Gene Expression Clustering in GDC 2.0

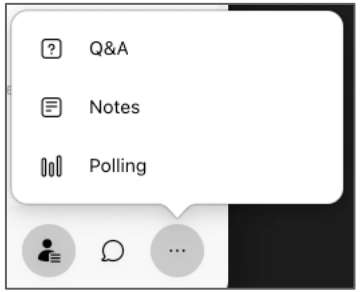
15 July 2024

Bill Wysocki, Ph.D. – GDC User Services Lead
Center for Translational Data Science
University of Chicago

NIH NATIONAL CANCER INSTITUTE

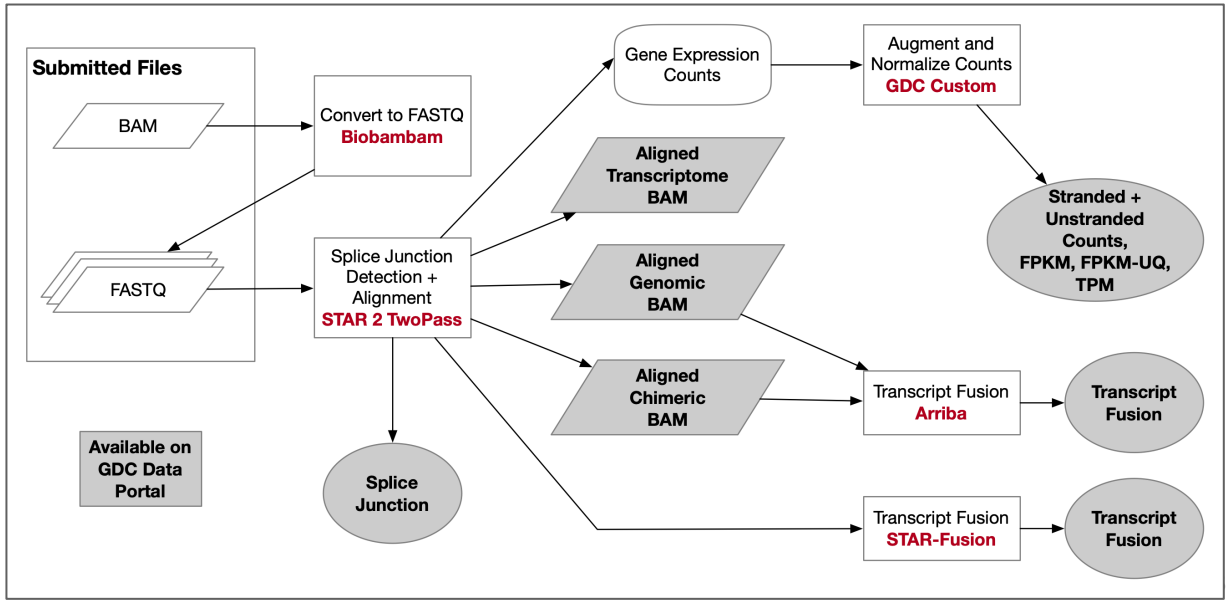
Agenda

1. *Overview of Gene Expression in the GDC*
2. *Demo: Gene Expression Clustering Tool*
3. *Demo: Gene Expression API*
4. *Downloading Gene Expression Files*
5. *Q&A*



Gene Expression in the GDC – Pipeline

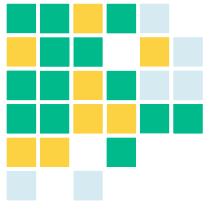
1. Tumor or normal reads are aligned to the GRCh38 reference genome with STAR (BAM x 3; Genomic, Transcriptome, Chimeric)
2. STAR is used to generate counts for each gene
3. Custom scripts in the GDC normalize the counts to:
   - TPM
   - FPKM
   - FPKM-UQ
4. Available via TSV file



Accessing Gene Expression Data in the GDC

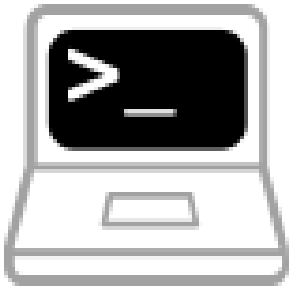
1. Gene Expression Clustering Tool

- Visualize gene expression data within your cohort
- Use a custom or curated set of genes

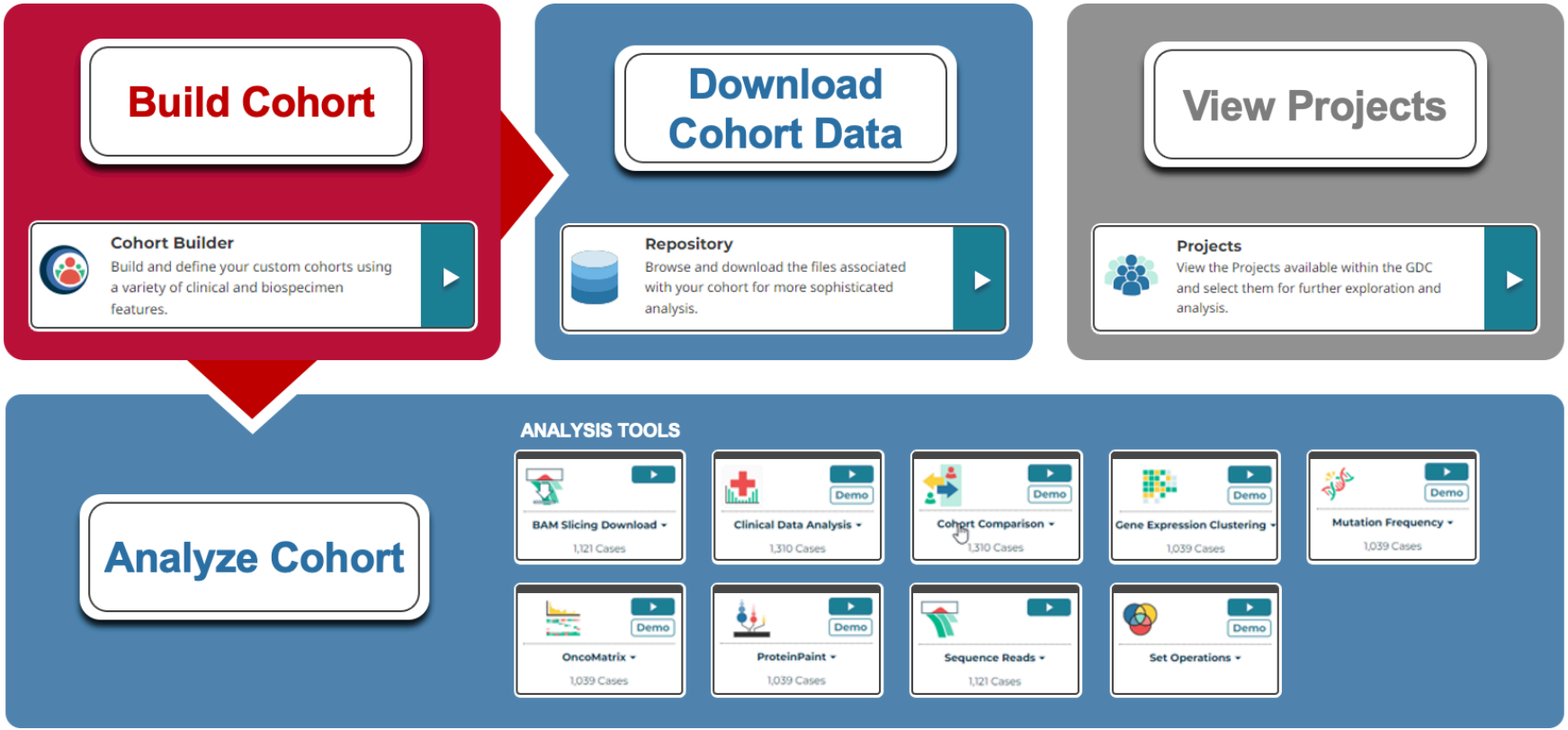


2. Gene Expression API

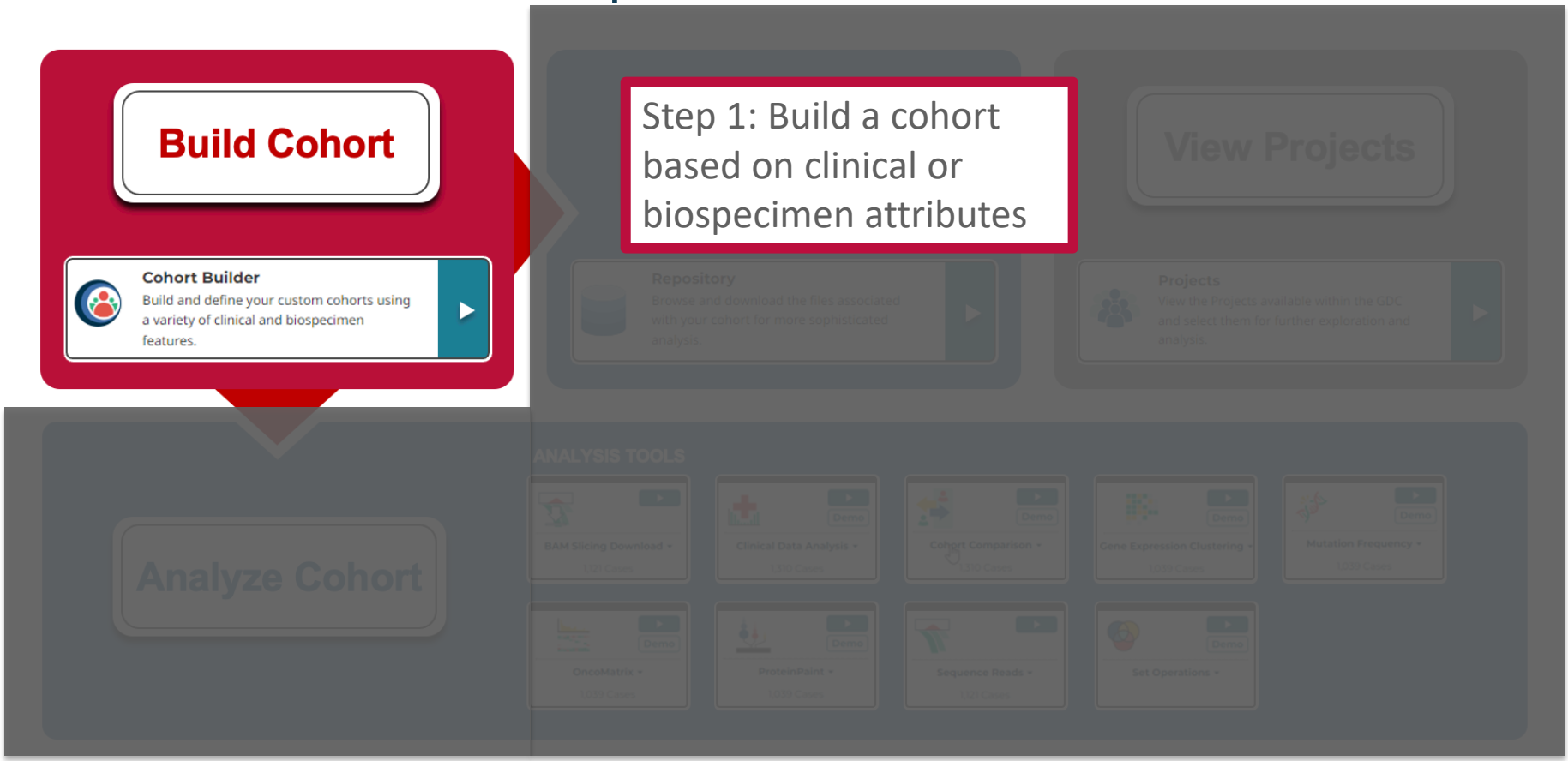
- Programmatically retrieve gene expression values
- Use a custom set of genes
- Values only



GDC 2.0 Workflow



GDC 2.0 Workflow: Step 1



GDC 2.0 Workflow: Step 2

Step 2: Use the cohort with tools in the analysis center.

Tools will be automatically applied to the cohort.

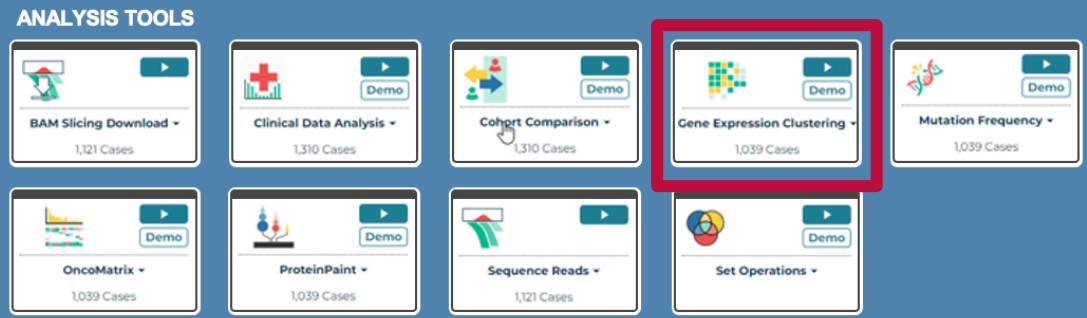
Download Cohort Data

Repository
Browse and download the files associated with your cohort for more sophisticated analysis.

View Projects

Projects
View the Projects available within the GDC and select them for further exploration and analysis.

Analyze Cohort



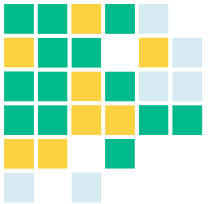
Demo Section

Demo:
Creating a Cohort for Gene Expression Analysis

Building a Cohort

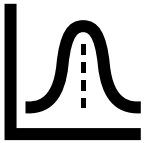
- **Goal:** Create two cohorts for gene expression analysis
  - General
    - Program
      - TCGA
  - **Cohort 1** - Tissue or Organ of Origin:
    - **lung, nos**
  - **Cohort 2** - Tissue or Organ of Origin:
    - **middle lobe, lung**

Demo: Gene Expression Clustering in the GDC Data Portal



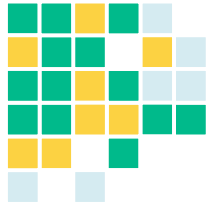
Gene Expression Clustering – Steps

Which cases have gene expression data?



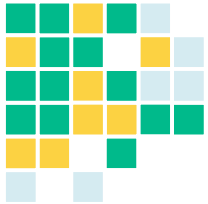
Which genes are most variably expressed?

What are the expression values?



Gene Expression Clustering

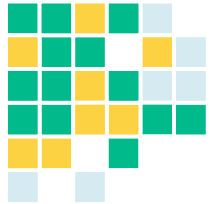
Tool Layout



- Heatmap Matrix
  - Rows – Genes
  - Columns – Cases
- Cell
  - Low → High
  - Hover:
    - Case identifier
    - Gene
    - Z-score transformed gene expression value
- Dendrogram
  - Represents how close expression profiles are between cases and genes

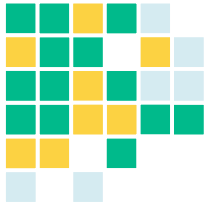
Toolbar – Clustering

- Clustering Method
- Dendrogram height/width
- Z-score cap
  - Controls color intensity threshold



Toolbar – Cases

- **Option:** Limit on case label length



- Cases are usually controlled by the cohort.

- Change or create a new cohort to change the cases in your matrix

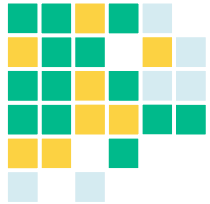
Toolbar – Genes

- **Option**: Limit on gene label length
- Starts with Cancer Gene Consensus genes

➢ Gene Set

  ➢ Edit Group

    - This changes the genes displayed and clustered in the matrix

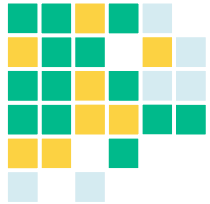


Edit Group – Methods

1. Enter genes in search box
2. Load top variably expressed
3. Pick from external database

Toolbar – Variables

- Add clinical/biospecimen variables to your matrix



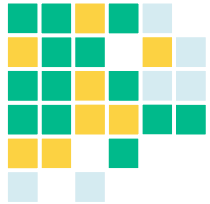
- **Option**: Limit on gene label length

- Select or search for clinical or biospecimen labels
    - Legend appears at the bottom

- Additional option to add a gene expression as a variable

Toolbar – Cell Layout

- Change colors of various visual features
- Visual options for font size / cell size

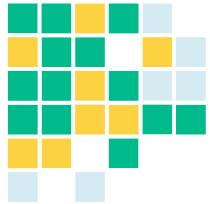


Toolbar – Legend Layout

- Change size of various legend features

Toolbar – Download

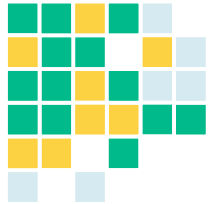
1. Download as Image (SVG)
2. Download as Data (TSV)
    - The TSV option is arranged in the opposite way as the matrix
    - Rows are cases, columns are genes



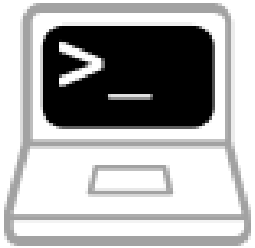
Gene Expression Clustering Misc. Functions

Misc. Functions

- **Zoom**
- **Click Gene**
  - Sort by gene expression level
  - Rename gene
  - Lollipop plot
  - Gene summary
- **Click Case**
  - Disco plot
  - Case summary

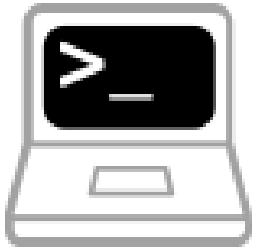


Demo:
Gene Expression API



Gene Expression API – Three Endpoints

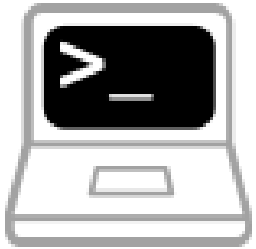
/gene_expression/{ ? }



1. **availability**
    - Displays whether a case or gene has GE data
2. **gene_selection**
    - Displays the most variably expressed genes
3. **values**
    - Displays the actual expression values

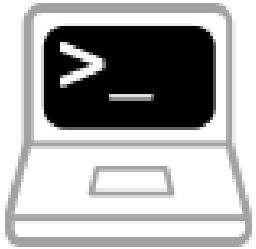
Gene Expression API – `availability`

1. List of cases
   - `case_ids`
   - Must be in **UUID** format
   - Can get from cohort builder export
2. List of genes
   - `gene_ids`
   - Must be in **Ensembl** format
   - Can get from gene set export



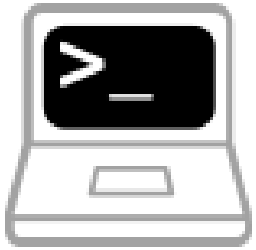
Gene Expression API – Command Syntax (curl)

- curl
- -X POST
- API URL + ENDPOINT
- -H “content-type: application/json”
- -d Payload.json



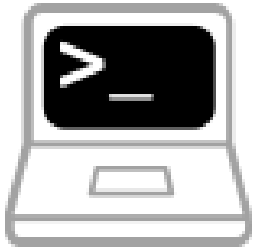
Gene Expression API – `gene_selection`

1. List of cases
   - `case_ids`
2. List of genes
   - `gene_ids`
   - `gene_type`
3. Top **n** of variably expressed genes
   - `selection_size`

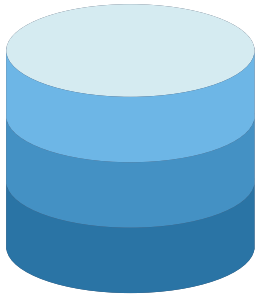


Gene Expression API – `values`

1. List of cases
   - `case_ids`
2. List of genes
   - `gene_ids`
3. Calculation type
   - `tsv_units`
   - Default: `uqfpkm`
   - Alt - `median_centered_log2_uqfpkm`

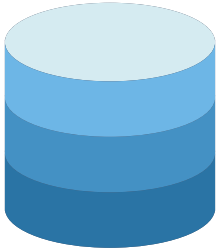


Demo: Downloading the Gene Expression Files



Gene Expression File Download

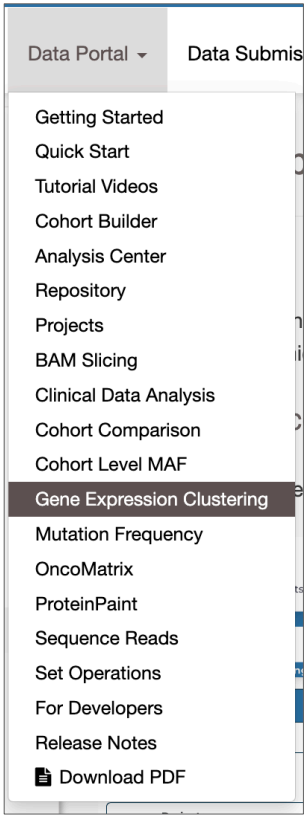
1. **GDC repository tool**
2. **Query:**
   - Experimental Strategy: RNA-Seq
   - Data Type: Gene Expression Quantification

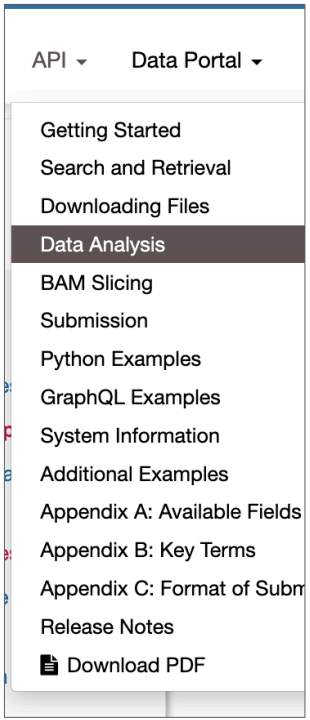


GDC Support

Gene Expression Clustering Documentation

https://docs.gdc.cancer.gov





GDC Support

- GDC Video Guides
- GDC User's Guides

https://docs.gdc.cancer.gov

- GDC Website

https://gdc.cancer.gov

- GDC Help Desk

Email:
support@nci-gdc.datacommons.io

Questions?

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

https://www.cancer.gov/

1-800-4-CANCER

Produced July 2024