Uncovering Patterns in Genomic Variation Data with Clinical Phenotypes Using the New GDC Correlation Plot Tool

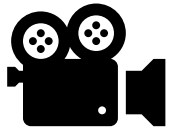
April 29, 2026

Bill Wysocki, Ph.D – Director of User Services; University of Chicago
Xin Zhou, Ph.D – Director of Data Visualization; St. Jude

NIH NATIONAL CANCER INSTITUTE

Webinar Details

- ***Webinar will be recorded***



- ***Recording and slides will be made available soon***

- ***Type any questions in the Q&A panel – they will be addressed at the end***

Agenda

1. *Introduction to GDC Data*
2. *Correlation Plot Tool*
3. *Correlation Plot Demo*
4. *Q&A session*

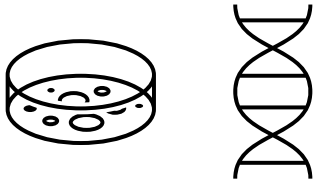
Introduction to GDC Data

What goes into the Correlation Plot

Two Main Data Types Available in Correlation Tool

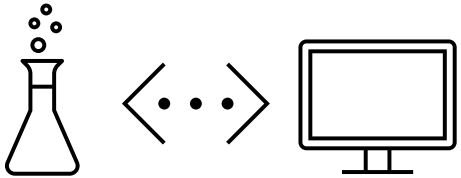
Directly Submitted Metadata

- Case properties
- Clinical properties



Molecular Data Generated by the GDC

- Somatic mutations
- Copy number estimates
- Gene expression data



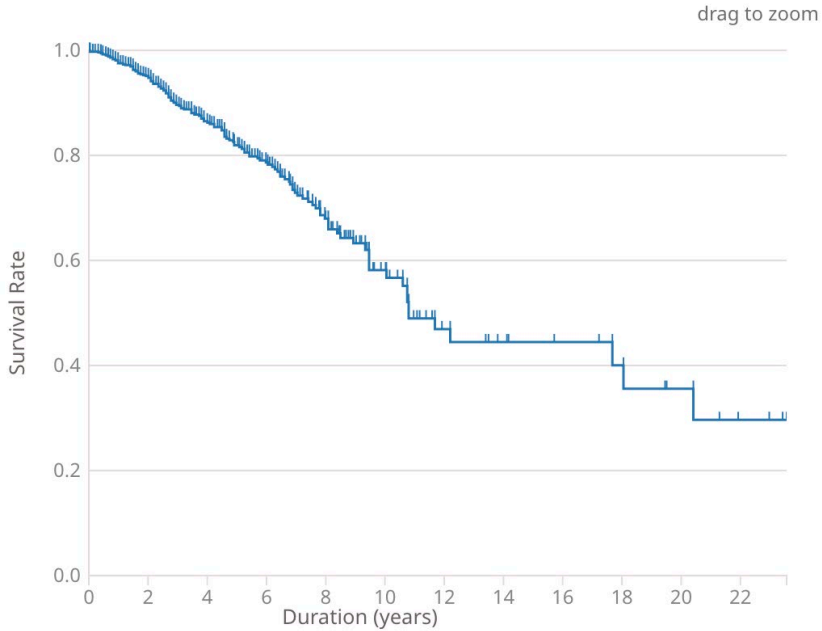
Directly Submitted Metadata (1/2)

Some types of GDC data are uploaded directly by external submitters with adherence to GDC's data model.

- ***Case Data***
  - Information directly about cancer patients and cell models
  - Primary site – Disease type – Lost to follow up
- ***Clinical Data***
  - Phenotypic data about a case that was collected in a clinical setting
  - Diagnosis – Demographic – Family History

Directly Submitted Metadata (2/2)

- ***Survival Data***
  - Derived from several clinical time-point values
    - Days to Death
    - Days to Last Follow Up



GDC-Generated Data (1/3)

Other types GDC data are derived from molecular data and generated using GDC's pipelines

- ***Somatic Mutation***
  - Single-nucleotide variation (SNV) data or small indels that were generated from somatic variant callers
  - Raw somatic variants are annotated and conservatively filtered to remove the chance of germline leakage
  - Data is available through the /ssm, /ssm_occurrences endpoints or in the "Masked Somatic Mutation" MAF files

Correlation Plot Example: SNV/Indel in TP53

GDC-Generated Data (2/3)

- ***Copy Number Estimate***
  - Classification based on copy number estimate integer
  - Amplification > Gain > WildType > Heterozygous Deletions > Homozygous Deletions
  - Based on several experimental strategies and workflows including:
    - ASCAT-NGS from Whole Genome Sequencing
    - ASCAT3 from SNP6 Array data
    - ABSOLUTE liftover from SNP6 Array data

Correlation Plot Example: KRAS CNV Gain

GDC-Generated Data (3/3)

- ***Gene Expression***
  - Generated from STAR alignment followed by STAR counts and FPKM-UQ
  - Values may be binned or continuous
  - Data is available through the /gene_expression/values endpoints
  - Originates from the “Gene Expression Quantification” RNA-Seq files

Correlation Plot Example: TP53 FPKM-UQ - 15 to < 20

More Information

- This is the subset of the data available at the GDC that can be analyzed with the GDC Correlation Plot Tool
- Many other data types are available from the GDC as other file types that can be downloaded from the repository

- For more information on all data available in the GDC:
  - Documentation Site: https://docs.gdc.cancer.gov
  - Explore the Repository:
    - https://portal.gdc.cancer.gov → Repository Link

Additional Information for Users

- GDC Website: https://gdc.cancer.gov
  - Previous webinars
- GDC Documentation: https://docs.gdc.cancer.gov/
  - User's guides
  - Pipeline overviews
  - Developer info
- Questions? Contact the GDC Help Desk
  - **support@nci-gdc.datacommons.io**

U.S. Department of Health & Human Services
National Institutes of Health | National Cancer Institute

https://www.cancer.gov/

1-800-4-CANCER

Produced February 2026