University of California
Santa Cruz

CGHub User Guide

January 11, 2011

Version 1.2

1 Introduction

CGHub is a genome sequence storage repository that provides large capacity and fast access to sequence data generated as part of NCI research projects. The repository is accessed over the Internet, providing simple but secure sequence data search and transfer services. The CGHub currently contains sequence data for The Cancer Genome Atlas (TCGA) cancer research study sponsored by the NCI and NHGRI. NCI plans to add more studies (such as TARGET and CGCI) over time.

For the TCGA study, sequence data is considered personally identifying information (PII) because it comes from human clinical samples. In order to support research use guidelines and enforce participant privacy rights, only NIH authorized users of TCGA can access the genome sequence data via protected CGHub Web Services. The metadata that describes the sequence and related analysis, captured as XML-based metadata descriptors, is publicly available via the CGHub Web Services.

1.1 Technical Support

For technical or operational requests, contact: support@cghub.ucsc.edu

1.2 References

CGHub Home Page: https://cghub.ucsc.edu/

TCGA Project Home Page at NCI: http://cancergenome.nih.gov/

TCGA Data Guide: https://wiki.nci.nih.gov/display/TCGA/TCGA+Data+Primer

TCGA Code Tables:http://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm

dbGaP TCGA Study Page: http://www.ncbi.nlm.nih.gov/gap?term=phs000178

1.3 Document Scope

This document describes how to search and download primary genome sequence data and related metadata from the CGHub repository.

Downloading sequence data from CGHub involves three user steps:

1. Use your browser to access your eRA account and generate a CGHub authentication credential
2. Query the metadata to get the download URLs of the desired analysis objects that contain the sequence data of interest
3. Invoke the GeneTorrent Client with the selected URLs to download

These steps will be described in detail in the subsequent sections.

Once a generated authentication credential has been obtained it can be re-used for an unlimited number of download operations until the credential expires. The credential will need to be periodically renewed; a new credential can be obtained at any time. Keep this credential in a safe and secure place on the computer used for downloading.

Note that the authentication credential is not required to query the CGHub metadata; the authentication credential is required only for download. Query functionality is available to the public.

Queries can be performed with either the *cgquery* script or the CGHub Web Services API.

- The *cgquery* script provides a simple but powerful way to execute queries and file download. It is recommended that new users start with the *cgquery* script before attempting to use the CGHub Web Services API.
- The CGHub Web Services API returns an XML result set. This method is useful when scripting or embedding queries into your site-specific tools.

2 System Overview

CGHub serves as the central repository for sequencing centers to store and for analysis centers to retrieve genome sequence data and its associated metadata.

There are four main groups that work together to support the CGHub repository: Tissue Source Sites, Biospecimen Core Resource Centers, Genome Sequencing Resource Centers, and Data Coordinating Centers

A **Tissue Source Site (TSS)** collects samples (tissue, cell or blood) and clinical metadata, which are then sent to a BCR. A TSS is identified by its TSS ID.

A **Biospecimen Core Resource (BCR)** is a TCGA center where samples are carefully catalogued, processed, quality-checked and stored along with participant clinical information.

The work of the BCR includes the following important functions:

- Serving as the interface between the TCGA program and the different Tissue Sources that are collecting tumor and matched normal controls
- Ensuring and verifying that TCGA human subjects protections and guidelines are adhered to and that all regulations are followed at each Tissue Source Site
- Examining of all biospecimens to ensure they meet rigorous standards for each tumor type (including percent necrosis and percent tumor nuclei)

- Reviewing of pathology to ensure accurate diagnosis and inclusion in TCGA
- Collecting clinical information for each sample and applying standardized terminology, definitions and formats that are caBIG compliant
- Extracting and distributing DNA and RNA from samples to each of the genomic characterization and sequencing centers

A **Genome Sequencing Center (GSC)** is a TCGA center that uses high-throughput methods to sequence tumor and normal samples provided by the BCR.

The **Data Coordinating Center (DCC)** is the central provider of TCGA data. The DCC standardizes data formats and validates submitted data.

The work of the DCC includes the following important functions:

- Protecting participant privacy and confidentiality through secure access to research and clinical information that are classified as controlled access datasets
- Developing data standards and controlled vocabularies
- Establishing informatics pipelines for dataflow from production centers to a central repository
- Developing new analytical and visualization technologies for different audiences to facilitate data analysis
- Coordinating project level activities.

The following describes the basic system workflow for creating and distributing genome sequencing data. This workflow is provided here to orient the reader to the end-to-end system operation and is for informational purposes only. More details can be found in the CGHub Submission Guide.

- A Tissue Source Site (TSS) submits a participant tissue sample to a Biospecimen Resource Center (BCR) where the sample is stored and prepared for sequencing.
    1. The analyte is shipped to a Genome Sequencing Center (GSC) with a unique identifier (UUID) for the sample aliquot.
    2. Metadata describing the biospecimen is sent to TCGA Data Coordination Center (DCC), who coordinates project level activities.
- GSC receives and processes the analyte from the BCR
    1. The analyte is sequenced by a high throughput sequencer to produce raw read coverage

    2. Additional processing, such as alignment, mapping, and expression counts are performed and captured in a BAM file.
- GSC submits the sequence information to CGHub
    1. A unique identifier (AnalysisUUID) for sequence data and metadata information for the single sample analysis is created by the GSC.
    2. This set of data (metadata and sequence data) is referred to as the Analysis Object.
    3. GSC performs validation of the sequence data and metadata prior to upload to ensure proper formatting.
    4. Once validated, the GSC will upload the Analysis Object to CGHub and initiate the workflow on CGHub to make the data available to downloaders.
- CGHub synchronizes with the TCGA DCC to determine status of sample.
    1. The DCC maintains the higher-level study attributes as well as clinical data associated with each sample.
    2. CGHub will query the DCC to verify that submitted data is associated with a valid sample that should be available to authorized researchers.
    3. CGHub will also query for redacted samples and suppress the sequence data from download.
- Researchers retrieve the CGHub catalog and perform queries across the public sequence metadata using selected metadata attributes such as cancer type, sequence type, source sequencing center, or date range.
- Researchers retrieve protected sequence data using their NIH authorized credentials, as long as the sequnce data is in the “live” state.

3 Software Installation

3.1 *cgquery* Script

The *cgquery* is a separate package from GeneTorrent and therefore must be downloaded and installed separately. Note: to permanently install this software you must be a System Administrator.

3.1.1 Prerequistes

Python 2.6.5 or later with SSL support.

Check for Python Version

```
$ python -c 'import platform; print platform.python_version()'
2.6.5
```

"2.6.5" or greater is required

Check if Python supports SSL

Use the following command to check if the python installation supports SSL.

```
$ python -c 'import socket; print hasattr(socket, "ssl")'
True
```

"True" is required

3.1.2 Download *cgquery* Software Package

From a browser:

- go to GeneTorrent Software Downloads web page (https://cghub.ucsc.edu/downloads.html)
- click on "cgquery-1.9.tar.gz" or latest version

From command line terminal window:

Command: `wget http://cghub.ucsc.edu/cgquery-1.9.tar.gz`

3.1.3 Install cgquery Script

Install *cgquery* with the following command:

Command: `sudo tar -C / -xzvf cgquery-1.9.tar.gz`

For technical or operational requests, contact: support@cghub.ucsc.edu

3.2 GeneTorrent Client Program

Note: to permanently install this software you must be a System Administrator.

Binary (ready-to-run) versions of the client are available for 64 bit versions of the following LINUX Systems:

- Centos6
- Centos 5.5
- SUSE11.4
- FedoraCore15
- Ubuntu11.4
- Ubuntu10.4

In addition, the source code for GeneTorrent can be found at **GeneTorrent-1.0.0.0-src.tgz** (https://cghub.ucsc.edu/GeneTorrent-1.0.0.0-src.tgz).

3.2.1 Prerequisites

The GeneTorrent software requires certain software packages to be installed to function. Curl and OpenSSL are required.

3.2.2 Download GeneTorrent Software Package

To find the GeneTorrent software installation package, go to GeneTorrent Software Downloads (https://cghub.ucsc.edu/downloads.html). Select a download according to your operating system. This will download a file named: GeneTorrent-1.0.0.0-YOUR_OS.x86_64.tar.gz (where YOUR_OS is the name of the operating system that you selected).

3.2.3 Install GeneTorrent Software

Install GeneTorrent with the following command:

Command: `sudo tar -C / -xzvf GeneTorrent-1.0.0.0.$YOUR_OS.x86_64.tar.gz`

Where $YOUR_OS is one of the operating systems listed in the OPERATING SYSTEMS SUPPORTED section.

```
]$ sudo tar -C / -xzvf GeneTorrent-1.0.0.0.Centos5.5.x86_64.tar.gz
[sudo] password for user:
usr/
usr/bin/
usr/bin/GTLoadBalancer
usr/bin/GeneTorrent
usr/bin/gtoinfo
usr/share/
usr/share/man/
usr/share/man/man1/
usr/share/man/man1/GeneTorrent.1
usr/share/GeneTorrent/
usr/share/GeneTorrent/GeneTorrent.openssl.conf
usr/share/GeneTorrent/dhparam.pem
```

In addition, there is an RPM package of GeneTorrent for Centos 6. This can installed using the following commands:

Command: `gunzip GeneTorrent-1.0.0.0-1.x86_64.rpm.gz`
Command: `sudo yum --nogpg install GeneTorrent-1.0.0.0-1.el6.CP.x86_64.rpm`

Note: Remember to answer ‘y’ to the yum prompt “Is this ok [y/N]:”

3.2.4 Check Installation

To verify that the installation succeeded and that the correct version of the software is installed type the following command.

```
[admin@centos6 ]$ GeneTorrent --version
GeneTorrent: Genomic Data Transfer Tool version 1.0.0.0
```

For technical or operational requests, contact: support@cghub.ucsc.edu

4 User Authentication

In order to download CGHub data, users must have a valid NIH or eRA account and must have received authorization from the TCGA project Data Access Committee (DAC). Please refer to the dbGaP study page for more information on applying for access to controlled-access data: http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v5.p5.

4.1 Authentication

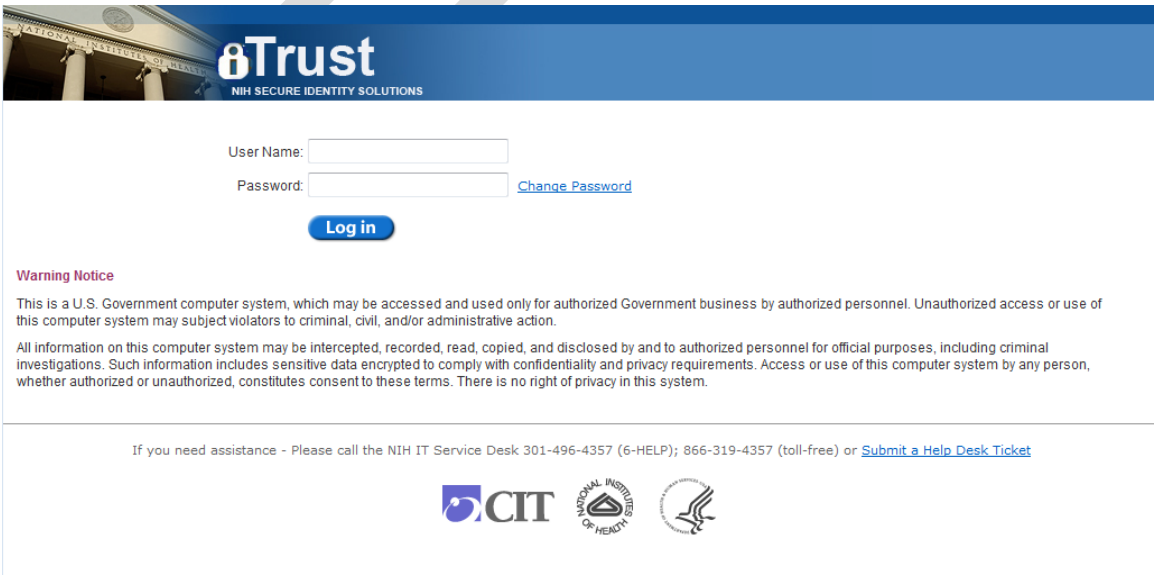
Steps 1 – 4 are only required to generate a new key. Download commands can be run repeatedly using the same key until it expires.

Step 1: Access URL for CGHub Authentication page.

The user begins to log in at the URL: https://cghub.ucsc.edu/secure.

Step 2: The CGHub system redirects the user to the NIH login page show below.

The user provides valid login credentials (Username and Password).



Step 3: NIH authenticates the user.

Step 4: CGHub generates and returns an authentication credential.

The user will be prompted to save the authentication credential in a local file.

This credential is valid for 365 days. You will need to repeat this process to generate a new credential whenever the credential expires or the user's account changes.

Pass the credential to the GeneTorrent download client as part of the sequence data download step.

5 Sequence Metadata Queries

Each sequence in the CGHub is described by metadata, which contains attributes such as the date, sample type, and disease abbreviation. (see section 5.1 for the list of searchable attributes).

Querying the sequence metadata is a powerful method of identifying sequences of interest.

cgquery is a command line tool that allows the user to construct and submit a querystring to search for sequences with one more sequence attributes. The results are formatted in either summary or complete view. (see section 5.2 *cgquery*)

cgquery uses public CGHub REST APIs. Users can call this interface directly through a browser or other tool such as wget or curl and develop custom GUIs or command line tools. (see section 5.4 for more detail)

5.1 Query Attributes

The querystring can include one or more of the attributes listed below using name=value pairs. Multiple queries can be combined using the '&' symbol. For example : "disease_type=OV&sample_type=01". Both attribute names and values are case sensitive. Values can use the "*" wildcard. For example, "disease_type=O*" will match any values beginning with "O".

CGHub also stores the original submission XML documents, which users can perform free form text search against any strings within the analysis.xml, run.xml and experiment.xml files. xml_text queries are case insensitive.

Name	Description	Example	Metadata Source
analysis_id	Globally unique ID (uuid) associated with a collection of sequencing data files and their metadata. This is the primary identifier for CGHub data. It is equivalent to an SRA SRZxxxx accession #.	analysis_id=d29389cd-72F8-48B1-bd33-7f525170c7b8	Generated by GSC prior to submission. Legacy data was assigned uuids when it was migrated from SRA into CGHub.
analysis_accession	Legacy SRZnnnnnn analysis accession # for data migrated from SRA.	analysis_accession=SRZ123456	analysis.xml <ANALYSIS accession=>
state	Determines if the submission is still being processed, is live or has been suppressed. Users can only download data with state=live. Valid values: submitted,	state=live	Managed by CGHub.

	uploading, live, suppressed, verifying_sample, verifying_files, bad_data		
Study	The study for which the data was generated. This value is used to determine which users can download the data. Valid values: phs000178	study=phs000178	analysis.xml <STUDY_REF refname=>
aliquot_id	Globally unique ID for the aliquot used by to generate this analysis. This is a primary identifier for each project's sample database (e.g. TCGA's DCC Portal). For TCGA, this will be a RFC4122 aliquot uuid managed by the DCC.	aliquot_id= a8C71dbb-a809-4415-bde8-f24f70557077	analysis.xml <TARGET refname=>
sample_accession	Legacy sample accession # for data migrated from SRA.	sample_accession=SRS123456	Analysis.xml <TARGET accession=>
legacy_sample_id	This may be any project specific ID. For TCGA, this will be the 7 part barcode retrieved from the DCC.	legacy_sample_id= TCGA-AB-2802-03C-01W-0731-08	TCGA DCC uuid to barcode mapping Web Service
disease_abbr	The short disease name. The set of legal TCGA values are defined by the DCC codes table at: http://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm	disease_abbr=OV	TCGA DCC uuidws Web Service <disease> <abbreviation>
participant_id	The ID of the human participant providing the sample. Previously represented by the TCGA 3 part barcode (e.g. TCGA-AB-2802**)**. For TCGA this will be a participant UUID provided by the DCC.	participant_id= a85391a4-c6bc-4cd7-b53b-6e7cebf044b5	TCGA DCC uuidws Web Service <participant href=>
sample_id	The ID for the root sample from which the aliquot was derived. Previously represented by the TCGA 5 part barcode (e.g. TCGA-AB-2802-03C-01W**)**.	sample_id= 972a7b73-9323-4b02-90dd-bb420a18681f	TCGA DCC uuidws Web Service <sample href=>

	For TCGA this will be a sample UUID provided by the DCC.		
analyte_code	The type of analyte (e.g. D=DNA, R=RNA, etc.). The set of legal TCGA values are defined by the DCC at: http://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm	analyte_code=D	TCGA DCC uuidws Web Service <analyteType>
sample_type	The type of sample (e.g. Blood Derived Normal, Primary solid Tumor, etc.). The set of legal TCGA values are defined by the DCC codes table at : http://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm	sample_type=01	TCGA DCC uuidws Web Service <sampleType>
tss_id	The Tissue Source Site The set of legal TCGA values are defined by the DCC codes table at : http://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm	tss_id=01	TCGA DCC uuidws Web Service <TissueSourceSite>
last_modified	Date the object was last modified.	1) "in the last month": last_modified=[NOW-1MONTH TO NOW] 2) "all September 2011 records": last_modified=[2011-08-31T23:59:59.99Z TO 2011-09-30T23:59:59.99Z]	Created by CGHub
center_name	Short names defined by the project (e.g. BI, BCM). The set of legal TCGA values are defined by the DCC codes table at : http://tcga-data.nci.nih.gov/datareports/codeTablesReport.htm	center_name=BI	analysis.xml <ANALYSIS center_name=>
alias	Name of the analysis object as defined by the submitting center.	alias= NA12878h.HiSeq.WGS.bwa.cleaned.recal.bam.header	analysis.xml <ANALYSIS alias=>
Title	Title of the analysis object as defined by the submitting center.	title= Reference alignment of M01005	analysis.xml <TITLE >
analysis_typ	The type of analysis. Uses	analysis_type=REFER	analysis.xml

e	the values from SRA 1.3 schema. For TCGA, this will be REFERENCE_ALIGNMENT	ENCE_ALIGNMENT	<ANALYSIS_TYPE>
library_strategy	The sequencing technique used. Values are defined by the SRA 1.3 schema: WGS WXS RNA-Seq WCS CLONE POOLCLONE AMPLICON CLONEEND FINISHING ChIP-Seq MNase-Seq DNase-Hypersensitivity Bisulfite-Seq EST FL-cDNA CTS MRE-Seq MeDIP-Seq MDB-Seq OTHER	library_startegy=WGS	experiment.xml <LIBRARY_STRATEGY>
platform	The machine used to generate the sequence data. Values are defined by the SRA 1.3 schema: LS454 ILLUMINA HELICOS ABI_SOLID COMPLETE_GENOMICS PACBIO_SMRT ION_TORRENT	platform=ILLUMINA	experiment.xml <PLATFORM>
filename	Name of a data file associated with an analysis.	filename=HG00099*	analysis.xml and/or experiment.xml <FILES>
xml_text	Free form search from the original submission XML documents.	xml_text= TCGA-AB-2802-03C-01W	analysis.xml, run.xml and experiment.xml

5.2 Using the *cgquery* Script

The *cgquery* script calls the Web services APIs and parses the XML output in a human readable form. It takes the same query string arguments as the direct Web Services API.

```
Usage: cgquery [options] <querystring>


Options:
  --version             show program's version number and exit
  -h, --help            show this help message and exit
  -o OUTPUTXML, --output-xml=OUTPUTXML
                        file in which to store raw xml output
  -s SERVER, --server=SERVER
                        CGHub server location, including protocol and
port,
                        e.g. https://cghub-01.ucsc.edu:20000
  -a, --attributes      query the /cghub/metadata/analysisAttributes
instead
                        of analysisObjects resource
  -i, --interactive     enable interactive mode
  -c CREDENTIAL, --credential=CREDENTIAL
                        file containing the GeneTorrent credential.
Only
                        required for interactive mode.
  -v, --verbose         enable verbose output


<querystring> should be the fully quoted query string, without the
question
mark separator, e.g. "disease_abbr=COAD".
```

5.3 Example *cgquery* Commands

The first query example searches for all sequences for a known participant.

The querystring is **"participant_id=6d72de06-232a-4983-a06c-eba6d82cb3f1"**

```
]$ ./cgquery "participant_id=6d72de06-232a-4983-a06c-eba6d82cb3f1"

================================================================================
    Script Version               : 1.9
    CGHub Server                 : https://cghub.ucsc.edu
```

```
    REST Resource                : /cghub/metadata/analysisObject
    QueryString                  : participant_id=6d72de06-232a-4983-
a06c-eba6d82cb3f1
    Output File                  : None
------------------------------------------------------------------------
    Results Returned             : 7
========================================================================

    Result 1
        analysis_id              : e29aa109-d508-4621-9a92-9f7ff7e0018f
        analysis_data_uri        :
https://cghub.ucsc.edu/cghub/data/analysis/download/e29aa109-d508-4621-
9a92-9f7ff7e0018f
        analysis_attribute_uri   :
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes/e29aa109-d508-
4621-9a92-9f7ff7e0018f
        last_modified            : 2011-06-13T07:00:00Z
        center_name              : BI
        state                    : live
        aliquot_id               : 087484e8-dc3e-461a-be5f-4217b7c39732
        Files
            filename             : C509.TCGA-55-1594-11A-01D-1040-
01.2.bam
            filesize             : 23140869502
            checksum             : d39c62bdb7abf0213b837fe738f81821
```

The query returns 7 results. The output has been truncated after the first result.

The next example does the same query but adds the –a option to show all attributes associated with the search results.

```
]$ ./cgquery -a "participant_id=6d72de06-232a-4983-a06c-eba6d82cb3f1"

========================================================================
    Script Version               : 1.9
    CGHub Server                 : https://cghub.ucsc.edu
    REST Resource                : /cghub/metadata/analysisAttributes
    QueryString                  : participant_id=6d72de06-232a-4983-
a06c-eba6d82cb3f1
    Output File                  : None
------------------------------------------------------------------------
    Results Returned             : 7
========================================================================

    Result 1
        analysis_id              : e29aa109-d508-4621-9a92-9f7ff7e0018f
        analysis_data_uri        :
https://cghub.ucsc.edu/cghub/data/analysis/download/e29aa109-d508-4621-
9a92-9f7ff7e0018f
        last_modified            : 2011-06-13T07:00:00Z
        center_name              : BI
        state                    : live
        study                    : phs000178
        library_strategy         : WXS
        platform                 : ILLUMINA
```

```
        sample_accession          : SRS127193
        legacy_sample_id          : TCGA-55-1594-11A-01D-1040-01
        disease_abbr              : LUAD
        analyte_code              : D
        sample_type               : 11
        tss_id                    : 55
        participant_id            : 6d72de06-232a-4983-a06c-eba6d82cb3f1
        sample_id                 : 99dc64d5-e61d-4e34-9972-f17d292aec3d
        aliquot_id                : 087484e8-dc3e-461a-be5f-4217b7c39732
        analysis_xml              : 134749 bytes of XML
        run_xml                   : 95798 bytes of XML
        experiment_xml            : 19466 bytes of XML
        Files
            filename              : C509.TCGA-55-1594-11A-01D-1040-
01.2.bam
            filesize              : 23140869502
            checksum              : d39c62bdb7abf0213b837fe738f81821
```

The query again returns 7 results. The output has been truncated after the first result.

The next query example searches for all sequences that have been modified in the last 5 months.

The querystring is **"last_modified=[NOW-5MONTH+TO+NOW]"**

```
]$ ./cgquery "last_modified=[NOW-5MONTH+TO+NOW]"

================================================================================
    Script Version            : 1.9
    CGHub Server              : https://cghub.ucsc.edu
    REST Resource             : /cghub/metadata/analysisObject
    QueryString               : last_modified=[NOW-5MONTH+TO+NOW]
    Output File               : None
--------------------------------------------------------------------------------
    Results Returned          : 1832
================================================================================

    Result 1
        analysis_id           : a03207f9-5fba-4e0f-a9e9-981b13e5c193
        analysis_data_uri     :
https://cghub.ucsc.edu/cghub/data/analysis/download/a03207f9-5fba-4e0f-
a9e9-981b13e5c193
        analysis_attribute_uri   :
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes/a03207f9-5fba-
4e0f-a9e9-981b13e5c193
        last_modified         : 2011-08-12T07:00:00Z
        center_name           : UNC-LCCC
        state                 : live
        aliquot_id            : 0031d433-d703-4b2c-9fdf-2920008eb457
        Files
            filename          : UNCID_339878.TCGA-A6-2681-01A-01R-
1410-07.110309_UNC3-
RDR300156_00080_FC_62J42AAXX.4.trimmed.annotated.translated_to_genomic.
bam
```

```
            filesize              : 1906592634
            checksum              : 6703e2b337ebac2a4e12cd09f0d1de77
```

This query returned 1832 results. The output has been truncated after the first result.

In order to refine the set of sequences of interest the next example searches for all sequences that have been modified in the last 5 months and have a sample type of 11 (solid tissue normal). The querystring is **"last_modified=[NOW-5MONTH+TO+NOW] &sample_type=11"**

```
]$ ./cgquery "last_modified=[NOW-5MONTH+TO+NOW]&sample_type=11"
================================================================================
    Script Version                : 1.9
    CGHub Server                  : https://cghub.ucsc.edu
    REST Resource                 : /cghub/metadata/analysisObject
    QueryString                   : last_modified=[NOW-
5MONTH+TO+NOW]&sample_type=11
    Output File                   : None
--------------------------------------------------------------------------------
    Results Returned              : 48
================================================================================

    Result 1
        analysis_id               : 187cc211-9de3-44aa-91f1-74e35a0bf98b
        analysis_data_uri         :
https://cghub.ucsc.edu/cghub/data/analysis/download/187cc211-9de3-44aa-
91f1-74e35a0bf98b
        analysis_attribute_uri    :
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes/187cc211-9de3-
44aa-91f1-74e35a0bf98b
        last_modified             : 2011-08-12T07:00:00Z
        center_name               : UNC-LCCC
        state                     : live
        aliquot_id                : 02222fea-f3da-4328-8188-8ef24d1f55e4
        Files
            filename              : UNCID_368056.TCGA-BH-A18U-11A-23R-
A12D-07.110714_UNC13-
SN749_0082_AD0DGMABXX.7.trimmed.annotated.translated_to_genomic.bam
            filesize              : 3085973081
            checksum              : 6cd3773bb591a255b93223e02daa022b
```

This search returns 48 results. The output has been truncated after the first result.

The final example shows how to search using the wildcard character "*". Suppose you have a partial aliquot id for a sample and you want all attributes associated with the sequence. The querystring is **"aliquot_id=*c90-a762-4df7-8b44-b4facd*"**

```
]$ ./cgquery -a "aliquot_id=*c90-a762-4df7-8b44-b4facd*"

================================================================================
    Script Version               : 1.9
    CGHub Server                 : https://cghub.ucsc.edu
    REST Resource                : /cghub/metadata/analysisAttributes
    QueryString                  : aliquot_id=*c90-a762-4df7-8b44-
b4facd*
    Output File                  : None
--------------------------------------------------------------------------------
    Results Returned             : 1
================================================================================

    Result 1
        analysis_id              : 8ca52fbf-60ab-477f-90cd-0704d6b5b42d
        analysis_data_uri        :
https://cghub.ucsc.edu/cghub/data/analysis/download/8ca52fbf-60ab-477f-
90cd-0704d6b5b42d
        last_modified            : 2011-08-12T07:00:00Z
        center_name              : UNC-LCCC
        state                    : live
        study                    : phs000178
        library_strategy         : RNA-Seq
        platform                 : ILLUMINA
        sample_accession         : SRS136282
        legacy_sample_id         : TCGA-BP-4342-01A-01R-1289-07
        disease_abbr             : KIRC
        analyte_code             : R
        sample_type              : 01
        tss_id                   : BP
        participant_id           : c0e26587-3191-4396-a20b-da9e7de213f4
        sample_id                : 04d3168b-ae52-400c-94fa-8ec43eface20
        aliquot_id               : 005a0c90-a762-4df7-8b44-b4facd06e9b1
        analysis_xml             : 2167 bytes of XML
        run_xml                  : 663 bytes of XML
        experiment_xml           : 2432 bytes of XML
        Files
            filename             : UNCID_374694.TCGA-BP-4342-01A-01R-
1289-07.110511_UNC11-
SN627_0085_BC01AFABXX.5.trimmed.annotated.translated_to_genomic.bam
            filesize             : 7978810705
            checksum             : 09e2d4d12bec19701b46b2b558a0b084

--------------------------------------------------------------------------------
    Parse Summary
        Parse Errors             :  0
        Parse Warnings           :  0
--------------------------------------------------------------------------------
```

5.4 CGHub REST API

The table below describes the query APIs available from CGHub.

	Resource URI	Purpose
Analysis Object Query	/cghub/metadata/analysisObject? {querystring}	Returns minimal fields to identify the desired object as well as a URI to download the data files.
Analysis Attribute Query	/cghub/metadata/analysisAttributes? {querystring}	Returns the full Analysis object attributes including the sample information and submission XML objects.

Details of the {querystring} are described in section 5.1 **Error! Reference source ot found.**.

The CGHub REST API returns query results in XML format. The results will be saved to a file if the –O <filename> option is used.

The **analysisObject** query returns minimal attributes to locate files for download. This format can be used to identify a list of sequences to pass as input to GeneTorrent to perform the transfer. It will return the following attributes:

- analysis_id
- last_modified
- center_name
- state
- aliquot_id
- files
    - filename
    - filesize
    - checksum
- analysis data URI (used to request a download of this object)
- analysis attributes URI (used to get the detailed attributes for this object)

The **analysisAttributes** query returns all of the analysisObject attributes as well as the following additional attributes:

- study

- sample_accession
- legacy_sample_id
- disease_abbr
- analyte_code
- sample_type
- tss_id
- participant_id
- sample_id
- XML documents from the original submission
    - analysis.xml
    - run.xml
    - experiment.xml

5.5 Sample REST Calls

The queries will return XML formatted data, including a URL to download the associated data files for each returned analysis object.

For example enter the following query into a Web browser:

```
https://cghub.ucsc.edu/cghub/metadata/analysisObject?aliquot_id=
c0cfafbc-6d07-4ed5-bfdc-f5c3bf8437f6
```

OR execute the command line:

```
wget --no-check-certificate -O xmlout
https://cghub.ucsc.edu/cghub/metadata/analysisObject?aliquot_id=
c0cfafbc-6d07-4ed5-bfdc-f5c3bf8437f6
```

The returned XML will look like:

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<ResultSet date="2012-01-09 09:19:22">
        <Query>aliquot_id:c0cfafbc-6d07-4ed5-bfdc-f5c3bf8437f6</Query>
        <Hits>1</Hits>
        <Result id="1">
                <analysis_id>8d0b3af6-0f78-4282-bde5-
c5238fa3d4c2</analysis_id>
                <state>live</state>
                <last_modified>2011-08-10T07:00:00Z</last_modified>
                <upload_date></upload_date>
                <center_name>BCCAGSC</center_name>
                <study>phs000178</study>
                <aliquot_id>c0cfafbc-6d07-4ed5-bfdc-
f5c3bf8437f6</aliquot_id>
```

```
                <files>
                        <file>
                                <filename>TCGA-CZ-5466-11A-01R-1502-
13_mirna.bam</filename>
                                <filesize>132976422</filesize>
                                <checksum
type="MD5">6ede057c98171b986a36e77a9bf7efc0</checksum>
                        </file>
                </files>

<analysis_attribute_uri>https://cghub.ucsc.edu/cghub/metadata/analysisA
ttributes/8d0b3af6-0f78-4282-bde5-c5238fa3d4c2</analysis_attribute_uri>

<analysis_data_urihttps://cghub.ucsc.edu/cghub/data/analysis/download/
8d0b3af6-0f78-4282-bde5-c5238fa3d4c2</analysis_data_uri
        </Result>
</ResultSet>
```

The results return the same attribute fields that a *cgquery* search does except that they are in XML format.

The same query can be done to return the full attribute set by substituting “analysisAttribute” for “analysisObject” in the query URI. This query saves the results to the file “xmlout”.

```
]$ wget --no-check-certificate -O xmlout
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes?aliquot_id=c0c
fafbc-6d07-4ed5-bfdc-f5c3bf8437f6
--2012-01-09 12:44:48--
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes?aliquot_id=c0c
fafbc-6d07-4ed5-bfdc-f5c3bf8437f6
Resolving cghub.ucsc.edu... 192.35.223.5
Connecting to cghub.ucsc.edu|192.35.223.5|:443... connected.
WARNING: certificate common name “*.cghub.ucsc.edu” doesn’t match
requested host name “cghub.ucsc.edu”.
HTTP request sent, awaiting response... 200 OK
Length: 3072 (3.0K) [text/xml]
Saving to: “xmlout”

100%[=======================================================================
==================================================>] 3,072       --.-
K/s   in 0s

2012-01-09 12:44:54 (6.74 MB/s) - “xmlout” saved [3072/3072]
```

Note that the results contain the raw XML in the <analysys_xml> section.

```
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<ResultSet date="2012-01-09 12:44:54">
      <Query>aliquot_id:c0cfafbc-6d07-4ed5-bfdc-f5c3bf8437f6</Query>
      <Hits>1</Hits>
      <Result id="1">
            <analysis_id>8d0b3af6-0f78-4282-bde5-
c5238fa3d4c2</analysis_id>
            <state>live</state>
            <last_modified>2011-08-10T07:00:00Z</last_modified>
            <upload_date></upload_date>
            <center_name>BCCAGSC</center_name>
            <study>phs000178</study>
            <aliquot_id>c0cfafbc-6d07-4ed5-bfdc-
f5c3bf8437f6</aliquot_id>
            <files>
                  <file>
                        <filename>TCGA-CZ-5466-11A-01R-1502-
13_mirna.bam</filename>
                        <filesize>132976422</filesize>
                        <checksum
type="MD5">6ede057c98171b986a36e77a9bf7efc0</checksum>
                  </file>
            </files>
            <sample_accession>SRS238705</sample_accession>
            <legacy_sample_id>TCGA-CZ-5466-11A-01R-1502-
13</legacy_sample_id>
            <disease_abbr>KIRC</disease_abbr>
            <tss_id>CZ</tss_id>
            <participant_id>5722df9f-5631-476d-a11b-
b3c1e9a40fbf</participant_id>
            <sample_id>f072dc5f-6b9a-4935-8ceb-f9f3fd7ebe73</sample_id>
            <analyte_code>R</analyte_code>
            <sample_type>11</sample_type>
            <library_strategy></library_strategy>
            <platform></platform>
            <analysis_xml><ANALYSIS_SET
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <ANALYSIS center_name="BCCAGSC" analysis_date="2011-07-27T18:28:48Z"
analysis_center="BCCAGSC" alias="M02084" accession="SRZ017243">
    <TITLE>Reference alignment of M02084</TITLE>
    <STUDY_REF accession="SRP000677"/>
    <DESCRIPTION/>
    <ANALYSIS_TYPE>
      <REFERENCE_ALIGNMENT>
        <ASSEMBLY>
          <STANDARD short_name="GRCh37-lite"/>
        </ASSEMBLY>
        <PROCESSING>
          <PIPELINE>
            <PIPE_SECTION section_name="Alignment">
              <STEP_INDEX>1</STEP_INDEX>
              <PREV_STEP_INDEX>NIL</PREV_STEP_INDEX>
              <PROGRAM>BWA</PROGRAM>
              <VERSION>0.5.7</VERSION>
              <NOTES/>
```

```
            </PIPE_SECTION>
          </PIPELINE>
          <DIRECTIVES>

<alignment_includes_unaligned_reads>true</alignment_includes_unaligned_
reads>

<alignment_marks_duplicate_reads>true</alignment_marks_duplicate_reads>

<alignment_includes_failed_reads>true</alignment_includes_failed_reads>
          </DIRECTIVES>
        </PROCESSING>
      </REFERENCE_ALIGNMENT>
    </ANALYSIS_TYPE>
    <TARGETS>
      <TARGET accession="SRS238705" refcenter="TCGA"
sra_object_type="SAMPLE" refname="TCGA-CZ-5466-11A-01R-1502-13"/>
    </TARGETS>
    <DATA_BLOCK>
      <FILES>
        <FILE checksum="6ede057c98171b986a36e77a9bf7efc0"
filetype="bam" filename="TCGA-CZ-5466-11A-01R-1502-13_mirna.bam"
checksum_method="MD5"/>
      </FILES>
    </DATA_BLOCK>
    <ANALYSIS_ATTRIBUTES>
      <ANALYSIS_ATTRIBUTE>
        <TAG/>
        <VALUE/>
      </ANALYSIS_ATTRIBUTE>
    </ANALYSIS_ATTRIBUTES>
  </ANALYSIS>
</ANALYSIS_SET>
</analysis_xml>
            <run_xml></run_xml>
            <experiment_xml></experiment_xml>

      <analysis_data_urihttps://cghub.ucsc.edu/cghub/data/analysis/dow
nload/8d0b3af6-0f78-4282-bde5-c5238fa3d4c2</analysis_data_uri
      </Result>
</ResultSet>
```

6 Sequence Download

To download sequence data files from the CGHub system securely, efficiently and reliably, users are required to use the GeneTorrent download client, an open-source application available at no cost to users.

6.1 Downloading Files using GeneTorrent Client

The GeneTorrent client runs as a command line tool on all systems

The download mode command line arguments are:

GeneTorrent -d [URI | UUID | .xml | .gto] -c credentials [-p path]

-d content-specifier, --download content-specifier

The content-specifier should be one of the following:

- A fully-qualified URI pointing to an analysis object at the GeneTorrent Executive.
- A UUID denoting an analysis object at the GeneTorrent Executive. In this case, GeneTorrent will construct a URI based on the default server, currently https://cghub.ucsc.edu.
- An XML file, which will be parsed to obtain a list of URIs.
- A .gto file directly, which eliminates all calls to the GeneTorrent Executive and causes a download to begin. In this case you do not need to supply access credentials using the -c option.

Note that multiple -d options can be specified on the same command line, in which case multiple analysis objects will be downloaded.

-c credential-file, --credentialFile credential-file

A credential file is required for download mode, unless you are directly passing a .gto file. Pass the full or relative path to the file containing the access credentials (security token) previously received from the User Authentication, (see section 4 for details).

-p path, --path path

Path to save data files in the gto file(s). UUID is part of the gto and will always be added to path, so data files will be found at path/UUID. The current directory will be used by default.

--maxChildren max-children

The maximum number of parallel children that should be spawned to perform the download. By default up to 8 children are used, but this number may be adjusted. In general you should have fewer children than you have CPU cores on your machine.

6.2 Query and Download

Step 1: Query for the desired files

```
% cgquery "center_name=CGHubTest"

================================================================================
    Script Version               : 1.9
    CGHub Server                 : https://cghub.ucsc.edu
    REST Resource                : /cghub/metadata/analysisObject
    QueryString                  : center_name=CGHubTest
    Output File                  : None
--------------------------------------------------------------------------------

    Result 1
        analysis_id              : FC31132C-CD53-44D5-A9F2-0F4AEE52965C
        analysis_data_uri        :
https://cghub.ucsc.edu/cghub/data/analysis/download/FC31132C-CD53-44D5-
A9F2-0F4AEE52965C
        analysis_attribute_uri   :
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes/FC31132C-CD53-
44D5-A9F2-0F4AEE52965C
        last_modified            : 0011-09-18 12:00:00.0
        center_name              : CGHubTest
        state                    : live
        aliquot_id               : 1643378B-CBAF-43E7-A9A0-157EB1FF3F02
        study                    : CGHUB
        Files
            filename             :
HG00114.mapped.ILLUMINA.bwa.GBR.low_coverage.20101123.bam
            filesize             : 11643625129


    Result 2
        analysis_id              : 0344FD0B-3435-4405-B284-C462849DF935
        analysis_data_uri        :
https://cghub.ucsc.edu/cghub/data/analysis/download/0344FD0B-3435-4405-
B284-C462849DF935
        analysis_attribute_uri   :
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes/0344FD0B-3435-
4405-B284-C462849DF935
        last_modified            : 0011-09-17 12:00:00.0
        center_name              : CGHubTest
        state                    : verifying_sample
        aliquot_id               : 38220E07-7242-4274-8155-24E45972A08F
        study                    : CGHUB
        Files
```

```
            filename              :
HG00099.unmapped.SOLID.bfast.GBR.low_coverage.20101123.bam
            filesize              : 24032208912
```

Step 2: Invoke GeneTorrent on the command line with the desired file URI(s) and user credentials. Multiple –d options can be provided to download multiple files.

```
GeneTorrent -v -c <credential> -d
https://cghub.ucsc.edu/cghub/data/analysis/download/FC31132C-CD53-44D5-
A9F2-0F4AEE52965C
```

The GeneTorrent client connects to the CGHub system, bi-directionally authenticates using the credential, verifies that the user is authorized to access the file, and starts the SSL-secured file transfer.

GeneTorrent, by default, runs silently to completion. If invoked with the '-v' (verbose) option, the user will see:

```
Welcome to GeneTorrent version x.y.z, download mode.
Status:  checking (r)   downloaded:          (         )
uploaded:          (         )
Status:  downloading    downloaded:  7.34MB (718.kB/s)
uploaded:  9.88kB (21.7kB/s)
Status:  downloading    downloaded:  14.4MB (1.33MB/s)
uploaded:  19.5kB (39.3kB/s)
```

Step 3: Verify that the download was successful.
When GeneTorrent finishes transferring the file, the user will find the BAM file in the current working directory, or in the location specified by the '-p' (path) option.

The file length and the MD5 checksum of the file can be verified using standard tools provided by the operating system.

6.3 Query and Download using *cgquery* Interactive Mode

The *cgquery* –i option will prompt the user to download one or more of the returned files using GeneTorrent. The interactive command requires the user credentials.

```
% cgquery –c mykey.pem –i "filename=HG00*"

================================================================================
    Script Version              : 1.9
    CGHub Server                : https://cghub.ucsc.edu
    REST Resource               : /cghub/metadata/analysisObject
    QueryString                 : filename=HG00*
```

```
    Output File                  : None
------------------------------------------------------------------------

    Result 1
        analysis_id              : FC31132C-CD53-44D5-A9F2-0F4AEE52965C
        analysis_data_uri        :
https://cghub.ucsc.edu/cghub/data/analysis/download/FC31132C-CD53-44D5-
A9F2-0F4AEE52965C
        analysis_attribute_uri   :
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes/FC31132C-CD53-
44D5-A9F2-0F4AEE52965C
        last_modified            : 2011-09-18T19:00:00Z
        center_name              : CGHubTest
        state                    : live
        aliquot_id               : 1643378B-CBAF-43E7-A9A0-157EB1FF3F02
        study                    : CGHUB
        Files
            filename             :
HG00114.mapped.ILLUMINA.bwa.GBR.low_coverage.20101123.bam
            filesize             : 11643625129

    Result 2
        analysis_id              : 0344FD0B-3435-4405-B284-C462849DF935
        analysis_data_uri        :
https://cghub.ucsc.edu/cghub/data/analysis/download/0344FD0B-3435-4405-
B284-C462849DF935
        analysis_attribute_uri   :
https://cghub.ucsc.edu/cghub/metadata/analysisAttributes/0344FD0B-3435-
4405-B284-C462849DF935
        last_modified            : 2011-09-17T19:00:00Z
        center_name              : CGHubTest
        state                    : live
        aliquot_id               : 38220E07-7242-4274-8155-24E45972A08F
        study                    : CGHUB
        Files
            filename             :
HG00099.unmapped.SOLID.bfast.GBR.low_coverage.20101123.bam
            filesize             : 24032208912


Enter index of URI to download (0 for all) or 'q' to quit
    [ 0] : All URIs
    [ 1] : https://cghub-
test.ucsc.edu/cghub/data/analysis/download/FC31132C-CD53-44D5-A9F2-
0F4AEE52965C
    [ 2] :
https://cghub.ucsc.edu/cghub/data/analysis/download/0344FD0B-3435-4405-
B284-C462849DF935

Index> 2

Processing URI
https://cghub.ucsc.edu/cghub/data/analysis/download/0344FD0B-3435-4405-
B284-C462849DF935

Executing command '/usr/bin/GeneTorrent -v -c mykey.pem -d
https://cghub.ucsc.edu/cghub/data/analysis/download/0344FD0B-3435-4405-
```

```
B284-C462849DF935'

Welcome to GeneTorrent version x.y.z, download mode.
Status:  checking (r)   downloaded:         (        )
uploaded:         (        )
Status:  downloading    downloaded:  7.34MB (718.kB/s)
uploaded:  9.88kB (21.7kB/s)
Status:  downloading    downloaded:  14.4MB (1.33MB/s)
uploaded:  19.5kB (39.3kB/s)
```