.
- 3) Stuart Schreiber from the Broad Institute aims to find small molecules that target KRAS directly through high-throughput screening.
- 4) Calvin Kuo from Stanford University is providing organoid model systems [12] that will be used for validating KRAS as a target in three-dimensional cell cultures.
- 5) Scott Powers from Cold Spring Harbor Laboratory is using sensor-based siRNAs to target RAS pathway components by RNAi delivery.

How are initiatives like CTD² and the RAS Program influencing the future of cancer research?

These collaborative projects are essential in solving many of the big questions in cancer research, given that cancer is such a complicated disease. There is room for individual innovation and single investigator/lab brilliance. However, insights are accelerated through the sharing of ideas and complementary technologies, such as organoid model systems, high-throughput screening, and data processing. As a result, big projects such as CTD² and the RAS program will be very helpful in moving the field forward.

More information on the RAS Program

Visit the RAS program website [9] to learn more about the RAS Program, including the individual projects that aim to address specific knowledge gaps and generate community-based resources.

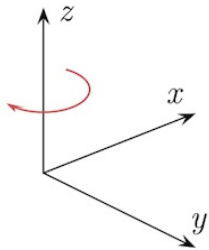
Additional References

- <https://profiles.ucsf.edu/frank.mccormick> [13]
- https://deainfo.nci.nih.gov/advisory/fac/0913/NFAC_RASupdate.pdf [14]

GUEST EDITORIAL

Functional Validation of Novel Cancer Driver Loci Using Three-Dimensional Organoid Cultures

J.T. Neal, Ph.D., Michael A. Cantrell, Ph.D., and Calvin Kuo, M.D., Ph.D.



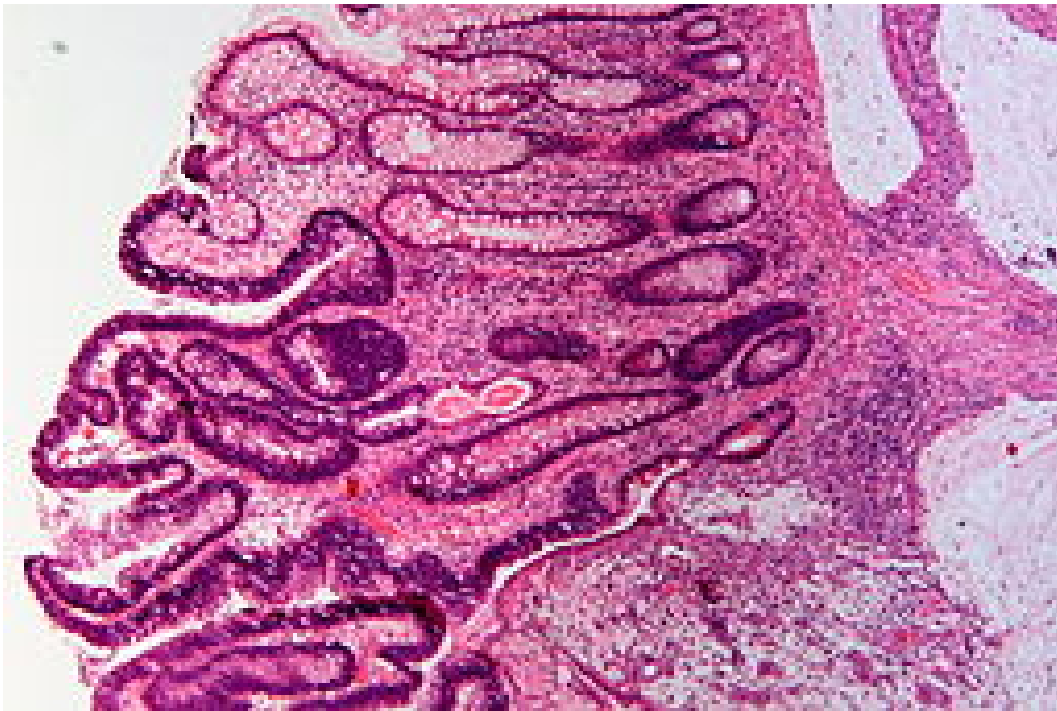
Initiatives such as [The Cancer Genome Atlas](#) ^[3] (TCGA), [Cancer Genome Characterization Initiative](#) ^[5] (CGCI), and [Therapeutically Active Research to Generate Effective Treatments](#) ^[4] (TARGET) are plumbing the depths of human cancer genomes. Their goal is to identify, from the thousands of mutations, rearrangements, and other alterations observed in cancer, those that are actually responsible for driving the development and progression of this complex family of diseases. Having generated petabyte-scale datasets, these projects are impressive in scope. This has created a pressing need for robust tools to functionally validate driver genes and multigenic driver modules (i.e., co-mutations) from among these data. Three-dimensional (3D) organotypic cultures (organoids) have emerged as one novel tool to address this need and study tumorigenesis and tumor progression.

Historically, *in vitro* functional validation of putative cancer drivers has relied on the use of transformed cell lines, which usually exhibit widespread DNA mutations and chromosomal rearrangements. Attempting to determine the effects of single up- or downregulated loci within a milieu of background alterations can create a significant signal-to-noise problem. Often, two-dimensional (2D) models also fail to recapitulate the complex microenvironment of cell-cell and cell-extracellular matrix interactions that are present in intact tissues. These interactions mediate cellular behaviors that are important for processes that sustain tumor progression, such as proliferation, epithelial to mesenchymal transition, migration, and invasion. Furthermore, responses to drug treatment can vary significantly between planar, 2D cell line models and more biologically accurate 3D culture models. *In vivo* mouse models have also been used for driver gene validation, but cost, strain variability, and generation time limit their application.

Organoid models address many of the shortcomings of cancer cell lines while retaining the genetic tractability and throughput associated with conventional monolayer cultures. Organoids have been developed for many tissue types, each with variable display of accurate *in vivo* architecture and cellular differentiation. Our group has pioneered air-liquid interface (ALI) organoid culture methodologies for the culture of diverse normal tissues and tumors from the mouse gastrointestinal tract^{1, 2}. In this system, organs are mechanically dissociated and suspended in a collagen gel that mimics the *in vivo* extracellular matrix environment. Importantly, this method optimizes oxygenation by directly exposing the collagen gel to air instead of submerging it beneath tissue culture medium. This setup significantly enhances growth of explants from gastrointestinal tissues. Cell aggregates expand rapidly to form cystic organoids that recapitulate the *in vivo* structure and differentiation of the

tissues from which they are derived. Organoids generated in this manner can be cultured indefinitely and are suitable for utilization in a wide variety of molecular and histological assays. Other key elements of this system are that (1) wild-type, non-transformed tissue can be easily cultured and manipulated alongside cancer tissue for assessment of normal developmental processes, and (2) both epithelial and mesenchymal components are included.

These organoid culture methods represent a powerful tool for translational genomics. Our most recent publication ^[15] in *Nature Medicine* documents the adaptation of our ALL system to evaluate specific genomic alterations in tumorigenesis and the acquisition of histological traits of tumors. The molecular techniques we developed allowed us to transform normal wild-type organoid cultures from diverse gastrointestinal tissues into adenocarcinomas of the colon, stomach, and pancreas *in vitro*.



Histopathological stain of an adenocarcinoma of the colon (image taken by Patho; licensed via Wikimedia Commons [CC-BY-SA-3.0](#) ^[16]).

We systematically assessed the oncogenic potential of individual genes within the 11p15.5 amplicon in human colorectal cancer. Surprisingly, the microRNA ^[17], *mir-483*, is a dominant oncogene within this interval. This result was unexpected because *mir-483* lies within an intron of *IGF2*, which itself was previously presumed to be the driver oncogene at this locus. However, in the absence of *mir-483* overexpression, *IGF2* did not induce transformation². Going forward, potential cancer driver genes from genome-scale surveys such as TCGA can be tested for the ability to induce tumorigenicity in organoids derived from normal tissues from a variety of organ systems.

As members of two National Cancer Institute initiatives, Cancer Target Discovery and Development ^[2] (CTD²) and Integrative Cancer Biology Program ^[18] (ICBP), we are

applying our organoid method to the discovery and validation of novel cancer driver genes. Our goal is to yield biological insights into the processes of tumor initiation and progression, as well as identify new targets for therapeutic intervention. In this vein, our collaborators in the Stanford CTD² Center have identified potential driver oncogenes from a variety of solid tumor types using systems biology approaches, which we are currently screening. By overexpressing or knocking down these genes individually or in physiologically relevant combinations within our organoid cultures, we hope to identify novel drivers of these diseases.

In addition to our work within the Stanford CTD² and ICBP initiatives, we are actively collaborating with others across the CTD² Network, including the Dana Farber Cancer Institute, the Broad Institute, and the University of California-San Francisco. These projects are leveraging organoid functional genomics approaches to identify novel “pan-cancer” targets responsible for progression of multiple types of cancer. Further, we have two independent collaborations with the RAS Program at the NCI Frederick National Laboratory of Cancer Research and the Emory University CTD² Center to perform high-throughput interrogation of small molecule therapeutics in cancer organoids.

Organoid model systems developed by our lab and others have significant potential to facilitate cancer biology research. They combine the experimental tractability of transformed cell lines with the accurate tissue architecture and differentiation of *in vivo* systems. We are also developing novel techniques to utilize organoids derived from patient tumors to study heterogeneity and resistance to therapy in a wide variety of neoplasms [19]. Such diverse organoid methods will provide the research community, as well as networks such as CTD² and ICBP, with powerful new approaches for drug discovery and the facile functional validation of cancer driver genes.

References

1. Ootani A, Li X, Sangiorgi E, Ho QT, Ueno H, Toda S, Sugihara H, Fujimoto K, Weissman IL, Capecchi MR, Kuo CJ (2009). Sustained in vitro intestinal epithelial culture within a Wnt-dependent stem cell niche. *Nature Medicine* 15(6): 701-6 (PMID 19398967)
2. Li X, Nadauld L, Ootani A, Corney DC, Pai RK, Gevaert O, Cantrell MA, Rack PG, Neal JT, Chan CW, Yeung T, Gong X, Yuan J, Wilhelmy J, Robine S, Attardi LD, Plevritis SK, Hung KE, Chen CZ, Ji HP, Kuo CJ (2014). Oncogenic transformation of diverse gastrointestinal tissues in primary organoid culture. *Nature Medicine*. 20(7): 769-77 (PMID 24859528)

[Back to Top](#)

NCI GENOMIC PROGRAM HIGHLIGHTS

From Patients to Data: A Glimpse into Tissue Processing and Clinical Data Collection



When genomics data from a project supported by the Office of Cancer Genomics (OCG) appear in the corresponding data matrix, the data are organized, labeled, and matched to patient information. The availability of high quality, clinically-annotated molecular data is crucial for the study of biologic factors that influence the progression and treatment responses of cancers.

What may not be readily apparent is that obtaining the data is a lengthy and rigorous process. From taking a tumor biopsy to sequencing tissue, care must be taken by participating clinicians, pathologists, and research personnel at every step.

The process begins in the clinic. Clinicians take tumor and matched normal tissues from patients that formally consented to donating samples for biological studies. The tissues are properly frozen and stored in a central facility and undergo pathological review to confirm disease identification. Whenever possible, donor information that poses little risk of exposing patient identities, such as gender, age, disease stage, and outcome, is collected over time.

From case-matched tissues of sufficient quality, researchers extract nucleic acids for sequencing and other genomic characterization. They perform a variety of bioinformatics analyses on the raw characterization data to determine the genetic alterations present in the tumors, but absent in the normal tissues. For each case, the analyzed genetic information, along with appropriate donor information, is shared with project investigators. The investigators use the data to find candidate causative mutations and make correlations between the genetic features of tumors and the characteristics or outcomes of patients (e.g., *EGFR* mutations occur more frequently in female lung cancer patients that never smoked). Raw and interpreted data are quality controlled and eventually deposited into project-specific databases for the broader research community.

The Need for Standardization

With so many steps and players involved in the tissue-to-data process of large-scale cancer genomics initiatives, a certain amount of standardization is needed. Prior to the establishment of such large consortia, individual laboratories performed many of the steps for their own studies by banking and processing tissue samples, molecularly characterizing those tissues, and collecting participant clinical data. This was possible because the number of participant donors within a given study was relatively small. As a result, individual labs successfully published results based on analyses of their own data.

The problem with this decentralized approach, however, was the extreme variability that existed between the lab protocols and data standards of individual groups. This hindered the ability to compare and integrate datasets across projects from different groups to produce larger, more statistically powerful datasets. Technical challenges also arose, such as systematic error generated when samples are processed in batches, which significantly impacted statistical validity.

Over the last decade, the National Cancer Institute (NCI) established multi-institutional genomics initiatives such as those run by OCG. As these projects grew in

size and scope, the NCI recognized the problem of variability between groups and its repercussions on data integration. NCI's solution was to impart a balanced level of centralization and uniformity. To start, NCI established centralized laboratories for tissue and data processing and identified a limited number of centers to perform molecular characterization (e.g., sequencing and expression analysis). By concentrating such work into core sites, the NCI could more readily implement a set of uniform requirements and best practices (i.e., Standard Operating Protocols; SOPs) to be followed by all sites/centers involved. For example, NCI provided clinical sites with template Institutional Review Board protocols and Informed Consent language, along with research pathology protocols. NCI also requested the tissue processing and molecular characterization centers adopt a uniform data labeling system as well as co-isolate DNA and RNA from the same piece of tissue (as opposed to isolating DNA and RNA from separate pieces of the same tissue). The SOPs help to minimize variability both within and across cancer genomics studies supported by NCI, including OCG projects.

Tissue Processing and Clinical Data Collection

Since 2009, the Research Institute at Nationwide Children's Hospital [20] (NCH) has contributed to the standardization of the tissue-to-data process. It serves as the tissue processing and clinical data collection core for many multi-institutional studies supported by the NCI's Center for Cancer Genomics (CCG), including both The Cancer Genome Atlas [3] (TCGA) and the genomic characterization programs (CGCI [5] and TARGET [4]) run by the Office of Cancer Genomics (OCG). NCH also serves as the biorepository for three NCI cooperative groups: the Children's Oncology Group, Gynecologic Oncology Group, and SWOG (formerly the Southwest Oncology Group).

NCH provides a uniform system for processing tissues systematically. They train the clinical sites participating in OCG and other NCI projects to follow biological quality control standards for the handling of sensitive materials. Nearly all OCG projects rely on rapidly frozen biomaterials, which are resected and frozen with minimal ischemia times and maintained at cryogenic temperatures. After accrual, clinical sites send the biological tissues to NCH, which follows over 150 SOPs that govern all procedures, including proper storage of biospecimen samples, pathology review, extraction of corresponding nucleic acids, and quality assurance of samples and analytes. NCH then ships DNA and RNA of sufficient quality to molecular characterization centers.

NCH also provides a uniform system for the collection and tracking of accurate, relevant donor information for each patient case (e.g., history of smoking and tumor stage of lung cancer patients). They deploy an electronic web-based Case Report Form system directly to clinical sites, where disease expert clinicians fill out the necessary demographic and clinical information. NCH translates these data, which may often be described using diverse wording, into a set of standardized terms. These terms, embedded as Common Data Elements, are registered at the NCI's Cancer Data Standards Registry and Repository (caDSR). This ensures that the same sets of vocabularies are used to describe all cases, which makes searching and cross-referencing the data straightforward. OCG studies using the NCH clinical data collection system include the CGCI projects, HIV+ Tumor Molecular Characterization Project [21] (HTMCP) and Burkitt Lymphoma Genome Sequencing Project [22] (BLGSP), as well as the functional genomics initiative, Cancer Target Discovery and Development

[2] (CTD²).

Molecular Characterization and Data Uniformity

Using uniform protocols and platforms (e.g., whole genome and transcriptome sequencing), the molecular characterization centers participating in OCG projects analyze the DNA and RNA extracted from case-matched tissues. They evaluate the molecular data for quality and accuracy. The centers provide this quality control information, along with the genomics data, to OCG's [Data Coordinating Center](#) [7] (DCC), the central organizer of OCG project databases.

In parallel, NCH also provides the DCC the de-identified donor information associated with each case studied. The DCC then cross checks all data, including molecular, clinical, logistical, and tissue processing, using a standard data model that is the same across all CCG programs. This uniformity enables the comparison of data not only between individual clinical sites collaborating within a single program, but also across multiple programs when the variables in question overlap. After genomic and clinical data pass inspection, the DCC eventually deposits the clinically annotated genomic data into project-specific databases for the research community.

What Remains

Although some challenges remain, the standardized practices implemented by OCG and other NCI-supported genomic initiatives make tissue processing and clinical data management more efficient, facilitate multi-project data integration, and ensure regulatory compliance. All parties involved, including the clinical sites, tissue processing and data collection cores, and the molecular characterization centers, are working together with NCI to refine the process and build on improvements.

Just as genomic characterizations efforts begin with patients, they hope to end with patients through their contribution to the development of targeted therapies and other genetically-informed treatment strategies.

[Back to Top](#)

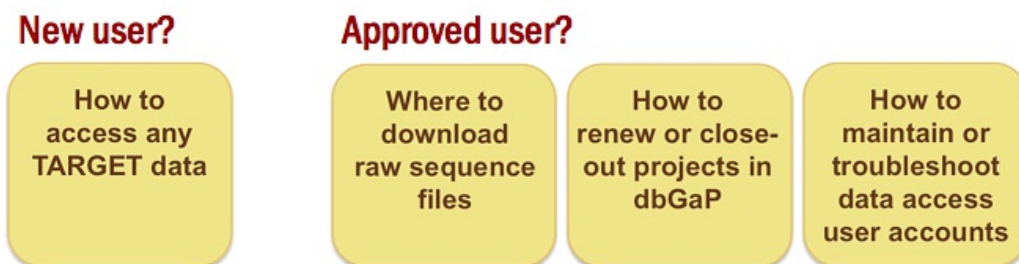
ANNOUNCEMENTS

OCG's New 'Guide to Accessing TARGET Data' Points Users in the Right Direction

Jessica Mazerik, Ph.D., and Shannon Behrman, Ph.D.

Whether you're a researcher wanting to use genomics data for the first time, or a seasoned user in need of updating your data access login password, OCG's new [Guide to Accessing TARGET Data](#) [23] is designed to make your life easier. It is a handy, easy-to-use resource that has a Q&A style format with graphics and embedded videos. The guide is accessible from the [Using TARGET Data](#) [23] webpage or you can bookmark it in your web browser for quick reference. Go ahead; take a peek at the Guide. After you look through it, we would appreciate suggestions and feedback via email: ocg@mail.nih.gov [24].

The [Guide to Accessing TARGET Data](#) [23] is intended to help all users – new, approved, or experienced – navigate through the Data Use Certification process, maintain user accounts, and download TARGET data. Starting at the top of the guide, there are clickable shortcuts that let you jump to the section most relevant to your needs:



New Users

Investigators that have never used OCG-generated data can click on the “How to access any TARGET data” box to start at the beginning of the guide. Here, you will learn about the two types of data available from TARGET: open and controlled access. Open access data are readily downloadable by anyone. Controlled access data can only be downloaded by researchers that obtained approval in the form of Data Use Certification. The links to open and controlled access data are color-coded in the Data Matrix for quick identification (see below).

The next section “How do I access open and/or controlled TARGET data?” contains a visual and interactive flowchart that walks you through the entire data access process. Regardless of what type of data you seek, the flowchart will help you find and download it.

The flowchart starts at the [TARGET Data Matrix](#) [25], because most data are easily accessible through its links:

TARGET DATA MATRIX				New user? START HERE		TARGET			
Disease	Patient Data	Gene Expression	Copy Number	Methylation	Other				
Acute Lymphoblastic Leukemia (ALL)									
Acute Lymphoblastic Leukemia (ALL) Phase I	Clinical File	Affymetrix U133 Plus 2	Affymetrix SNP 500K		Whole Genome	Whole Genome Lite	mRNA-seq	Targeted Re-sequencing	Kinome
	Case Matrix	DCC Open*	DCC Open* DCC Controlled†		View cases DCC Controlled† NCBI SRA†	View cases NCBI SRA†	DCC Open* DCC Controlled† NCBI SRA†	Sequence Trace Linking Table† 125 Genes*	Kinome Linking Table† 113 genes*

The blue links indicate open access data:

**Visit the DATA MATRIX and
click on blue links for
open access data**

The orange links designate controlled access data, accessible to only approved users with the correct login credentials. To access these data requires navigating through the Data Use Certification process to gain approval; this is where the flowchart comes into play.

The flowchart provides an instructional video from the database of Genotypes and Phenotypes (dbGaP). Through recorded screen sharing and audio, this tutorial shows users in a step-by-step fashion how to apply for access to controlled data. For users looking to cut to the chase, a summary of the video's most important points are also highlighted:

Apply for Access to Controlled TARGET Data through dbGaP:
Summary of dbGaP tutorial for accessing data**

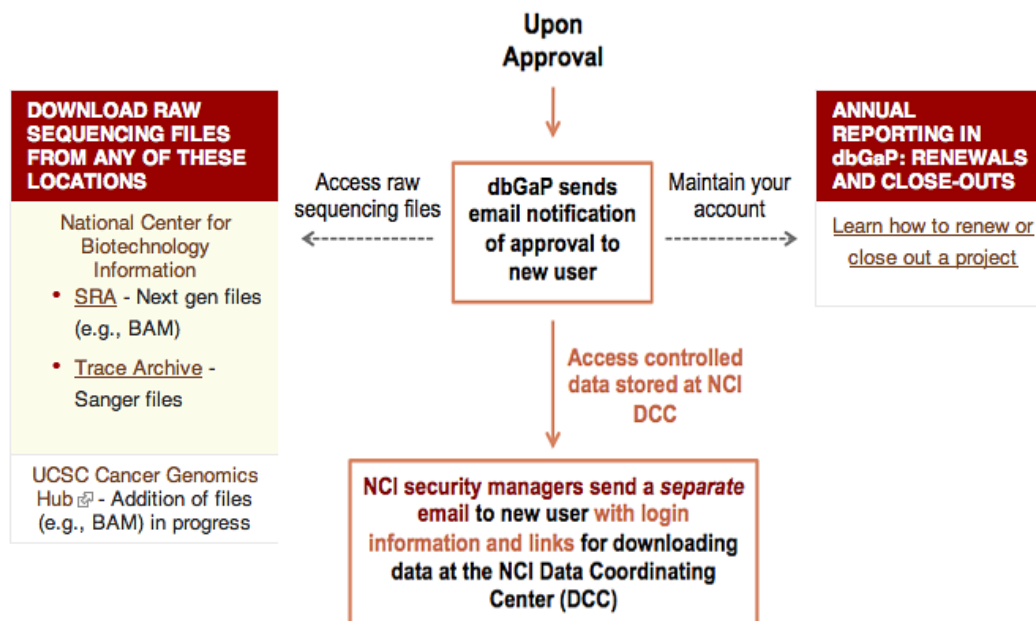
1. [Login to dbGaP](#) using eRA Commons credentials.
 - If you do not have an eRA Commons account, register for one on the dbGaP authorization page
2. Create a new research project page to fill out data access request
 - Input specific project info and datasets you wish to access
 - Add all other required information
3. Review the project-specific Data Use Limitations
4. Attach all required documents (e.g. IRB form)
5. Review and submit request
6. Signing official accredits request

Approval status?
Users can track approval status on dbGaP project page.
dbGaP will also send email notifications.

****Apply separately for access to CGCI data. The process is the same, but the Data Use Limitations are slightly different**

Approved Users

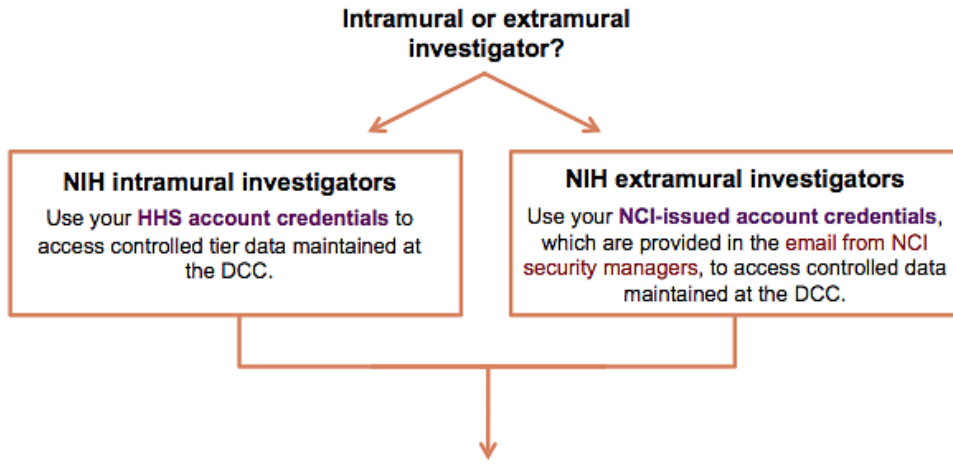
The remaining pieces of the flowchart summarize how to go from receiving approval notifications to downloading the data:



When you are first approved, you will be notified by an email from dbGaP. If you do not receive an email within about 4 weeks, contact dbGaP. Your dbGaP account initiates your ability to access data, but it also has to be maintained **through annual reporting**. The flowchart provides a link to information further down on the page that tells you how to maintain your dbGaP account. It is important to follow these instructions; otherwise you may lose access to the data one year after you gain approval.

One type of controlled access data, primary sequencing files, are housed at repositories external to the National Cancer Institute (i.e. National Center for Biotechnology Information and University of California, Santa Cruz). You cannot access these login pages directly from the TARGET Data Matrix. The flowchart conveniently links to each location where you can login and download sequencing files.

All other controlled access TARGET data are housed at NCI's Data Coordinating Center (DCC). You can access these data from the TARGET Data Matrix. Shortly after receiving your dbGaP approval notification, you will receive a second email from the NCI's security managers. This email will contain your specific login credentials and links to access controlled data stored at NCI's DCC. Your login credentials will depend on whether you are an external investigator or an internal NIH investigator, and the flowchart covers this important distinction:



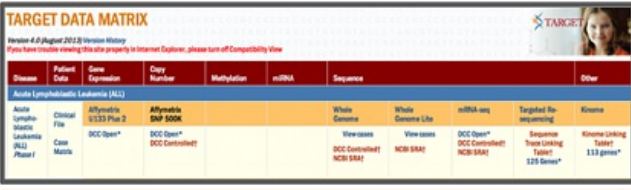
After OCG's new flowchart has guided you through the Data Use Certification process for controlled access data, head to the TARGET Data Matrix to download the different types of data.

Download controlled access data via links in email or the the DATA MATRIX

[Visit the TARGET Data Matrix](#)

Examples of controlled access data available:

- Specific phenotype or genotype data for each case
- VCF or MAF files from whole genome sequencing (WGS)
- VCF or MAF files from mRNA-seq
- Raw or low level chip-based data that may identify an individual
- Information linking all sequence traces to an individual



Experienced Users

Finally, don't forget about the new guide once you have access. The flowchart also has information about how to maintain user accounts...

	NCI Account (DCC data)	eRA Commons (Grants, NCBI, CGHub)
Register password or login	http://password.nci.nih.gov	https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login
If account inactive	Login above	Login above
Account expires in	120 Days	180 Days
Troubleshooting	ncicbiit@mail.nih.gov ; 240-276-5541	http://era.nih.gov/commons/faq_commons.cfm

...and video tutorials that teach you how to renew and close out data access projects in dbGaP.

[Back to Top](#)

Source URL:<https://ocg.cancer.gov/news-publications/e-newsletter-issue/issue-12>

Links

[1] <https://ocg.cancer.gov/printpdf/taxonomy/term/76> [2] <https://ocg.cancer.gov/programs/ctd2> [3] <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga> [4] <https://ocg.cancer.gov/programs/target> [5] <https://ocg.cancer.gov/programs/cgci> [6] <https://ocg.cancer.gov/programs/ctd2/data-portal> [7] <https://ocg.cancer.gov/data/data-policies> [8] <https://commonfund.nih.gov/bd2k> [9] <https://www.cancer.gov/research/key-initiatives/ras?redirect=true> [10] <https://frederick.cancer.gov> [11] <https://pubmed.ncbi.nlm.nih.gov/22934764/> [12] <https://ocg.cancer.gov/news-publications/e-newsletter-issue/issue-12#396> [13] <https://profiles.ucsf.edu/frank.mccormick> [14] https://deainfo.nci.nih.gov/advisory/fac/0913/NFAC_RASupdate.pdf [15] <https://pubmed.ncbi.nlm.nih.gov/24859528/> [16] <https://creativecommons.org/licenses/by-sa/3.0/> [17] <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/microrna?redirect=true> [18] <http://icbp.nci.nih.gov/> [19] <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/neoplasm?redirect=true> [20] <https://www.nationwidechildrens.org> [21] <https://ocg.cancer.gov/programs/cgci/projects/HTMCP> [22] <https://ocg.cancer.gov/programs/cgci/projects/burkitt-lymphoma> [23] <https://ocg.cancer.gov/flowchart/guide-accessing-data> [24] <mailto:ocg@mail.nih.gov> [25] <https://ocg.cancer.gov/programs/target/data-matrix>