Frederick National Laboratory
for Cancer Research

The Center for Strategic Scientific Initiatives Data Coordinating Center

[CSSI DCC]

Mary K. Anderson, Ph. D.
Corinne Zeitler, Ph.D.
Charles Shive
Joan Pontius, Ph.D.
Anita Undale, MD, Ph.D.
Maureen Dyer, Ph.D.
Deb Hope, Ph.D.
Andrew Quong, Ph.D.

9/30/2015

Contents

Table of Figures and Tables

Vision

The Center for Strategic Scientific Initiatives (CSSI) data coordinating center (DCC) will be a durable repository for data from disparate sources that span the range from molecular to phenotypic and are often generated from the latest advances in technology. The DCC will have a portal for access to the data with the interactive functionality of browsing, searching and downloading data from these studies. Users will be able access a website to upload data along with the corresponding minimally complete set of metadata, the information that describes the data.

Overview

The overall goal for the creation of the DCC is to enhance and expedite cancer research by creating a data sharing environment (DSE) that provides access to research data as well as tools to verify data format and interact with that data. Use of the DCC to disseminate the data will facilitate collaboration and future elaboration of the originating research. The majority of the projects funded by CSSI utilize new and emerging technologies in addition to existing technologies to support emerging initiatives. To overcome the challenge of housing and mining data obtained from disparate technologies, the DCC will be will be required to be more flexible than many of the resources for 'Omics scale data available to date. The DCC is envisioned to be an integrated, web-based data store that can manage disparate data types generated by projects currently supported or initiated by CSSI for use by the cancer research community.

The DCC will serve the research community by –

- Hosting diverse biologic data generated in support of cancer research by the CSSI, including data from cell lines, animal models, animal/human xenograft, and clinical samples
- Providing the ability to search between different sets of physical, cellular, phenotypic, and molecular data and metadata.
- Allowing the dissemination of data to the larger research community
- Providing a scalable, persistent resource to house data.

Key Features and Benefits

I. Researchers will be able to accession data into the store and provide the required set of metadata. The metadata is the set of information that describes the study that was performed, the type of experiment or assay performed, the input and output information whether observation or calculated and the pedigree and/or provenance of the input information so that the data from a variety of experiments can be included.

II. Researchers will be able to access data in the store either for inspection or for download without overly obtrusive access policies so that data can be re-used.

III. Researchers will be able to query across submitted datasets by using information within the data and metadata.

IV. Information about any controls or standards used and the file types included in the datastore can be provided so that a consumer of that data will have confidence in structure and integrity of the information.

V. Information submitted will be validated to ensure that it meets the quality criteria for the repository on several levels to ensure consistency across data sets.

VI. Information will conform to standard ontologies and metadata specifications in common use such as the ISA-Tab format.

VII. The DCC will make use of a flexible data model to allow the submission of novel data types to create a diverse datastore and that can accommodate new data types.

VIII. The DCC will provide a more permanent location to house data that is not subject to project completion or contract expiry so that the data persists for continual use by the research community.

IX. The DCC will serve as a central repository for eligible data generated by multiple projects and initiatives.

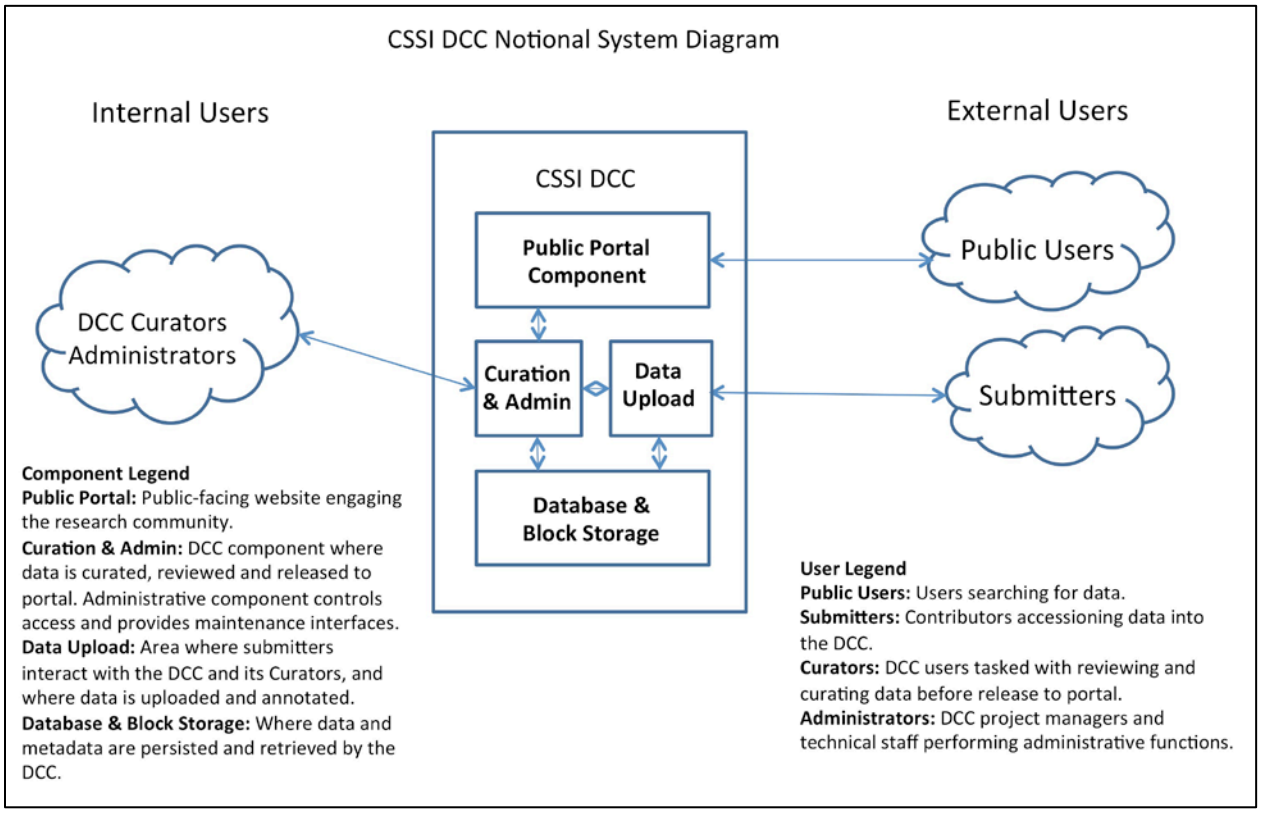

Figure 1 Proposed System Diagram

Methodology

The project team has analyzed and characterized the data that has been developed to date on completed and in-progress projects sponsored by CSSI in order to highlight commonalities in data, metadata and other file attributes. Characteristics of this data are summarized in the following section.

In addition, we examined existing portals that have been developed at the NIH and compared and contrasted the functionality. Many common features found in these portals and repositories will be adapted for inclusion in the CSSI DCC. Data collections examined included:

1. The Cancer Genome Atlas (TCGA)
   https://tcga-data.nci.nih.gov/tcga/
2. The Cancer Proteome Atlas (TCPA)
   http://app1.bioinformatics.mdanderson.org/tcpa/_design/basic/index.html
3. Library of Integrated Network Based Cellular Signatures (LINCS)
   http://lincs.hms.harvard.edu/db/datasets/
4. The Immunology Database and Analysis Portal (ImmPort)
   https://immport.niaid.nih.gov/immportWeb/display.do?content=AboutImmPort
5. NIH Microbiome Project
   http://hmpdacc.org/micro_analysis/microbiome_analyses.php
6. NCBI BioProject
   http://www.ncbi.nlm.nih.gov/bioproject/
7. Office of Cancer Clinical Proteomics Assay Portal (CPTAC)
   https://assays.cancer.gov/
8. NICHD Data and Specimen Hub (DASH)
   https://dash.nichd.nih.gov/

From the review of the listed sites and a combination of brainstorming sessions within the project team and discussions with external stakeholders, we developed and vetted a list of use cases of proposed features that can be grouped around the major functionality of the site. In addition, a strong consideration of the ISA- Investigation/Study/Assay (ISA-Tab) format and its associated validation tools was considered as an approach to data format and metadata validation (http://isatab.sourceforge.net/format.html). ISA-Tab is being considered for use due to its ability to incorporate assay and study information into the metadata for a dataset. This will allow users of the DCC to not only view the data, but understand how the data was generated.

A glossary of terms (Appendix 1) was created to ensure common usage of technical and descriptive terms and processes that will be used in the subsequent software development phase.

Data Inventory and Characterization

The current data sets in the possession of FNLCR, and that expected from future CSSI (and PS-ON) contracted projects were reviewed prior to planning and development of prototype software to store cancer research related data. These data sets, as listed in the table below, were characterized with respect to data types, file types, file size, and the metadata information describing the data and the

experiments used to generate the data. Pending or prospective data from projects not yet under way is included for certain projects; projected dataset size is approximate.

Table 1 Data Characterization Summary

YT	Project Description	Data Type	File Type	Size	Potential Metadata
11-212NS	Physical Characterization of Parameters in Biospecimens (CTC phase 1)	Cell images, spreadsheets of cell number, relative fluorescence of selected biomarkers and blood draw information	jpg xls txt	< 1 TB	Imaging settings, Instrument name and version, Biomarker staining intensity Clinical data mapping to biospecimen (non PII)
11-225	Genomic Analysis of PS-ON Cell Lines	Sequencing: miRNA, exome, mRNA	Fastq Bam Vcf Txt html	<1 TB	Cell line, Passage number, Passage conditions, Instrument type and files, Reference sequence, probe sets
12-129NS	HCS of Physical Based Properties in Biospecimens Phase 2	cell images and spreadsheets of cell number, relative fluorescence of selected biomarkers and blood draw information, cfDNA	TBD	TBD	Imaging settings, Instrument name and version, Biomarker staining intensity Clinical data mapping to biospecimen (non PII)
12-138	Physical Based Properties of PS-OC Cell Line Panel	Cellular images	Tif Jpg Xls Txt py	~1.0TB	Cell line, Passage number, Passage conditions, Imaging settings, Instrument name and version, Calculation methods
13-125NS	ImmunoMRM Assay Pilot	Mass spec data	sky Wiff t2d qgd spc mzXML raw PSM prot	~ 1.0 TB	Instrument files, Peptide spectrum and match, protein assembly and relative abundance, checksum
13-134NS	CaSIX	Physical characterization data, targeted genomic data, targeted proteomic data from human colon cancer xenografts	TBD	TBD	TBD

YT	Project Description	Data Type	File Type	Size	Potential Metadata
14-083	Proteomic Characterization of PS-OC Cell Line Panel	Mass Spec data	sky Wiff t2d qgd spc mzXML raw PSM prot	~1.0 TB	Cell line, Passage number passage conditions Imaging settings Instrument name and version Peptide spectrum and match, protein assembly and relative abundance, Phosphorylation sites
14-107NS	Thrombosis in Cancer Patients	Elisa and anticoagulant assays	Txt Xls or tsv	<1 TB	Calculation methods, normalization, Clinical data mapping to biospecimen
14-152NS	Prospective Biospecimen Collection Phase 2	Mass spec, proteomic	raw mzML PSM prot meta	~ 8 TB	Clinical data files, mapping of biospecimens to iTRAQ labels (where applicable), folder and file naming conventions

Proposed Use Cases

A use case is narrative text that describes an interaction of the user with the system, focusing on the value the user gains from the system. Typically they are written in the form of: As a [user role] I want to [goal] so I can [reason]. For example: As a [cancer biologist] I want to [browse available data sets] to [evaluate high content image data]. For the report, the use cases are presented with the feature that desired (goal) and the rationale (reason) for the feature.

A list of use cases that describe the major functionality for a data-sharing environment was developed to refine the requirements for the DCC software and infrastructure prototype that will be developed in Phase II of this project. These use cases describe how to store, view, and annotate within the repository and how to distribute the data through download. They further describe how the different types of user will interact with the data and the site to achieve a certain goal.

The major categories of functionality described are:

I. Portal
   a) Data Download
   b) Data visualization
   c) Search/Query
II. Data Upload
III. Data Integration/mapping
   a) Data Validation and Format control
IV. Documentation
   a) Context specific help
   b) End User and Training Documentation
V. Site Administration and Dashboard

Both the data inventory and the use cases will provide foundation for the development team to understand and refine the requirements to incrementally create a fully featured DCC during subsequent phases and provide a sound foundation for technical estimation needed to create the software development plan.

The complete list of use cases is included in the Appendix 2.

Table 2 Use Case Categorization and Prioritization Summary

Primary Category	Importance	Count
Administration	1	7
	2	10
	3	4
Administration Total		21
Documentation	1	9
	2	2
	3	1
Documentation Total		12
Integration/mapping	1	17
	2	15
	3	13
	4	11
Integration/mapping Total		56
Portal	1	12
	2	22
	3	13
	4	12
Portal Total		59
Upload	1	2
	2	1
	3	1
Upload Total		4
Grand Total		**152**

General Recommendations for Use Case Prioritization

In order to ensure a mature vision for the proposed software, a diverse set of stakeholder voices was considered to develop as inclusive a set of proposed functionality as possible. The list produced was longer than could be completed in the proposed period of development for the prototype; therefore not all features envisioned will be included in the initial prototype.

I. Use cases were created and categorized based on how the team envisioned the research community would want to interact with the underlying data; many of the cases have substantial

overlap with functionality in the list of repositories used as models. Stakeholder vision should be continuously validated through frequent contact throughout the project.

II. The Customer vision for future support and maintenance beyond the prototype should be discussed prior to the commencement of development because many features envisioned have a human support component that cannot be overlooked. Many of the sites that were analyzed for comparison have full time staff associated; for example TCGA has 3-4 FTE who assist investigators submitting data and curate that data upon submission.

III. The initial prioritization of Use Cases included in this report should be confirmed they will be used as input to the planning for the phase II software development planned for early in FY16.

a. The initial prioritization was completed by the project team and each use case was given as score by the team from 1 to 4 using the following scale:

1 = The DCC must have the feature
2 = The DCC should have the feature
3 = The DCC could have the feature
4 = The DCC will not have the feature

b. The scoring is based on the assessment of the project team during phase I. The score itself does not indicate that the feature will be present in phase II, however the score will assist in prioritizing features for the phase II team.

c. The priorities of use cases in each category are summarized in Table 2 with each use case and its score available in Appendix 2.

Preliminary Design Considerations

The technical approach used for the CSSI DCC will require a complex balance between software and hardware concerns because decisions made during the design phase may be difficult to undo at a later date. This balance was considered when prioritizing the desired functionality.

I. An Agile development approach is recommended because it mandates the frequent assessment of the features by stakeholders throughout the development of the system. This approach leads to increased stakeholder satisfaction and reduces the overall risk to the project by ensuring that evolving requirements stay true to the end users' needs.

II. The DCC is fundamentally a storage destination for a multitude of data types, where all data is properly annotated so that a user may search data sets. Certain assumptions must be made about storage, bandwidth and availability for uploading, downloading and ensuring that the data is persisted properly within the data store. A robust storage and hosting solution needs to be carefully considered to ensure expandable capacity, high availability, acceptable performance and adequate system security. The server virtualization, storage and network infrastructure at

the FNLCR is well equipped to provide the basic hosting needs for the DCC, and FNLCR staff will be consulted during the prototyping phase to assist with configuration of the purchased system.

Infrastructure issues will also drive cost projections of the proposed system; one area for future consideration will be the desire to have an offsite backup. Any approach will need to consider the overall cost for backup of ~50TB (or more) and if off-site backup is desired. The approach must also adhere to the data retention and disaster recovery policies in place at the FNLCR. However, offsite backup is not required for phase I due to the system size and the hardware purchased, as well as the fact that the data will still be housed in several other DCCs initially.

III. The DCC requires an administrative layer built on top of the data storage and retrieval layer. This administration layer must provide for basic validation, notification and audit infrastructure to ensure data movement and changes within the system are adequately traced. The exact workflow the system must support for each user type will be determined during further design.

IV. The data staging and portal components of the DCC are the most complex. Built upon the data storage and made available through the administration layer, both horizontal and vertical scalability must be built into the system. This requires careful evaluation of the data storage, housekeeping, memory and bandwidth usage to ensure compactness of the DCC over time.

V. Due to the potential number of data types, modern database technologies must be evaluated in order to ensure maximum flexibility of the DCC. Phase II will not be limited to evaluating a fully custom approach. Open source solutions such as LabKey Server will be considered, as well as some NCI and NIH sponsored DCCs. Examples include TCGA's enhanced portal and NCBI's BioProject. Such offerings will be reviewed for their coverage of our DCC use cases, as well as their ability to import and export data over APIs. Levels of customizations will also be reviewed. The chief criterion for consideration is their pre-existing flexibility in handling varieties of scientific data types.

VI. An evaluation of a blend of an open source solution with custom development to fill in gaps or extend functionality will be made to determine the most efficient and cost effective approach to achieve the desired functionality. Extensive customization of existing software may prove more cost prohibitive than an all custom solution. Prototyping across several approaches will ensure that the selected approach can be delivered on time and within budget.

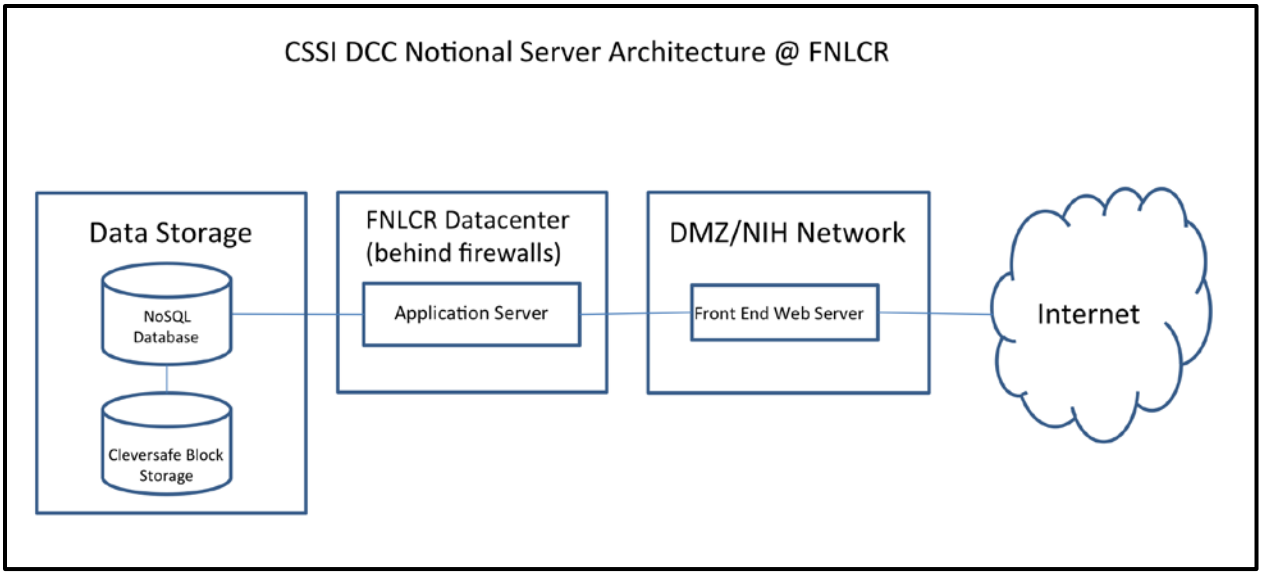

Figure 2 Proposed Architectural Diagram

Hardware

DCC hardware to house the prototype was purchased in FY 2015 as part of a composite purchase to support high performance computing that is to be housed at the ATRF: 50TB of Cleversafe Object Storage capacity. Cleversafe storage provides a highly reliable object-based storage solution, with the capacity to grow to multi-petabyte capacity. Object Based Storage systems (OBS) have emerged as the right choice for balancing scale, complexity and costs. Cleversafe storage uses Representational State Transfer (REST) interfaces to applications, has the capability to encrypt and encode on write. The erasure coding protects the data against drive, server, rack and site failures and is more efficient than data protection schemes that combine Redundant Array of Inexpensive Disks (RAID) and replication. The erasure coding approach which makes multiple slices of the data on the storage system provides the failsafe access that would be given a traditional backup. The slices are written across multiple devices in the system so that a single device failure does not compromise the data. The Cleversafe technology also allows the data to reside in multiple copies on multiple devices across multiple sites which gives it the high reliability rate.

Object storage was selected over traditional transactional storage because it allows relatively inexpensive, scalable, and self-healing retention of massive amounts of unstructured data providing the flexibility that will be required to house the volume of biologic data envisioned for the CSSI DCC as it grows over time. Cleversafe also allows rolling updates to maximize system up-time, provides built-in data integrity monitoring and reduced data loss compared to traditional RAID 6 storage configurations, RAID is only capable of controlling up to 64 terabytes of data before an organization begins to experience loss of data and bit errors. Finally, data are randomly encrypted prior to dispersal across multiple nodes to ensure data is secure in a system breach. Examples of the use of object storage include such diverse data collections as the storage of photos on Facebook, songs on Spotify, or files in online collaboration services, such as Dropbox.

Appendix 1: CSSI Data Coordinating Center Glossary

Term	Definition/Description
Accession data	To enter new data into the data store
Agile	An iterative method of software development where features are augmented over time until a mature application is developed. Its focus is repeated incremental deliveries of working software
Assay	An investigative method used to measure characteristics of a sample
Assay result	A measurement resulting from an assay
Biospecimen	A biological sample used in an assay
Class Identifier	A way of indicating data objects of the same type, could be a tag or field in the metadata
Clinically-Associated	Pertaining to a specimen with direct association to a patient interaction (ie- tumor sample, blood sample)
Clinically-derived samples	Samples immediately derived from patients
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CTC	Circulating Tumor Cell
CSSI	Center for Strategic Scientific Initiatives
Data Curation	The active and on-going management of data through its life cycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time
DCC	Data Coordinating Center
DSE	Data Sharing Environment
Data	The assay results from the experiments (primary) and resulting files after normalization or processing (derived)
Database	An organized collection of data and metadata
Data coordination framework	A model used to compile the data and meaningfully associate it with related data
Data dictionary	A collection of descriptions of the data objects within a data store that may also contain a mapping of data elements from one data set to another
Data table	A set of data whose values are represented with descriptions in vertical columns and individual entries in horizontal rows
Data model	A method of representing how data elements relate to one another within a datastore
Derived data	Data which does not represent an assay result, but rather is derived from an assay result or from other data (i.e. normalized or processed data)
Data store	Any repository of data objects , can be physical- such as a filing cabinet, or electronic-such as a database
Genomic	Pertaining to DNA or RNA sequences of a given sample
ImmunoMRM	Multiple Reaction Monitoring using Mass Spectrometric quantification of peptides in a mixture isolated by antibody purification
ISA	Investigation/Study/Assay- a data standard used for the characterization of molecular data
Metadata	The information that describes the raw or processed data in the data store. This may include, but is not limited to, clinical, experimental or processing information.
Normalized Data	Data which has been adjusted so that it meets standards across multiple data sets
Portal	A web interface for users to access data and associated files in the data store
Proteomics	Pertaining to the protein sequences and structure and function found in a given sample
Provenance	The origin of a sample or of data
PS-ON	Physical Sciences Oncology Network (previously known as PS-OC, Physical Sciences in Oncology Center)
Query	A defined method for data retrieval from a database

Term	Definition/Description
Report	A pre-configured, displayable set of data returned from a query against a database
Stakeholder	Someone who has an interest in the system under consideration who may or may not directly access it
Test case	An explicit proposed scenario, intended for use by software developers and testers, consisting of itemized steps describing the interaction with an interface and the desired outcome of those steps
User	Someone who interacts with the system under consideration
Use Case	A set of user stories that when combined describe a larger system goalist of actions or goals that define the interactions between the user and the system
User Story	A description of functionality that includes the type of user, how they will interact with the system and for what purpose
Wet Lab	Lab where biological or chemical samples are manipulated.
Validation	The process of verifying that file contents and format are consistent with a set of format and content specifications

Appendix 2: List of Use Cases

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Administration	Documentation	Everyone	want be able to see updates for bug fixes related to the DCC and its datasets.	so I know when problems are solved and can proceed with my work	1
Administration	Documentation	System Admin	want to be able to view usage reports through the portal over a period of time or per center/project	I can assess the change of usage over time to plan for future needs in terms of disc space or software needs	1
Administration	Documentation	System Admin	want to be able to view usage reports through the portal over a period of time or per center/project	I can assess number of centers and individuals downloading data, and assess what studies are used the most	1
Administration	Documentation	Everyone	want to be able to track the progress of fixing bugs related to the coordinating center and its datasets	so I know when problems are solved and can proceed with my work	1
Administration	Portal	System Admin	have an access control policy in place for all the hosted data	to ensure that personally identifiable information is kept from unauthorized users	1
Administration	Portal	System Admin	have access control for the uploaded documents	assign different levels of access to all entities involved in DCC	1
Administration	Upload	System Admin	want analysts and developers of the coordinating center to have access to a minimum of disk storage, CPU and access to the database and files of the coordinating center	analysts and developers of the data coordinating center can efficiently run scripts, programs and analysis of production data and pending data.	1
Administration	Documentation	End User	want to be able to view user feedback such as bug reports and user suggestions	know about and get updates about reported problems and suggestions for the interface or datasets	2
Administration	Documentation	System Admin	want to be able to view usage reports through the portal over a period of time or per center/project	provide monthly progress reports	2
Administration	Documentation	System Admin	want to be able to categorize bugs and data issues	so I can detect trends in problematic issues, and be proactive in their remediation	2
Administration	Documentation	End User	be able to view user feedback such as bug reports and user suggestions	know about and get updates about reported problems and suggestions for the interface or datasets	2

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Administration	Interface	System Admin	Create users and user groups	so that I can assign different permissions to them	2
Administration	Interface	System Admin	track log-ins and data changes	so I can do audit tracking and roll backs	2
Administration	Portal	Everyone	create or register for an account for myself	so I can access restricted features of the DCC	2
Administration	Portal	End User	save my searches associated with my account	so I can return to favorite or common searches	2
Administration	Portal?	System Admin	create roles and permissions	so that I can separate out access based on logged user	2
Administration	Upload	System Admin	want to categorize and enumerate the most common validation failures including in terms of what center (submitter) or data sets are failing (audit trail on the submission action)	be proactive in making changes that will avoid failures in the future data submissions	2
Administration	Integration/mapping	End User	track all the changes made to the data set or the content	be able to access the previous or a specific version or be able to obtain details of the changes	3
Administration	Upload	System Admin	want to be confident that only trusted users have access to protected files	data submitters can share protected files with colleagues, but without public access	3
Administration	Upload	submitter	have options for who can see my data set	share my data set before it is publicly accessible.	3
Administration	Upload	submitter	want various ways to access upload validation reports, such as through email, through an API	I can select the most convenient method to review my validation reports	3
Documentation	Documentation	Everyone	Know the definition of terms and vocabulary being used at the DCC including: data types, metadata, ontologies, copy number variation etc. and ensure that they are being used consistently.	plan to use those controlled vocabularies that will allow my data set to be appropriately labeled and become integrated with other studies -- commonly used terms will not be taken out of context and misused.	1

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Documentation	Documentation	Everyone	want terms which classify data within the DCC to be different from the commonly used adjectives for that data	when discussing data and how it is categorized, there will be no ambiguity as to whether context of a term being used is the data's category or a mere description of the data. For example, if the DCC assigns data to categories of: primary data and derived data, and subsequently, we find that dataset X was originally categorized as primary data but actually had been normalized, we need a way to say that dataset X is not primary data, even though the DCC had categorized it in the category named primary data.	1
Documentation	Documentation	Everyone	The user manual to include: the definition of terms used at the DCC to be organized in a glossary; description of submission, access and QC including how datasets were validated; navigating the portal	I can look up terms easily because they are sorted alphabetically perform my own QA and understand the QA that was performed on the datasets	1
Documentation	Documentation	Submitter	The user manual to include: documentation describing submission, access and quality control at the DCC; file formats and specifications; example submissions	my submission can be efficient and consistent with the setup at the DCC my upload will pass validation	1
Documentation	Documentation	System Admin	have a set sample test cases which represent a variety of valid and invalid data submissions	can test new software implementations to see if legacy data can be processed properly with the pending software changes and detect errors in submissions that do not pass file and format specifications	1
Documentation	Documentation	System Admin	have access to an itemized list of requirements for a data coordinating center	use the DCC requirements as a basis for planning new projects	1
Documentation	Documentation	Everyone	have access to definitions for terms, and have these terms be consistent with their use within the DCC	minimize ambiguity in communications	1

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Documentation	Documentation	submitter	review example submission formats (from a help page/user guide)	efficiently plan my data submission format	1
Documentation	Portal	Everyone	help section with Tutorials/training manuals, Glossary, FAQ, policies and regulations etc.	maneuver the DCC	1
Documentation	Administration	End User	track all the changes made to the data set or the content	be able to access the previous or a specific version or be able to obtain details of the changes	2
Documentation	Documentation	System Admin	have a simplified method to keep track of requirements for the DCC	the documentation of tracking requirements does not become unwieldy	2
Documentation	Documentation	End User	a link to the website with future opportunities for funding, publications, etc.	get relevant information for new projects	3
Integration/mapping	Portal	End User	want a checksum or md5 file for my file downloads	confirm that my download was successful by comparing the size of my download with expected size of download	1
Integration/mapping	Portal	End User	have a search function across data sets	to allow me to find relationships in the data	1
Integration/mapping	Portal	End User	ability to search data by typing in keywords or metadata terms.	obtain reports or data for further analysis	1
Integration/mapping	Upload	submitter	have a formal way to represent a machine readable data standard that describes an experiment (For example as provided by the use of ISAtab)	describe the steps of how the data was determined, the experiment was performed and how each sample was processed	1
Integration/mapping	Upload	submitter	add metadata at either the study or experiment level	associate the attributes at the right level	1
Integration/mapping	Upload	submitter	tag my experiment or study with identifiers for attributes	(ie protein, genomic, physical characterization, imaging) so that I can group like experiments and studies	1
Integration/mapping	Upload	Everyone	want well-defined specifications for genome annotations which is consistent across all datasets	I know what human genome assembly was used when genome coordinates are provided, I know what NCBI gene ID is being referred to when gene symbols are provided, etc	1

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Integration/mapping	Upload	submitter	verify(test) that the format and contents of my submission are valid before upload, for example, through a distributed validator that I can run on my computer before submission	know of any inconsistencies of my submission such as inclusion of unintentional files, misnamed files, format problems, deviations from controlled vocabularies before I go to the trouble to upload them.	1
Integration/mapping	Upload	End User	have English text description of column headers in tab delimited text file	know what columns represent in tsv files	1
Integration/mapping	Upload	End User	want fields and their values to not include carriage returns	a set of variables can be represented on one line of text	1
Integration/mapping	Upload	End User	want fields and their values to not include HTML tags	I don't have to edit the values provided in fields to get text only	1
Integration/mapping	Upload	End User	want text files to be in tsv rather than csv format	so that fields can include commas without hindering the parsing of data	1
Integration/mapping	Upload	End User	I want samples to have identifiers that make it easy for me to look at the identifier and know something about the metadata of the sample. For example OvCa3 is a sample from ovarian cancer.	I can easily know if something is not right, for example, if I think I have samples from one tissue and see an identifier representing another tissue. And so I easily know if the sample is of interest to me.	1
Integration/mapping	Upload	End User	samples to have unique identifiers that map to their parent samples, which also have unique identifiers	retrieve datasets representing identical samples or parents of samples	1
Integration/mapping	Upload	End User	give my sample an alias	use own identifiers but still meet the standard	1
Integration/mapping	Upload	End User	files in data submissions that represent text only should be in machine readable format (text file) rather than Word or PDF, and files in data submissions that represent columns of values should be submitted as tab delimited text files, instead of excel spreadsheets.	use UNIX and parsing scripts such as Perl,Python and R to access the data and text, and at the same time, can always use excel to open the file	1
Integration/mapping	Upload	End User	have data sets where the flat files are free of non-ASCII characters	use UNIX and parsing scripts such as Perl,Python and R to access the data and text.	1
Integration/mapping	Integration	End User	have access to the associated reference data sets for a study	can integrate probe sets, gene sets and other data with study data	2

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Integration/mapping	Portal	End User	be able to download data sets based on what metadata they are mapped to- such as cell line name or ID	retrieve those data sets that match my research interests	2
Integration/mapping	Portal	End User	be able to browse metadata and have access to files that match my selections	be able to download data sets that match my research interests	2
Integration/mapping	Portal	End User	ability to represent the my query as a URL	I can share my query with a colleague by including it in an email	2
Integration/mapping	Portal	End User	access the quantitative data from the images stained with particular antibody/antibodies across samples of interest	compare the expression levels of markers across the samples of interest and controls	2
Integration/mapping	Upload	Everyone	want metadata to use a non-redundant controlled vocabulary, for things with identical meaning, for example, I want 'Male' to be represented by 'Male' and not a variety of terms including: 'Male', 'MALE', 'male', 'M, m', 'XY' and '1'.	can reliably use DCC metadata to combine studies	2
Integration/mapping	Upload	Everyone	have metadata include controlled vocabularies that are mapped to established ontologies	integrate my data and labels with the data from the other collaborators	2
Integration/mapping	Upload	End User	want data sets to be mapped to metadata from established, known ontologies, rather than metadata invented by the DCC	1.understand how the DCC came up with their metadata categories and terms 2.integrate data sets from the DCC with other online resources that used established ontologies	2
Integration/mapping	Upload	submitter	be able to submit the metadata for my submission as a flat file included in the submission. The DCC has software that parses the metadata from the flat file.	submit a data set without having to fill out online forms for each entry, and so that I can avoid errors in filling out forms	2
Integration/mapping	Upload	Everyone	have an enforced case-insensitivity for data entry but for download have one consistent case	I don't have to reconcile upper and lowercase formats	2

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Integration/mapping	Upload	End User	have text files where \n is used between lines instead of \r	so that I can use UNIX commands such as wc to know how many lines are in a file	2
Integration/mapping	Upload	End User	no empty spaces, brackets, parenthesis and special characters in directory and file names	use UNIX or other appropriate platform to access the data	2
Integration/mapping	Upload	End User	have two fields for reporting values, the unit of measure and the value of that unit	don't have to parse the unit name from the value name from a reported value.	2
Integration/mapping	Upload	End User	know what file format and version each file is	don't have to open a file, or guess from its name in order to know what format it is (need to standardize file names)	2
Integration/mapping	Upload	submitter	uploading links to literature and publications relevant to the research projects	have all the relevant information for the project in one place for others to review	2
Integration/mapping	Portal	End User	have links to articles published with the data	so that I can see conclusions previously drawn	3
Integration/mapping	Portal	End User	have an ad hoc search tool that allow me to do complex searches (and/or not, if, nested)	so I can determine if there are data available that match different stringencies of description	3
Integration/mapping	Portal	End User	want the ability to query the contents of a file for example, from a VCF file: select lines based on:1) genomic information such as gene or locus, 2)genomic function such as transcript, exon, UTR, miRNA, 3) genomic effect such as frame shift, missense, nonsense.	I can retrieve lines from a file that pertain to my research interests, for example, as a geneticist wanting to retrieve lines from a VCF file.	3
Integration/mapping	Portal	End User	want the ability to query based on file-specific metadata including but not restricted to:1) genomic information such as gene or locus, 2)genomic function such as transcript, exon, UTR, miRNA, 3) genomic effect such as frame shift, missense, nonsense.	I can retrieve genomic data that pertains to my research interests as a geneticist.	3

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Integration/mapping	Portal	End User	generate a query to get the data with expression levels of proteins of interest across specific samples	compare the expression levels of protein markers across the samples of interest and controls	3
Integration/mapping	Portal	End User	want the user interface to be able to map from metadata describing a chromosome locus to a gene, and from a gene to its chromosome locus	can use a gene ID to retrieve files that match the locus for that gene, and I can use a locus to get files that match the genes that have that locus.	3
Integration/mapping	Portal	System Admin	have the DCC make use of off the shelf tools such as LabKey	I don't have to design an interface from scratch, and so the DCC can be easily integrated with other online instances using the same open source software.	3
Integration/mapping	Upload	submitter	want a mechanism to provide annotations that are outside of the scope of those in the standard vocabularies and ontologies	if I have an unusual annotation, I will have a mechanism to include it in my submission	3
Integration/mapping	Upload	submitter	submit files associated with my publication	I can share my files with the research community and also include links to my entry at the DCC in my publication.	3
Integration/mapping	Upload	submitter	uploading publications as a result of the uploaded and analyzed data	have all the relevant information for the project in one place for others to review	3
Integration/mapping	Upload	End User	have unique column headers per file (note-ISA Tab will be an exception)	columns can be resorted within a file without losing the meaning of the contents. (NOTE, an exception is ISA tab, column order is important)	3
Integration/mapping	Upload	End User	want the metadata of a study and its files to include: author, submitting center, institution, project, experiment type, file type and disease	efficiently retrieve data sets that match my queries per these fields	3
Integration/mapping	Upload	End User	want to know what samples were dropped a study and why	I can understand and review the steps of the study	3
Integration/mapping	Integration	System Admin	I want sample identifiers to be devoid of metadata	When there is a sample swap, and all the metadata was wrongly assigned, we can update the metadata of a sample without changing the sample ID.	4

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Integration/mapping	Portal	End User	access images stained with particular antibody/antibodies across samples of interest	compare the presence or absence of staining of antibodies across the samples of interest	4
Integration/mapping	Portal	End User	want to download and mirror an entire dataset	so that I can perform integrated analysis on the dataset	4
Integration/mapping	Portal	End User	download the updated data sets of a study associated with a publication as one download, as opposed to downloading the original data sets	use advances in past research results to perform my own research more efficiently	4
Integration/mapping	Portal	End User	generate a query to get the data with genes of interest down-regulated across specific samples with fold differences, the parameters for setting up the baseline (either using a control sample/s or using house-keeping genes) will be determined by me	compare the fold differences in the specific gene expression levels across samples of interest	4
Integration/mapping	Portal	End User	generate a query to get the data with genes of interest down-regulated across specific samples, the parameters for setting up the baseline (either using a control sample/s or using house-keeping genes) will be determined by me	compare the expression levels of specific genes across samples of interest, perform pathway analysis	4
Integration/mapping	Portal	End User	generate a query to get the data with genes of interest up-regulated across specific samples with fold differences, the parameters for setting up the baseline (either using a control sample/s or using housekeeping genes) will be determined by me	compare the fold differences in the specific gene expression levels across samples of interest	4
Integration/mapping	Upload	submitter	create an experiment	to associate with a study to house my data	4
Integration/mapping	Upload	End User	want to be able to select data sets and images that I can annotate on my own, for example, if i find an image that is interesting	I can share my expertise regarding a dataset or image without needing to download everything and make my own website just to share my point of view	4

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Integration/mapping	Upload	submitter	navigate (magnify etc.) and annotate images such as histology images, CT scan images	upload the annotated images for collaboration and submission	4
Integration/mapping	Upload	End User	I want to be able to construct my own identifiers for samples, constructing sample aliases, using values for their metadata at a point in time	my sample IDs are a reflection of the metadata that I find important	4
Portal	Documentation	Everyone	have access to an online User Guide	have one central location where I can keep up to date on file format specifications, upload procedures as well as how to access data set submissions	1
Portal	Documentation	End User	want a diagram of how data is uploaded, processed and downloaded by a final user of the data	so I can quickly understand the flow of data from the submitter, into the system and to the end user	1
Portal	Documentation	End User	to be provided a concise overview of the DCC	can quickly decide if the DCC overlaps with my interests	1
Portal	Documentation	System Admin	have a concise description of a study's representation at the DCC and anticipate file submission	assess the completion of a study.	1
Portal	Documentation	End User	want to be able to report bugs or errors	so that they can be remedied	1
Portal	Download	End User	download all files for a study/experiment as a single file, for example a zip or tar file	don't have to click on various links on a browser to get a data set, and so that I know I have everything for a data set (user must set the level at which all associated files get selected, not at the individual file level)	1
Portal	Portal	System Admin	have a user interface and documentation that passes 508 compliance	pass government requirements for accessibility	1
Portal	Upload	submitter	have version control assigned to my upload	communicate with collaborators about which revision of data they are using	1
Portal	Upload	submitter	upload huge data sets using the internet	upload and update my data sets without needing to FedEx a hard drive to the DCC	1

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Portal	Upload	submitter	upload one or more than one file in a single submission	so that I can associate all files to a submission	1
Portal	Upload	submitter	want to verify a checksum or md5 file for my file uploads	I can confirm that my upload was successful	1
Portal	Upload	Everyone	verify that the format and contents of my submission are valid after upload but before making public (positive acceptance)	any invalid submissions will not be made public or completely processed	1
Portal	Documentation	End User	know if there exists a detailed description of what protocols and controls were used in each assay	can review the protocols and assess if the protocol is substantiated	2
Portal	Documentation	End User	have a concise description of file formats, and version numbers, available per study	know ahead of time if the data set has files that are of interest to me	2
Portal	Documentation	End User	want a Help button or Tips on the webpage	so I can orient myself on the web interface	2
Portal	Documentation	System Admin	easy access to sample submissions to be used as test cases in software development	can use scripts or programming to access the sample data sets, as opposed to needing to download them from a cloud or use a GUI for access. (Agile CI test process to be used?)	2
Portal	Documentation	End User	have DCC genomic data available as BAM, GFF or BED format or other standard format	I can upload data sets to a genome browser such as UCSC/JBrowse or my own genome browser	2
Portal	Integration	Everyone	link to the associated protocol or SOP used for the processing and analysis of the samples	access the specific protocol and or SOP for the specific sample or data	2
Portal	Integration	End User	want items represented on the interface to include links to associated items	can navigate from a study, to the samples for that study and *vice versa*, ie from a display of sample information to the entry for the study that the sample was used in	2
Portal	Integration	End User	want items represented on the interface to include a method to download the displayed information as machine readable format	I can retrieve the information displayed on the interface without having to strip HTML tags or other formatting information that is external to the content.	2

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Portal	Integration	End User	want items represented on the interface to consist of ASCII characters	I can cut and paste from the interface and not have to worry about picking up non-ascii characters, for example, non-ascii representations of quotes or hypens	2
Portal	Integration	End User	want a notification when new data sets are available	so I can update my research in a timely manner.	2
Portal	Integration/mapping	End User	want my downloads to include a file with a date	so know the exact query and the date I ran the download	2
Portal	Interface	End User	programmatically access the database through a standard format/api	exchange and pull data	2
Portal	Interface	End User	want the ability to query and retrieve data sets programmatically, without using a browser	I can write scripts to retrieve my data sets	2
Portal	Portal	End User	want an efficient way to access results of commonly used queries	I don't have to learn how to construct queries on my own	2
Portal	Portal	End User	to save the query that generated my downloaded dataset	share with others my method used to retrieve data, as well as redo the same query at a later time.	2
Portal	Upload	submitter	replace one file in a previous submission	update a file in a submission without having to update an entire submission	2
Portal	Upload	submitter	remove one file from a previous submission	update a file in a submission without having to update an entire submission	2
Portal	Upload	submitter	upload a brand new file to be included in a previous submission	update a file in a submission without having to update an entire submission	2
Portal	Upload	submitter	upload data sets using a web interface	don't need to know UNIX to deposit data	2
Portal	Upload	submitter	receive notification that my data has been uploaded successfully	so that I know the DCC receive the data	2
Portal	Upload	submitter	receive notification that the data has been validated (is compliant) and will be made available to users	so that I know my data met the correct standards and is available for use in the DCC	2
Portal	Upload	submitter	have a error message or notification that data I am trying to upload doesn't meet format requirements as it's uploading	so that I do not have to wait for an entire upload to complete (if its large) before being notified that the format is incorrect	2

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Portal	Documentation	End User	want a visual representation of the experimental pipeline for each study	can understand a complicated experimental procedure	3
Portal	Download	End User	I want to download datasets incrementally, as opposed to one huge download	so that I won't lose everything if my computer crashes in the middle of a huge download	3
Portal	Download	End User	I want to download updates to a dataset, instead of an entire dataset	I don't have to download an entire dataset just to get the new files	3
Portal	Integration	Everyone	group the samples by the collection criteria such as time of collection, method of collection, preservation method, storage details, processing details	perform quality metrics across the samples of interest	3
Portal	Integration	End User	want a preview of my query selection before download, including a size estimate	I have an idea of how large my query is, and I can decide to be more stringent, or less stringent in my query criteria.	3
Portal	Integration	End User	have a visual and familiar representation of genomic data, for example, as provided by the genome browser JBROWSE	I can download annotations that pertain to regions of the genome that I am interested in, as well as visualize trends across a chromosome	3
Portal	Integration	End User	want to be able to be able to perform the same query at two different time points and immediately see if the results have changed	will know if there has been an update and that the results from my previous query is outdated.	3
Portal	Integration	End User	download all the original data sets of a study associated with a publication as one download, ie one dataset freeze, as opposed to downloading the study's current data set which may have been revised since the date of publication	use advances in past research results to perform my own research more efficiently	3
Portal	Portal	End User	want to be able to use Boolean Logic to combine query terms	I have flexibility to retrieve data sets that match my selected values for metadata.	3
Portal	Portal	End User	want to be able to save my query preferences	don't have to re-enter my query preferences each time I visit the DCC website	3

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Portal	Upload	submitter	have the ability to sign off/comment before releasing the finalized data (including metadata) reports	share the finalized data with collaborators and public	3
Portal	Upload	submitter	package metadata in the format of choice and as per specific requirements	share my data set with collaborators before making it available to the public	3
Portal	Upload	submitter	upload data via an api in an automated process	automate data upload and do real-time data acquisition	3
Portal	Download	End User	if the filetype is a huge file, and my query only pertains to a few lines in the file, I want the ability to retrieve only the relevant lines of a file	I don't need to parse relevant data from huge files each time I do a query.	4
Portal	Integration	End User	export charts and graphs created	use data from the repository in my research/work	4
Portal	Integration	End User	view returned query results as a graph or chart format	explore relationships within the data sets	4
Portal	integration	End User	save charts and graph created associated with my account	so I can easily return to my previous work	4
Portal	Integration	End User	uploading presentations relevant to the research projects and be able to work on it further	collaborate with other researchers	4
Portal	Integration	End User	developing new analytical and visualization technologies for different audiences to facilitate data analysis	establish collaborations to work on new projects, get help with study design	4
Portal	Integration	End User	generate a query to get the data with genes of interest up-regulated across specific samples, the parameters for setting up the baseline (either using a control sample/s or using house keeping genes) will be determined by me	compare the expression levels of specific genes across samples of interest, perform pathway analysis	4
Portal	Upload	submitter	create a study without adding data to it	create a placeholder for study	4
Portal	Upload	End User	want a place to upload data for other trusted users	so that I can collaborate with other trusted users without making our data public	4

Primary Category	Secondary Category	User	Use Case	Rationale	Importance
Portal	Upload	System Admin	ensure the uploaded data complies with rules and regulations (IRB approvals, FDA guidelines, HIPAA regulations etc)	control access of the data to all entities involved in DCC	4
Portal	Upload	End User	create/upload presentations resulting from the data from DCC	collaborate with other researchers	4
Portal	Upload	Everyone	data sets in Word documents, PDF, Excel spreadsheets	can export to personal drive for sharing with external researchers, for further analysis for presentation	4
Upload	Integration/mapping	Submitter	user should acknowledge that there is no PHI or PII in the data	HIPAA Compliance	1
Upload	Upload	submitter	upload data sets from the command line (without using a web interface)	automate data uploads	1
Upload	Administration	submitter	want to be able to validate my files with any new specifications before the specifications become a requirement and have them be accepted as valid submissions.	I can continue to upload datasets that do not follow new specifications while changing my pipeline to adopt the new specifications	2
Upload	Integration	submitter	want a lag time between the announcement of pending changes in specification requirements and enforcing of submission requirements	I can continue to upload datasets that do not follow new specifications while changing my pipeline to adopt the new specifications	3

Appendix 3: Data Characterization

YT	Project	File types	Other information	Approx. Size (TB)
13-125NS	ImmunoMRM	raw	(raw): The original mass spectrometry(MS) instrument files	1.0 TB
		mzML	(mzML): HUPO-PSI standard raw data files generated from the original MS instrument files	
		PSM	(PSM): Peptide-Spectrum Match data from the Common Data Analysis Pipeline (CDAP)	
		prot	(prot): Protein assembly data and protein relative abundance	
		meta	(meta): Clinical data files, mapping of biospecimens to iTRAQ labels (where applicable), folder and file naming conventions	
		Skyline (.sky, .view) files	Checksum files are included in all downloads for verification.	
		Zip (.zip) files		
		pdf files	Some of these are stored on Acronis Access and Panorama Dashboard	
		powerpoint (.ppt. .pptx)		
		Excel Files (.xls, .xlsx)		
		library files (.clib)		
		Mass spec files (.wiff, .t2d, .qgd, .spc, .mzXML)		
14-107NS	Thrombosis	Excel (.xls, .xlsx) files	ELISA output	<1.0 TB
		Text (.tsv)	Anticoagulant assay	
			Potential to have immunoMRM data in the future	
12-129NS	CTC II	TBD	High content analysis of circulating tumor cells	TBD
			Circulating free DNA from patient samples, quantity and sequence variation	
			Mass cytometry analysis of circulating tumor cells	
14-152NS CPTAC	CPTAC Study 1	.t2d .mgf .mzML .cal .cksum .XML .bin .txt .raw .wiff .mzXML	The Unbiased Discovery Working Group attempted to identify the components of the NCI-20 test sample, a mixture containing 20 human proteins, using different mass spectrometry experimental platforms. Three samples of intact NCI-20 (1A) and three samples of trypsin digested NCI-20 (1B) were sent to each laboratory during each of two successive weeks to identify the proteins using their own protocols with any available instruments. Instrument platform diversity was highest in this initial study.	0.1

YT	Project	File types	Other information	Approx. Size (TB)
	study 2	.raw .mzML .cksum	The Unbiased Discovery Working Group attempted its first use of a Standard Operating Procedure (SOP v1.0) in re-analyzing the trypsin-digested NCI-20 mixture at sites that had installed Thermo LTQ or Orbitrap instruments. Three "1B" samples were provided to each participating laboratory on two successive weeks. The use of an SOP (v1.0) controlled the rate of MS/MS acquisition but proved less effective in generating similar identified peptide counts across instruments.	0.01
	study 3	.t2d .mzML .raw .wiff .cksum	Significant changes from Study 3 include; a new *Saccharomyces cerevisiae* (yeast) reference proteome, the SOP version 2.0, and a bioinformatic infrastructure to collect raw data files and to identify peptides and proteins. Study 3 tested these tools in a small-scale methodology test. The Unbiased Discovery Working Group tested a new version of the SOP v2.0 for the complex yeast lysate sample introduced in this study. Each instrument was evaluated for SOP v2.0, specifying 184-minute RPLC analyses, in two replicates of NCI-20 followed by yeast lysate. For this study only, an Applied Biosystems QSTAR-Elite accompanied the LTQ and Orbitrap instruments.	0.034
	study 5	.raw .mzML .cksum	Study 5 extended upon Study 3 by revising the SOP to version 2.1 and adding a new sample: yeast sample spiked with BSA. The Unbiased Discovery Working Group probed the yeast sample in depth and evaluated the impact of spiking a small amount of BSA into samples. Each LTQ and Orbitrap instrument produced six RPLC analyses of yeast and six of BSA-spiked yeast, with NCI-20 samples present as QC mixtures. The study showed no negative effects from the spikes on other identifications, but the need for SOP v2.1 to specify flow rate was demonstrated by an outlier instrument.	0.06
	Study 6	.raw .mzML .cksum	Study 6 built upon Study 5 by including spikes of the Sigma UPS 1, a mixture of 48 human proteins in equimolar concentration. The Unbiased Discovery Working Group evaluated the sensitivity of spiked protein detection by evaluating samples of yeast with Sigma UPS 1 spiked at five levels. The yeast, Sigma UPS 1, and five spiked levels were each analyzed three times by RPLC, and NCI-20 samples were present as QC mixtures. The Sigma UPS 1 proteins were not detected at the lowest concentration, but each of the other concentrations generated UPS 1 detections, increasing with concentration. This Study reflects the final SOP v2.2 specification for LTQ and Orbitrap instruments.	0.121

YT	Project	File types	Other information	Approx. Size (TB)
	Study 7	.wiff .mzML .cksum	Executed by the CPTAC Experimental Design and Statistics: Verification Studies Working Group, was a multi-laboratory study designed to assess performance metrics of multiplexed, protein-based Multiple Reaction Monitoring (MRM) assays, including recovery, precision, and limits of detection and quantitation. Although individual laboratories have demonstrated that MRM coupled with stable-isotope dilution mass spectrometry has suitable assay performance for quantitative measurements of candidate protein biomarkers in plasma, reproducibility and transferability of these assays across multiple laboratories has yet to be demonstrated. CPTAC Study 7 was an advanced consortium-wide (8 laboratories participated) follow-up study to the initial "feasibility" Study 4. Study 7 encompassed a three-tiered experimental protocol that progressively introduced sample preparation variables likely to affect inter- and intra-laboratory reproducibility, transferability, and sensitivity. The experimental design varied from 1) spiking digested plasma with 10 signature peptides representing 7 target proteins (samples prepared centrally at NIST), 2) spiking digested plasma with the 7 digested target proteins (samples prepared centrally at NIST), and finally 3) spiking the 7 intact target proteins into un-digested, neat plasma followed by combined digestion (samples prepared at each participating site). In Study 7, using common materials and standardized protocols, it was demonstrated that MS-based assays of proteins in plasma can be sensitive and highly reproducible across laboratories and instrument platforms. Here we provide the Study 7 data files to be used as benchmarks against which individual laboratories can compare their performance and evaluate new technologies for biomarker verification in plasma.	0.008
	Study 8	.raw .mzML .cksum .txt	The Unbiased Discovery Working Group re-evaluated the yeast lysate sample without the requirement of an SOP. Each group was asked to produce three RPLC separations of the yeast sample at both high and low concentrations using their own protocols. The same quantity of yeast was analyzed in Studies 5, 6 and 8, with an additional 5x (high load) sample analyzed in Study 8. The LTQ and Orbitrap instruments generated a broad range of identified peptides as groups used their own lab-specific protocols.	0.026

YT	Project	File types	Other information	Approx. Size (TB)
	Study PTM	.raw .mzML .cksum	The Post-Translational Modification Working Group evaluated the reproducibility of a lectin-based glycopeptide enrichment and identification workflow on the conditioned medium of luminal and triple-negative breast cancer cell-lines. Three sites and two different instrument platforms analyzed lectin-enriched and PNGase-F treated aliquots from digested conditioned media (CM) by LC-MS/MS. The CM samples were initially trypsin-digested and aliquoted at a single site and then distributed to 3 sites. Each site separated the 10 CM samples, in duplicate, by chromatography using lectins Aleuria aurantia (AAL) and Sambucus nigra agglutinin (SNA), which generated 40 fractions. The samples were deglycosylated using PNGaseF and analysed in duplicate, yielding a total of 80 MS/MS data sets per site. The lectin affinity chromatographic enrichment and LC-MS/MS protocol can be accessed from here. N-deglycosylated glycopeptide identifications provided on N-glycosylation sites and N-Glycosites were compared between sites and breast-cancer derived samples.	0.144
14-152NS CPTAC	TCGA Colorectal Cancer Scientific Data	raw mzML PSM prot meta (raw): The original mass spectrometry(MS) instrument files (mzML): HUPO-PSI standard raw data files generated from the original MS instrument files (PSM): Peptide-Spectrum Match data from the Common Data Analysis Pipeline (CDAP) (prot): Protein assembly data and protein relative	Comprehensive evaluation of TCGA colorectal cancer tumors with primary instrument files and derived secondary data files compiled and presented in forms that will allow further analyses of the biology.	0.07
	Proteogenomics of Colorectal Cancer Nature 2014		Proteomes of colon and rectal tumors, previously characterized by The Cancer Genome Atlas (TCGA), were analyzed and integrated proteogenomic analyses were performed. Protein identifications in the format of IDPicker assemblies are provided for the 95 tumor samples along with the original mass spectrometry data.	0.73
	TCGA Ovarian Cancer		Two CPTAC Proteome Characterization Centers, Johns Hopkins University and Pacific Northwest National Laboratory, analyzed 174 ovarian cancer TCGA samples to characterize the cancer proteome. Complementary observations of the glycoproteome and phosphoproteome were explored in 122 and 69 of the TCGA samples, respectively.	0.78
	TCGA Ovarian Cancer CompRef Samples		Comparison and Reference (CompRef) control samples were analyzed to monitor the consistency of mass spectrometry instrument performance throughout the TCGA ovarian cancer proteome study. Five proteome and 4 phosphoproteome iTRAQ experiments were performed at Pacific Northwest National Laboratory. Six proteome iTRAQ experiments were analyzed at Johns Hopkins University.	0.112

YT	Project	File types	Other information	Approx. Size (TB)
	TCGA Breast Cancer	abundance (meta): Clinical data files, mapping of biospecimens to iTRAQ labels (where applicable), folder and file naming conventions Checksum files are included in all downloads for verification.	The CPTAC, TCGA Cancer Proteome Study of Breast Tissue analyzed the proteomes and phosphoproteomes of 105 TCGA tumor samples, these data include observations from each of the 4 breast tumor subtypes: luminal A, luminal B, HER2E and basal-like.	1.53
	TCGA Breast Cancer CompRef Samples		CompRef samples were analyzed in iTRAQ experiments along with the TCGA Breast Cancer sample iTRAQ experiments to monitor the consistency of mass spectrometry instrument performance. Proteome and phosphoproteome analyses were completed on two human-in-mouse xenograft reference samples, P5 (basal) and P6 (luminal).	0.276
	TCGA Colorectal Cancer		The goal of the CPTAC, TCGA Cancer Proteome Study of Colorectal Tissue is to analyze the proteomes of TCGA tumor samples that have been comprehensively characterized by molecular methods. Ninety-five TCGA tumor samples were used in this study.	0.736
	TCGA Colorectal Cancer CompRef Samples		Comparison and Reference (CompRef) control samples were analyzed to monitor the consistency of mass spectrometry instrument performance throughout the TCGA Colorectal Cancer and the Normal Colon Epithelium studies. A total of 32 interstitial CompRef measurements were made, 20 during the analysis of the 95 TCGA tumor samples and 12 during the analysis of the 30 normal colon samples.	0.231
	Normal Colon Epithelium Samples		Non-tumor, colon tissue samples (ascending and descending) were obtained from 30 patients. Each sample was analyzed with label free global proteomic profiling.	0.44
	System Suitability (CompRef) Study		The objective of the "System Suitability (CompRef) Study" was to validate mass spectrometry protocols used at each Proteome Characterization Center (PCC). Four human-in-mouse xenograft samples were used.	1.03
	Time Course Colon Cancer (0,10,30,60)		Phosphoproteome and proteome study of human colon adenocarcinoma core biopsy punches at 4 time points (0, 10, 30, and 60 minutes ischemia)	0.154
	Time Course Breast Cancer (0,60)		The Time-Dependent Proteome Studies evaluate changes in the proteome and phosphoproteome when there is a delay in freezing tissues following sample (tumor) excision. The 4 biospecimens used are from human-in-mouse breast cancer tumor samples at two time points, 0 and 60 minutes.	0.173
	Time Course Breast Cancer (0,5,30,60)		Proteome and Phosphoproteome data from human-in-mouse breast cancer tumor samples at 4 time points, 0, 5, 30, and 60 minutes. Twenty-four xenograft samples are included in this study.	0.55

YT	Project	File types	Other information	Approx. Size (TB)
	Time Course Ovarian Cancer (0,5,30,60)		The Time-Dependent Proteome and Phosphoproteome Studies of 4 patient-derived ovarian cancer tumors at 0, 5, 30, and 60 minutes.	0.07
	Embedding Media Study (OCT)		The goal of the Optimal Cutting Temperature (OCT) Embedding Media Study was to measure the effect of OCT embedding on peptide identification. One OCT embedded sample and one snap frozen sample were compared.	0.014

Additional PS-ON and CSSI Data Information			
Project	**Samples**	**File Type(s)**	**Submission Size**
Genomic Characterization - mRNA	39 cell lines	raw reads fastq each including report summarizing reads (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)	0.3 TB
		cufflinks output - gtf files	
		vcf file	
Genomic Characterization - miRNA	39 cell lines	fastq	0.15 TB
		trimmed fastq	
		bam	(0.1 TB is output of miRanalyzer)
		unmapped fastq	
Genomic Characterization - exome	39 cell lines	raw reads and trimmed reads fastq, each including report summarizing reads (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)	0.3 TB
		bam - HG19	
		vcf	
		bed	
Physical Characterization - Atomic Force Microscopy (AFM)	30 cell lines	txt (tsv)	0.3 GB
		xlsx	
Physical Characterization - Morphology Images	29 cell lines	tif jpg	30 GB
Physical Characterization -	30 cell lines	txt	

Additional PS-ON and CSSI Data Information			
Project	**Samples**	**File Type(s)**	**Submission Size**
Morphology Data		xls	
Physical Characterization - Motility	30 cell lines	tif	336 GB
		xls	
		txt	
		py	
Physical Characterization - Traction force	29 cell lines (NCI-PBCF-CRL1740 LnCAP clone FGC did not grow)	tif	682 GB
		lif	
		jpg	
Physical Characterization - Traction force - Summary	29 cell lines	xls	
Proteomic Characterization - TBD	30 cell lines		
caSix - TBD			
CTC-Phase 1 - TOTAL	patient blood samples	jpg	0.1 GB
		xls	
		txt	
CTC-Phase 1 - Blood Collection Tube (BCT) comparing anticoagulants	patient blood samples		
CTC-Phase 1 - Time to Assay	patient blood samples		
CTC-Phase 1 - Streck vs CellFree	patient blood samples		
CTC-Phase 1 - single cell low pass sequencing	single cells of one patient	single excel spreadsheet	< 0.001 GB
CTC-Phase 2 - TBD			
Publication from Nature, Scientific Reports		power point	7.9 GB
Sci Rep. 2013;3:1449. doi: 10.1038/srep01449.		tif	
A physical sciences network characterization of non-tumorigenic and metastatic cells.		xls	
		pdf	
		docx	

Additional PS-ON and CSSI Data Information		
Project	**Primary Data**	**Derived Data**
Genomic Characterization - mRNA	reads prepared using TruSeq Stranded RNA Sample Preparation Kit, and bar-coded with individual tags. Library preparation performed using a semi-	bam files
		transcript fasta, gff, bed
		insertion/deleti

Additional PS-ON and CSSI Data Information		
Project	**Primary Data**	**Derived Data**
	automated 96-well plate method, with washing and clean-up/concentration steps performed on the Beckman Coulter Biomek NX platform and with ZR-96 DNA Clean & ConcentratorTM-5 plates .	on: bed
		cufflinks output
		gene fusions
		vcf
Genomic Characterization - miRNA	reads prepared using TruSeq Small Total RNA Sample Prep Kit, and bar-coded with individual tags. Library preparation will be performed using a semi-automated 96-well plate method, with washing and clean-up/concentration steps performed on the Beckman Coulter Biomek NX platform and with ZR-96 DNA Clean & ConcentratorTM-5 plates.	bam files
		miRanalyzer output
		diff expression
Genomic Characterization - exome	reads from libraries prepared using Agilent SureSelectXT2 Human All Exon V5	bam files
		mutation candidates
		vcf
Physical Characterization - Atomic Force Microscopy (AFM)	AFM measurements of single cells from cell lines	Young's Modulus
		Cantilever Spring constant
Physical Characterization - Morphology	Images of cells from cell lines.	area
		circularity
		aspect ratio
Physical Characterization - Motility	Images of cells from cell lines.	end to end distance
		total distance
		speed
Physical Characterization - Traction force	Images of cells from cell lines, labeled with CellTracker Green CMFDA (Invitrogen) and DRAQ5 (Cell Signaling Technology) to label cytoplasm and cell nucleus respectively. A 63x, 1.2 NA water immersion objective on a laser scanning confocal microscope (Leica TCS SP5, Wetzlar, Germany)	Cell Volume
		Traction Force
		Nuclear Volume
		Cell Area
Proteomic Characterization - TBD	mass spec data	Peptide sequence(?)
	peptide ID, MW, numeric measurement (TOF? intensity, m/z)	
	will need reference data that identifies peptides	

Additional PS-ON and CSSI Data Information		
Project	**Primary Data**	**Derived Data**
	growth characteristics	
	images for back comparison- physical measurements	
CaSix - TBD	genomic data- defined set of genes	
	proteomic data	
	image data	
	pathology report including diagnosis, location of tumor, histology reports etc.	
CTC-Phase 1 - Blood Collection Tube (BCT) comparing anticoagulants	images of cells	CTCs/mL
		relative nuclear size
		relative cytokeratin intensity
CTC-Phase 1 - Time to Assay		
CTC-Phase 1 - Streck vs CellFree		
CTC-Phase 1 - single cell low pass sequencing	sequenced reads from single CTC (not provided)	numerical value for chromosome segments
CTC-Phase 2 - TBD	6 high content parameters	target analysis on CFD and CTCs
	Sequencing of Cell Free DNA	
	Sequencing of CTCs	
	Cytof- mass cytometry, quantitive epitope characterization	
Publication from Nature, Scientific Reports	Some of the methodologies according to supplemental material:	
Sci Rep. 2013;3:1449. doi: 10.1038/srep01449.	1. Morphology	
	1. Differential Interference Contrast Microscopy	
A physical sciences network characterization of non-tumorigenic and metastatic cells.	2. Single-cell Tomographic Imaging and 3D Morphometry	
	3. Partial Wave Spectroscopy	
	4. Matrix Stiffness (seems to actually be a measure of proliferation?)	
	5. CD44 and lipid raft distribution assays	
	2. Motility and Mechanics	
	1. Measurement of Maximum Displacement	
	2. 1D Migration: Device Preparation and Time-lapse Microscopy	

Additional PS-ON and CSSI Data Information		
Project	**Primary Data**	**Derived Data**
	3. 2D Migration: 2D Collagen Substrate	
	4. 3D Migration: 3D Collagen Matrix	
Thrombosis	marker identification within plasma (4) (place holder)- elisa	
Immunomrm (Ras quantitation)	http://www.cancer.gov/research/key-initiatives/ras/ras-central/blog/ras-quantitative-assays	

PS-ON and CSSI Data Information								
Reference Data	**Primary Data and subject matter**	**Derived.1**	**Derived.2**	**Derived.3**	**Derived.4**	**Derived.5**	**Derived.6**	**Notes**
	jpg image of circulating tumor cell	txt						
human genome assembly (version TBD)	low pass sequencing reads? (not provided)	bam file (not provided)	screen redundant mappings (not provided)	final set of markers and windows?(not provided)	excel spreadsheet jpg of cell			
	fastq.R1, R2	fastq trimmed fastq trimmed 16-25	bam miRanalyzer output txt (describing read length)					
	raw reads fastq.gz	fastqc.zip - a report of raw reads generated using http://www.bioinformatics.babraham.ac.uk/projects/fastqc/						
	raw reads fastq.gz	trimmed readsfq.gz	fastqc.zip - a report of trimmed reads generated using http://www.bioinformatics.babraham.ac.uk/projects/fastqc/					

PS-ON and CSSI Data Information								
Reference Data	**Primary Data and subject matter**	**Derived.1**	**Derived.2**	**Derived.3**	**Derived.4**	**Derived.5**	**Derived.6**	**Notes**
	raw reads fastq.gz	trimmed readsfq.gz (GUESSING THEY USED TRIMMED READS)	tophat output including:					
			bam					
			bam.bai					
			bed					
HG19 build of the human genome	*R1 and R2 refer to forward and reverse strands. files prefaced with run531 indicate the second of two runs.*	trimmed_reads/*.R1.trim.fq.gz	*mark.bam	txt files (annotations)				
	raw_reads/*R1.fastqc.gz	trimmed_reads/*.R1.trim.fq.gz						
	raw_reads/*R2.fastqc.gz							
	R1 and R2 refer to forward and reverse strands. files prefaced with run531 indicate the second of two runs.	trimmed_reads/*.R1.trim.fq.gz	trimmed_reads/*R1.trim.fq_fastqc.zip					
	raw_reads/*R1.fastqc.gz	trimmed_reads/*.R1.trim.fq.gz	trimmed_reads/*R2.trim.fq_fastqc.zip					
	raw_reads/*R2.fastqc.gz		(a report of raw reads generated usinghttp://www.bioinformatics.babraham.ac.uk/projects/fastqc/)					

PS-ON and CSSI Data Information								
Reference Data	**Primary Data and subject matter**	**Derived.1**	**Derived.2**	**Derived.3**	**Derived.4**	**Derived.5**	**Derived.6**	**Notes**
	R1 and R2 likely refer to run1 and run2, as contracted.?	raw_reads/R1.fastqc.zip						what is the difference between R1 and R2?
	raw_reads/*R1.fastqc.gz	raw_reads/R1.fastqc.zip						
	raw_reads/*R2.fastqc.gz	(a report of raw reads generated usinghttp://www.bioinformatics.babraham.ac.uk/projects/fastqc/)						
(none)	Volume.lif	Cell__.tif					cell_line_PBCF.xls	
(none)	Volume.lif	Cell__.tif	CellArea__.tif		crop_CellArea.tif		cell_line_PBCF.xls	
(none)	Volume.lif	Nucleus__.tif					cell_line_PBCF.xls	
(none)	Volume.lif	BeadsCell__.tif		BeacsC__.tif	crop_BeadsC__.tif	BeadsC__#constrained.jpg	cell_line_PBCF.xls	
(none)	Volume.lif	BeadsCell__.tif		BeacsC__.tif	crop_BeadsC__.tif	BeadsC__#unconstrained.jpg	cell_line_PBCF.xls	
(none)	Volume.lif	BeadsCell__.tif		BeacsC__.tif	crop_BeadsC__.tif	BeadsC__#displacement.jpg	cell_line_PBCF.xls	
(none)	Traction.lif	BeadsReleased__.tif	BeadsRef__.tif	BeadsR__.tif	crop_BeadsR__.tif		cell_line_PBCF.xls	
(none)	Date_condition_cell_instance.txt	Date_cell_line_For upload.xlsx	CELL_LINE_AFM.xls					
(none)	CELL_LINE_plate_ti	Results_from_CELL_LIN	Motility.py	motility.summary.txt				*note, there is no

PS-ON and CSSI Data Information								
Reference Data	**Primary Data and subject matter**	**Derived.1**	**Derived.2**	**Derived.3**	**Derived.4**	**Derived.5**	**Derived.6**	**Notes**
	me_cell_instance.tif	E_plate_condition.xls						mapping from derived data back to the original tif file used to derive these values
(none)	CELL_LINE_plate_time_condition.tif	area data.txt aspect ratio data.txt circularity data.txt	summary.txt					*note the proliferation README refers to several numbers of frames per tif file. Not clear how it was decided which frames were used for morphology results. Were all of them used? If not, which ones were used?
(none)	CELL_LINE_plate_time_condition.tif	Leidos project counting results.xlsx						Content is counts of touching cells, count of single cells, number of frames per cell line, condition, time and plate. *note the README refers to several numbers of frames per tif file. xls spreadsheet provides number of frames read, but does not map back to what frame had what value.
(none)	CELL_LINE_plate_time_condition.tif	Cell counting timex.xlsx						Content is the number of cells counted per plate per time.time is in hours, this is not indicated in the file, only in the README