GDC BAM Slicing

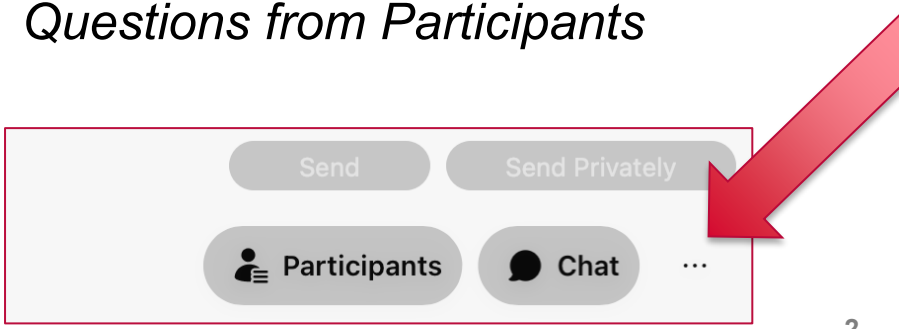
26 June 2023

GDC Monthly Webinar

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

GDC BAM Slicing Agenda

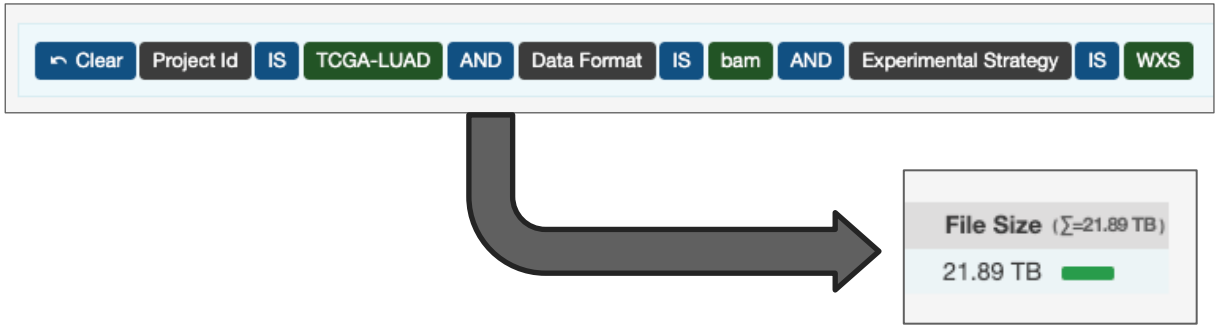
1. *Introduction to BAM Slicing*
2. *BAM Slicing from the Data Portal*
3. *BAM Slicing from the GDC API*
4. *BAM Slicing tips and troubleshooting*
5. *Questions from Participants*



1. Introduction to BAM Slicing

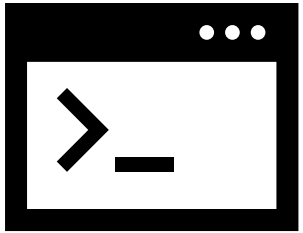
BAM Format Alignments

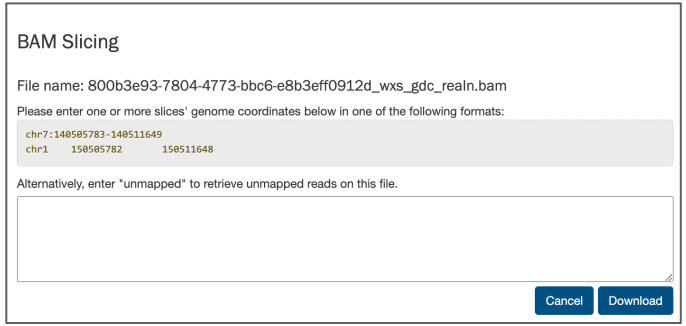
- Sequence Read Alignments in GDC are stored in Binary Alignment Map (BAM) format
- BAM files can be large in size, and the entire genome may not be required for your study
  - WGS Alignments in GDC – *Mean: 164 GB; Median: 103 GB*



GDC BAM Slicing – Decrease in File Size

- BAM Slicing is a feature that allows for a certain region or set of regions to be downloaded from the GDC in BAM format
  - Region can include entire chromosomes, specific ranges within chromosomes, and specific genes.
- Slicing can be performed using the API or using the GDC Data Portal





BAM Slicing Method Comparison

Description	Portal	API
Use genomic coordinates to specify region	✓	✓
Use GENCODE v36 gene symbols (e.g. KRAS) to specify region		✓
Run on small number of BAM files	✓	✓
Run on large number of BAM files		✓
Can be accessed programmatically		✓
User interface available	✓	
Requires dbGaP access	✓	✓

BAM Slicing Caveats

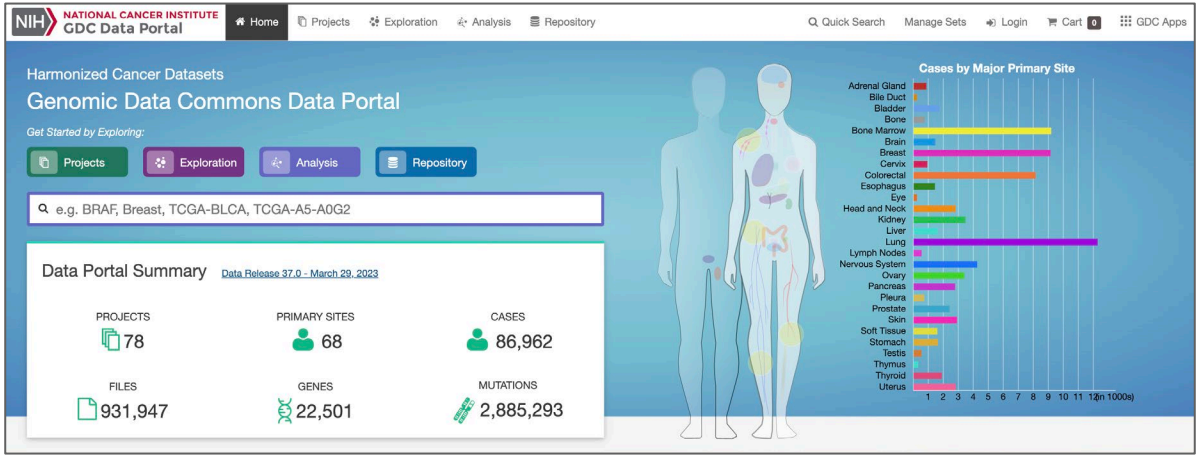
- All BAM files in the GDC are controlled-access. All slices of BAM files are controlled-access.
- BAM slicing cannot be performed on the RNA-Seq transcriptome BAM, due to their lack of sorting.
- Occasionally BAM files may be missing their index file (BAI) and cannot be sliced. These will need to be downloaded in full before they can be sliced.

2. BAM Slicing Using the Data Portal

User Interface

GDC Data Portal – BAM Slicing

- User interface can accept coordinates for a specific file.
- Recommended only if a few BAM files need to be sliced due to manual effort.
  - Great option for testing single BAM for larger pipeline



BAM File Selection

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository

Files | Cases

Add a File Filter

Search Files

e.g. 142682.bam, 4f6e2e7a-b...

Data Category

	# Files
sequencing reads	134,282

Data Type

	# Files
Aligned Reads	134,282

Experimental Strategy

	# Files
RNA-Seq	67,064
WXS	34,419
miRNA-Seq	16,822
WGS	14,285
Targeted Sequencing	1,252

Clear | Data Format | IS | bam

Files (134,282) | Cases (21,373)

Primary Site | Project | Data Ca

Showing **1** - **20** of **134,282** files **3.76 PB**

File UUID	File Name
88ee176a-a7e7-48f2-b332-6425d32bde07	800b3e93-7804-4773-bbc6-e8b3eff0912d_wxs_gdc_realn.bam
b581bca5-a250-4517-af9e-446c29a37608	11ed8e05-8f30-460a-b502-01ae09504315.rna_seq.chimeric.gdc_realn.bam

BAM File Operations

FL 800b3e93-7804-4773-bbc6-e8b3eff0912d_wxs_gdc_realn.bam

Add to Cart | BAM Slicing | Download

File Properties

Name	800b3e93-7804-4773-bbc6-e8b3eff0912d_wxs_gdc_realn.bam
Access	controlled
UUID	88ee176a-a7e7-48f2-b332-6425d32bde07
Data Format	BAM
Size	21.23 GB
MD5 Checksum	197a382e60601a21ed4eeb014bd4120d
Archive	--
Project	TCGA-BRCA

Data Information

Data Category	Sequencing Reads
Data Type	Aligned Reads
Experimental Strategy	WXS
Platform	Illumina

Showing **1** - **1** of **1** associated cases/biospecimen

Associated Cases/Biospecimen

Entity ID eg. TCGA-13*, *13*, *09

Entity ID	Entity Type	Sample Type	Case UUID	Annotations
TCGA-BH-A0BM-01A-11W-A071-09	aliquot	Primary Tumor	c2aeee6c-ba10-4cf9-b560-c739580f7bf3	0

Show 10 entries

« ‹ 1 › »

BAM Slicing

FL 800b3e93-7804-4773-bb

Add to Cart | BAM Slicing | Download

File Properties

Name	800b3e93-7804-4773-bbc6-e8b3eff0912d_wxs_gdc_realn.bam
Access	controlled
UUID	88ee176a-a7e7-48f2-b332-6425d32bde07
Data Format	BAM
Size	21.23 GB
MD5 Checksum	197a382e60601a21ed4eeb014bd4120d
Archive	--
Project	TCGA-BRCA

Data Information

Data Category	Sequencing Reads
Data Type	Aligned Reads
Experimental Strategy	WXS
Platform	Illumina

Showing **1** - **1** of **1** associated cases/biospecimen

Associated Cases/Biospecimen

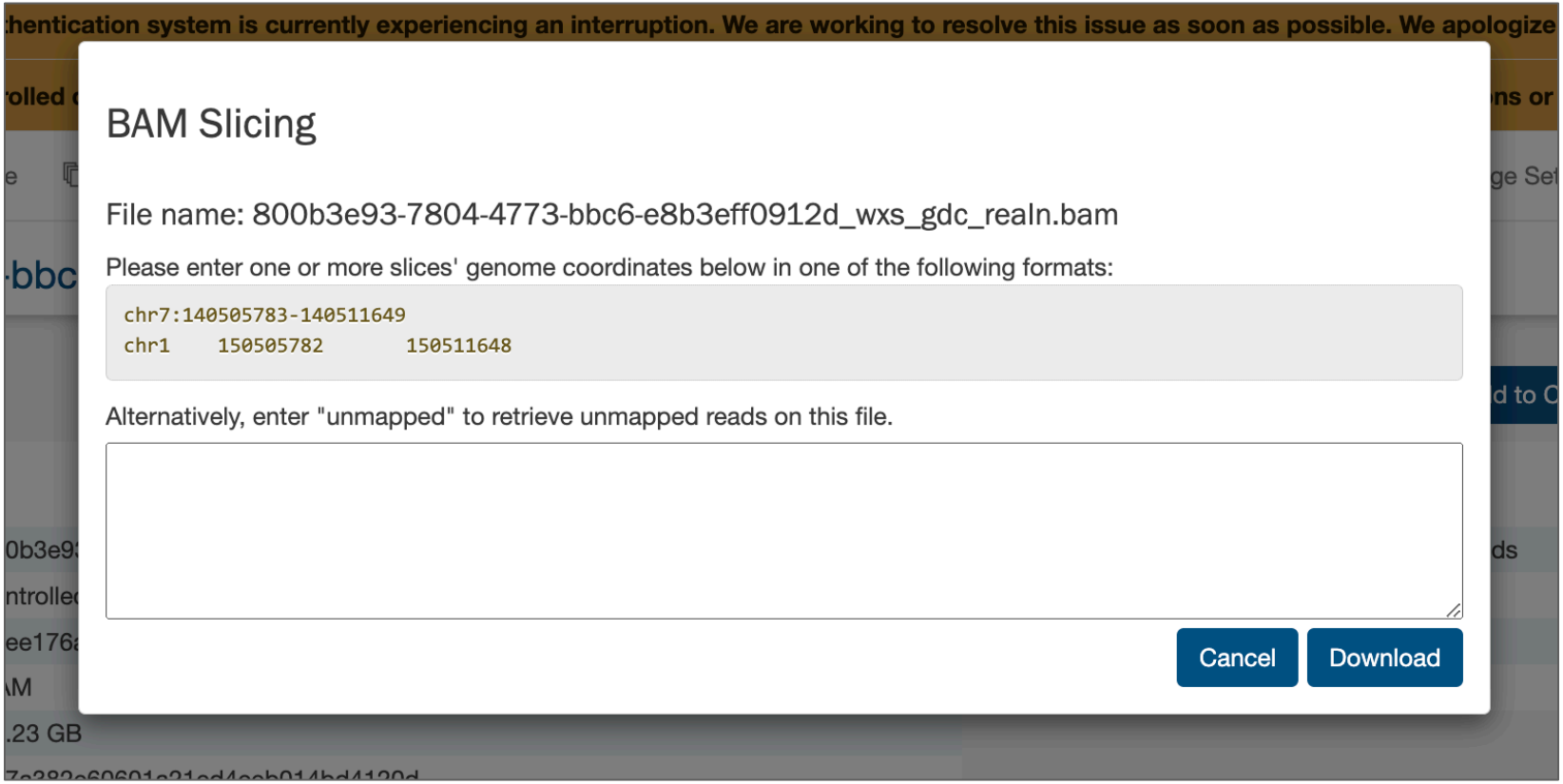
Entity ID eg. TCGA-13*, *13*, *09

Entity ID	Entity Type	Sample Type	Case UUID	Annotations
TCGA-BH-A0BM-01A-11W-A071-09	aliquot	Primary Tumor	c2aeee6c-ba10-4cf9-b560-c739580f7bf3	0

Show 10 entries

« ‹ 1 › »

BAM Slicing Genome Coordinate Entry



BAM Slicing Results

BAM Slicing

File name: 800b3e93-7804-4773-bbc6-e8b3eff0912d_wxs_gdc_realn.bam

Please enter one or more slices' genome coordinates below in one of the following formats:

```
chr7:140505783-140511649
chr1    150505782       150511648
```

Alternatively, enter "unmapped" to retrieve unmapped reads on this file.

```
chr19
chr2:1-20000
```

This will download:

- Entirety of Chromosome 19
- Bases 1 through 20,000 for Chromosome 2

Cancel | Download

BAM Slice Download

Download preparation in progress. Please wait…

Cancel Download

BAM Slicing

File name: 800b3e93-7804-4773-bbc6-e8b3eff0912d_wxs_gdc_realn.bam

Please enter one or more slices' genome coordinates below in one of the following formats:

```
chr7:140505783-140511649
chr1    150505782       150511648
```

Alternatively, enter "unmapped" to retrieve unmapped reads on this file.

```
chr19
chr2:1-20000
```

Cancel Download

3. BAM Slicing Using the API

Command Line Interface

BAM Slicing with the GDC API

- Recommended method for downloading slices from many BAMs.
  - Can be automated with bash scripting, Python, etc.

- Also allows for GENCODE v36 gene symbols to be specified.

```
(base) Bills-MacBook-Pro-2:BAM_Slice_Webinar billwysocki$ curl --header "X-Auth-Token: $token"
 'https://api.gdc.cancer.gov/slicing/view/88ee176a-a7e7-48f2-b332-6425d32bde07?gencode=KRAS'
--output gene.bam
```

BAM Slicing API Call Structure (curl GET)

```
curl # Default -XGET
--header "X-Auth-Token: $token" # Token in $token

https://api.gdc.cancer.gov/slicing/view/{UUID}? # URL
regions=chr1:start-end& # Region 1
regions=chr2:start-end& # Region 2
gencode={gene_symbol}   # Gene symbol

--output {file_name}.bam # Output file name
```

BAM Slicing API Call Structure (curl POST)

```
curl -XPOST # Specify POST
--header "X-Auth-Token: $token"
--header "Content-Type: application/json" # New Header
https://api.gdc.cancer.gov/slicing/view/{UUID}

--data @slice.json # Payload File
```

```
{
    "regions": [
        "chr19",
        "chr2:1-20000"
    ],
    "gencode": [
        "KRAS"
    ]
}
```

```
--output {file_name}.bam
```

BAM Slicing API Call Example

```
curl --header "X-Auth-Token: $token"
'https://api.gdc.cancer.gov/slicing/view/88ee176a-a7e7-48f2-b332-6425d32bde07?
region=chr2:1-20000&chr19&gencode=KRAS’
--output example_slice.bam
```

This will download:

- Bases 1 through 20,000 for Chromosome 2
- Entirety of Chromosome 19
- The KRAS gene on Chromosome 12

3. BAM Slicing Tips and Troubleshooting

BAM Slicing Tips (1/2)

- GDC BAM slicing does not return multiple copies of data when the same region is requested multiple times.
  - Example: A query for chr12 and KRAS (on chr12), will return the same result as just chr12.

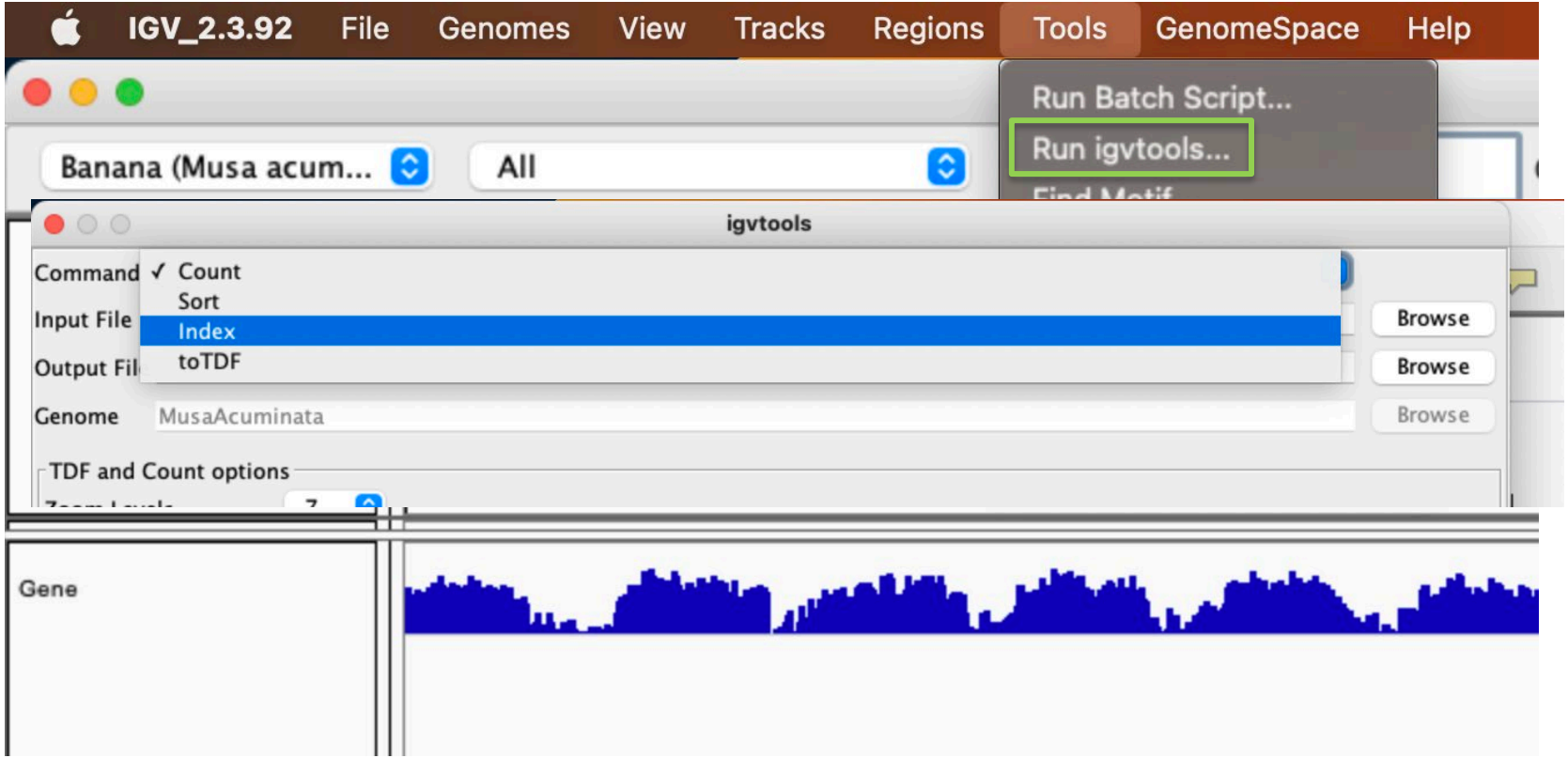
```
MD5 (KRAS_and_chr12.bam) = c14ceeb6c41196331ef18572d8fbc008
MD5 (chr12.bam) = c14ceeb6c41196331ef18572d8fbc008
```

- No region or gene specified will return only the BAM header.
- A request for an empty region will not return an error, just an empty BAM.

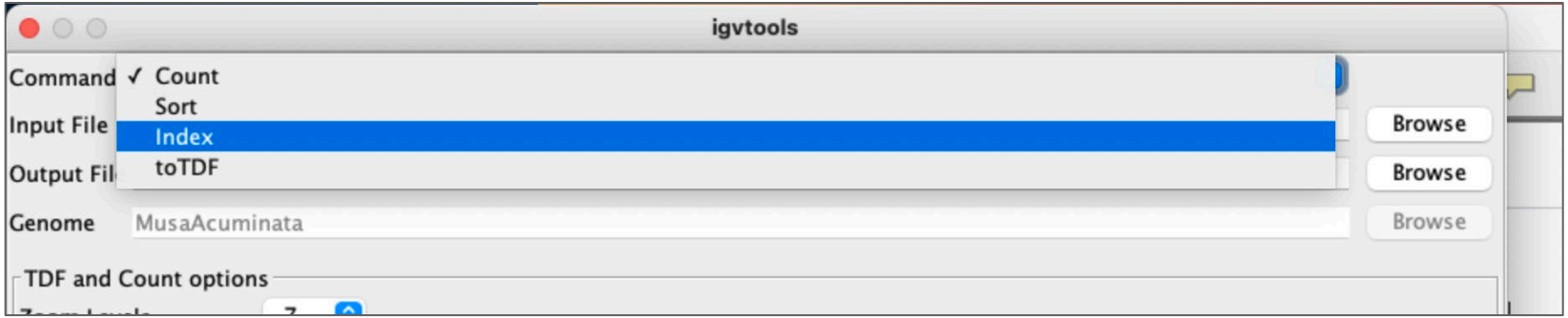
BAM Slicing Tips (2/2)

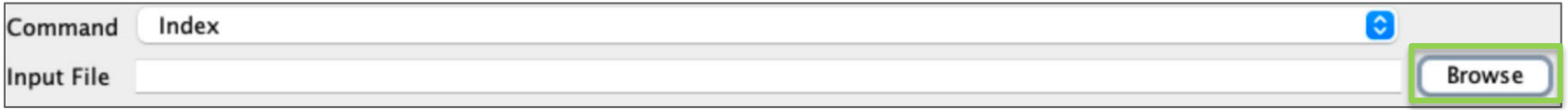
- Reads that overlap the region of interest will be provided in full.
- Reads that do not fall within the region of interest, but are paired with reads that do, will not be provided.
- No BAI files are created during BAM slicing, users must create their own:
  - Linux/Mac - `samtools index slice.bam`
  - Windows – IGV Demo

IGV Tools


IGV_2.3.92
File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
Run Batch Script...
Run igvtools...
Banana (Musa acum...
All
igvtools
Command
Count
Sort
Index
toTDF
Input File
Output Fil
Genome
MusaAcuminata
Browse
Browse
Browse
TDF and Count options
Gene

IGV Index


igvtools
Command
✓ Count
Sort
Index
toTDF
Input File
Output Fil
Genome
MusaAcuminata
Browse
Browse
Browse
TDF and Count options


Command
Index
Input File
Browse


Close
Run

BAM Slicing Troubleshooting

- Not sure if a BAM slice worked: `$ file slice.bam`

```
example_slice.bam:     Blocked GNU Zip Format (BGZF; gzip compatible), block length 9181
gene_and_region.bam:   Blocked GNU Zip Format (BGZF; gzip compatible), block length 9181
get_regions_slice.bam: Blocked GNU Zip Format (BGZF; gzip compatible), block length 9181
test_failure.bam:      JSON data
```

- Inspect BAM : `$ samtools view slice.bam`
- Find GENCODE symbols:
  - https://portal.gdc.cancer.gov/exploration
  - https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files

Questions from Participants

Useful Links

- GDC Portal – https://portal.gdc.cancer.gov
- GDC Documentation – https://docs.gdc.cancer.gov
  - BAM Slicing Docs – https://docs.gdc.cancer.gov/API/Users_Guide/BAM_Slicing/
- GDC Website – https://gdc.cancer.gov
- GDC Help Desk – support@nci-gdc.datacommons.io
- IGV – https://software.broadinstitute.org/software/igv/

U.S. Department of Health & Human Services

National Institutes of Health | National Cancer Institute

https://www.cancer.gov/

1-800-4-CANCER

Produced June 2023