GDC Office Hours Questions and Answers
January 25, 2021
2:00 PM - 3:00 PM (EST)

1. Where can I find the GDC Data Portal?

The GDC Data Portal is hosted at https://portal.gdc.cancer.gov

2. How do I obtain access to controlled access data?

The process to apply for access to controlled access data is outlined in GDC website: https://gdc.cancer.gov/access-data/obtaining-access-controlled-data/ The process to access controlled data generally involves two steps: 1) apply for an eRA Commons ID; the ID will be your username and password for the GDC Data Portal and 2) apply for project access through dbGaP. dbGaP and eRA Commons use a project's 'phs' id for authorization to access a given project; if users cannot find the phs ID for a given project, they can contact the GDC Support Helpdesk at support@nci-gdc.datacommons.io

3. Is there a way to browse what controlled-access data are available for a project?

Users can navigate to the Repository page of the GDC Data Portal, click on the 'Cases' tab, and then click on a specific project in the 'Projects' facet. Controlled access files will have a locked padlock icon next to file name. Or, users can also go to the 'File' tab of the Repository page and click on 'controlled' for the 'Access' facet to browse controlled data for a project or set of cases. Here is also a link to browse all controlled access data in the GDC.

4. How can I download variant calls from specific variant callers in the GDC?

From the Repository page of the GDC Data Portal, click on the 'File' tab, and then choose the desired caller(s) from the 'Workflow Type' facet to view MAF and VCF files from the specified variant caller(s). The GDC uses five different variant callers so consensus can be found for likely variants from genomic samples, although there is no scientific consensus at this time for which caller is best. Users can view information for caller parameters used in the GDC harmonization DNA-Seq pipeline here. The five different variant calling pipelines are:

- MuSE
- MuTect2
- VarScan2
- SomaticSniper
- Pindel

5. How can I find and download project-specific data, including supplement files, metadata, biospecimen and clinical data?

One way to find project clinical data is from a given case's page (such as this case for the TCGA-KIRC project). There are "Clinical" and "Biospecimen" Sections of the case page, where users can browse through tabs with clinical and biospecimen data available for the case. There are also "Clinical Supplement File" and "Biospecimen Supplement File" subsections, respectively. Available in these subsections are BCR Biotab files that consist of aggregate clinical and biospecimen data across all cases of the given project. Additionally, XML files will have clinical data for the specific case, including data or metadata that may be excluded from the aggregate clinical BCR Biotab files.

BCR Biotab files and individual case XML files can also be found from the GDC Data Portal, by filtering down the 'Data Format' facet to 'bcr xml' or 'bcr biotab' in the 'Files' tab of the Repository page.

Another method is to retrieve clinical, biospecimen and metadata for cases associated with files in a user's 'Cart' page in the GDC Data Portal. After adding these files to their GDC Data Portal 'cart' and navigating to the 'Cart' page, users can click on the following links to download:

i. **Biospecimen** information for cases associated with files in the cart, such as aliquot and sample ids. Users can download information as TSV or JSON files.
ii. **Clinical** information for cases associated with files in the cart, such as available exposure and family history data for cases. Users can download information as TSV or JSON files.
iii. A **Sample sheet** which contains information about the files in the cart, such as project and case IDs, in TSV format.
iv. **Metadata** information in JSON format for files in the cart, including file size, case IDs and any annotation information for cases associated with the files.

Lastly, users are welcome to utilize the GDC Community Tool, **GDC TSV Downloader**, which will download clinical and biospecimen data for files specified in a GDC download manifest. Directions and installation of the GDC TSV Downloader can be found at the GDC Community Tools page.

6. How can I use case sets within the GDC Data Portal?

On the Exploration page of the GDC Data Portal, users can toggle filters from the tabs on the left-hand side to create a custom 'set' of cases. This will also update visualizations (i.e. pie charts, survival plots, OncoGrid etc.) on the Exploration page relative to the cases being filtered. Users can then click the 'Save/Edit Case Set' link to save a set of cases that have been filtered down for later reference, data download, and data analysis. In the case of saving case sets, users can also save the 'top N' number of cases as well.

Users can upload custom case sets, either from the 'Manage Sets' link at top of GDC Data Portal page, or from the 'Upload Case Set' links on the Exploration or Repository pages of the GDC Data Portal. After uploading a case set, users can view all files in the Repository page for cases in the set and can further refine results in the Exploration or Repository pages with facets.

Additionally, users can also save sets of genes and mutations as well, covered in the GDC Documentation site here. Please note that saved sets are stored for browser's current session, and that any saved sets will be removed from the browser with each new GDC data release. It is therefore recommended to save sets locally. To download a set, users can click 'Manage Sets' link at top of GDC Data Portal save sets as a TSV for future reference, as well as view and manage sets.

7. What are the GDC Data Portal's Analysis tools?

After saving case sets per the instructions above, users can navigate to the Analysis page of the GDC Data Portal and perform set operations, clinical data analysis and cohort comparisons on case sets. For instance, in the Clinical Data Analysis tab, users can create and customize summary bar plots and survival pots on chosen your case set. Users can remove or add data items, and aggregate data bins for categorical data types (i.e. vital status) as well as edit ranges for continuous data (i.e. age).

Users can also perform set operations on saved case, gene or mutation sets using the Set Operations tab. Users can choose 2 or 3 sets of the same type and produce a Venn diagram that allows users to inspect the intersection and unique values of the sets. From this analysis, users can download the sets, create new sets and inspect files associated with the intersection and unique set values in the GDC Data Portal Repository.

Comprehensive documentation on using data analysis tools in the GDC Data Portal can be viewed here, and users are welcome to review past webinars that cover these tools here at the GDC Support Webinar page.

8. What does the Data Submission Process generally entail?

The GDC accepts data from different cancer study groups that submit molecular, clinical and biospecimen data. Submitters work with the GDC to create new data types that may not be present in the GDC Data Dictionary so that proposed study can fit within the GDC Data Model. The submitted molecular data is subsequently processed and harmonized with bioinformatics pipelines and is then hosted on the GDC Data Portal. Future submitters should ensure the proposed study for submission meets one or more of the following criteria:

i. Studies designed to provide a definitive answer to an important question in cancer genomic biology
ii. Studies that fill in gaps in knowledge of rare cancers, or special pathologic or therapeutic circumstances

iii. Genomic studies that are incorporated into clinical trials, that are well-annotated, and possess outcome data
iv. Genomic study data that underlie well-known published results of high impact

More information on data submission to the GDC can be found at the following links:

https://docs.gdc.cancer.gov/Data_Submission_Portal/Users_Guide/Checklist/

https://gdc.cancer.gov/submit-data/requesting-data-submission

9. Where can I find information about the Publication pages and GDC Community Tools?

Publications that feature GDC hosted data can be found at this link, which also can be found from the main GDC Website by clicking the 'About the Data' header, and then clicking 'Publications' in the drop down banner. Users can filter through publications using keywords or program name on the Publications page.

In addition to displaying download manifests for data hosted at the GDC Portal, Publication pages also may have download manifests for other data from the publication that may be relevant (e.g. supplementary tables) but does not fit into the GDC Data Model. Data featured on the publication pages can be downloaded using their associated manifest files and with the GDC Data Transfer Tool, with data being either open or controlled access. Users are welcome to edit the manifests to only include files they would like to download.

Controlled access data on Publication pages will require appropriate access to the project(s) that the data is sourced from. For instance, publications with controlled access data from PanCanAtlas projects will require access to each project in the PanCanAtlas separately. Users can follow the instructions here to request access to project data.

GDC Community Tools are third party tools that are not developed or maintained by the GDC, but can be useful for users of the GDC. The list of tools can be found here from the main GDC website. Available tools include **gdc-readgroups**, which helps users submit BAM files to the GDC, and the **GDC TSV Downloader**, which will download biospecimen and clinical data for given files in a GDC manifest file. Users should be aware that because GDC Community tools are not maintained by the GDC, users should contact authors of these tools for issues and assistance.

10. Where is the User Support section of GDC website?

The GDC User Support site is hosted at https://gdc.cancer.gov/support. The support webpage hosts links to past webinar videos and other guides on how to make use of resources from the GDC. Also pertinent is the 'Resources for TCGA Users page, https://gdc.cancer.gov/resources-tcga-users, that has information such as how to use the TCGA Legacy Archive and information

tables for TCGA Tags, descriptors for data levels, center codes, etc. You can contact GDC Support Helpdesk at support@nci-gdc.datacommons.io.

11. Is there a place or way to find out which GDC datasets were most frequently downloaded by users?

This data is not public, but users are welcome to check the GDC Publications page to see how GDC data is being used and by whom.