



Universidad de Granada

**MÁSTER UNIVERSITARIO OFICIAL EN
CIENCIA DE DATOS E INGENIERÍA DE
COMPUTADORES**

**BIOLOGÍA COMPUTACIONAL CON BIG DATA-OMICS E
INGENIERÍA BIOMÉDICA**

Guión I: GDC Portal

Profesores:

Francisco Carrillo Pérez

Daniel Castillo Secilla

Luis Javier Herrera

Ignacio Rojas Ruiz

4 de febrero de 2020

Índice

| | |
|---|-----------|
| 1. Objetivos | 1 |
| 2. Introducción | 1 |
| 3. Pantalla de inicio y pestañas | 1 |
| 4. Seleccionar, explorar y descargar datos | 5 |
| 4.1. Selección y filtrado de los datos | 5 |
| 4.2. Guardado de los datos como un set | 10 |
| 4.3. Añadir datos al carrito y descarga | 10 |
| 5. Análisis de datos dentro de GDC Portal | 15 |

1. Objetivos

Durante el desarrollo de este guión se pretenden alcanzar los siguientes objetivos principales:

- Familiarizarse con la plataforma GDC.
- Descargar datos desde la plataforma.
- Familiarizarse con las herramientas de análisis que proporciona la propia plataforma.

2. Introducción

La iniciativa de GDC Data Portal (1) fue iniciada en Chicago y fue financiada por el gobierno de los Estados Unidos. El objetivo era poder dar acceso a los datos que se encontraban en distintos proyectos y homogeneizarlos. De esta forma no solo se presentarían en una plataforma visual y fácil de usar, si no que se podrían utilizar para análisis posteriores.

Se pueden encontrar diversos tipos de datos en GDC. Algunos requieren pedir permiso de acceso, indicando qué uso se va a hacer de los mismos, ya que pueden incorporar información personal sobre los pacientes. Sin embargo, otros son de libre acceso y serán en los que nos centraremos nosotros para trabajar.

3. Pantalla de inicio y pestañas

Cuanto accedemos a **GDC Portal** (<https://portal.gdc.cancer.gov/>), nos encontramos con la pantalla de inicio en la que se pueden ver distintas opciones. Una vista general se puede encontrar en la Figura 1.

En la parte superior podemos encontrar las distintas pestañas que dan acceso a distintas funcionalidades, las cuales se pueden observar en las Figuras 2 y 3. La funcionalidad de cada pestaña se detalla a continuación:

- **Home:** La pantalla de inicio donde nos encontramos actualmente.
- **Projects:** Acceso a los distintos proyectos que contiene la plataforma. Cada proyecto puede estar asociado a un único cáncer o varios, y puede contener un único o distintos tipos de datos.
- **Exploration:** Nos permite explorar todos los pacientes que contiene la plataforma.

- **Analysis:** Colección de herramientas de análisis que nos proporciona la propia plataforma.
- **Repository:** Visualización de todos los ficheros que contiene la plataforma, ya no por paciente.
- **Lupa:** Nos permite hacer búsquedas usando IDs de pacientes, tipos de cáncer, etc.
- **Manage Sets:** Nos permite administrar nuestros *sets* creados (qué es un *set* se explicará más adelante en este guión).
- **Login:** En el caso de haber pedido permiso para acceder a archivos que no sean públicos podremos loguearnos desde aquí.
- **Carrito de la compra:** En el carrito de la compra podremos ir añadiendo los archivos para luego descargarlos.

Ya dentro de la pantalla de inicio podemos acceder a algunas funcionalidades. A la izquierda de la pantalla, como se puede observar en la Figura 4, tenemos otro acceso a las pestañas superiores, una barra de búsqueda y algunas estadísticas de los datos que contiene GDC.

En el lado derecho, como se puede observar en la Figura 5, podemos observar una manera muy visual de seleccionar los datos de los tipos de cáncer que contiene la plataforma de GDC. Si pulsamos sobre la barra del cáncer que queremos, nos llevará a todos los pacientes que se contengan dentro de ese tipo de cáncer.

Guión I: GDC Portal

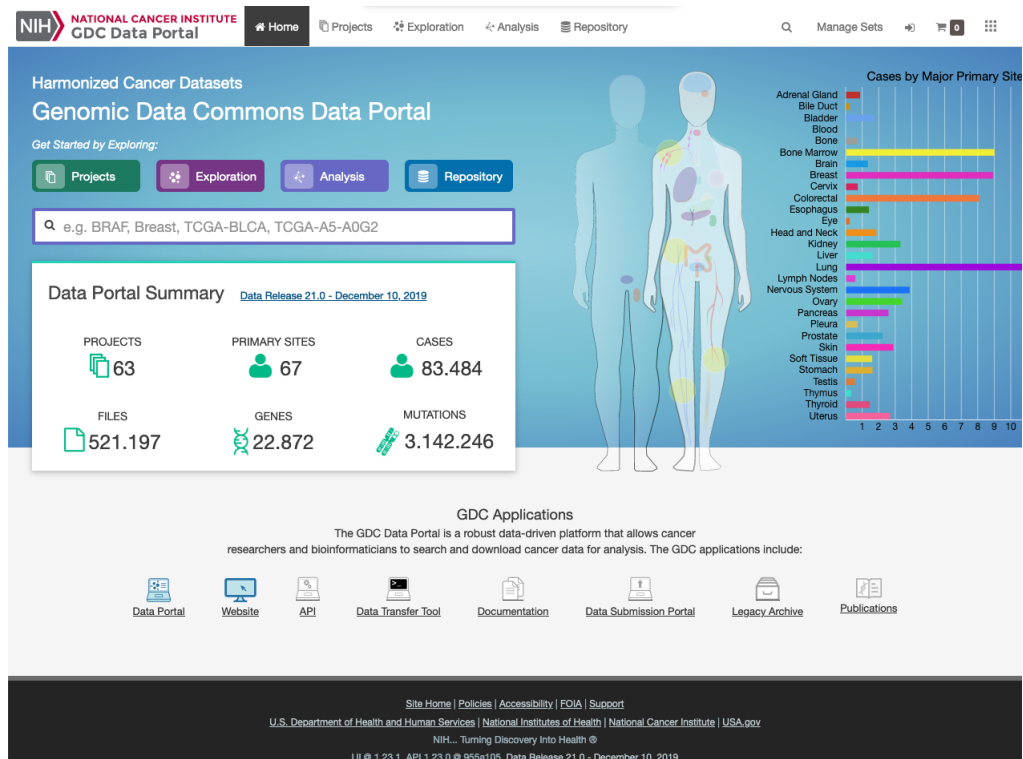


Figura 1: Pantalla de inicio del GDC Portal.



Figura 2: Primeras pestañas que nos encontramos en la parte superior.



Figura 3: Segundas pestañas que nos encontramos en la parte superior.

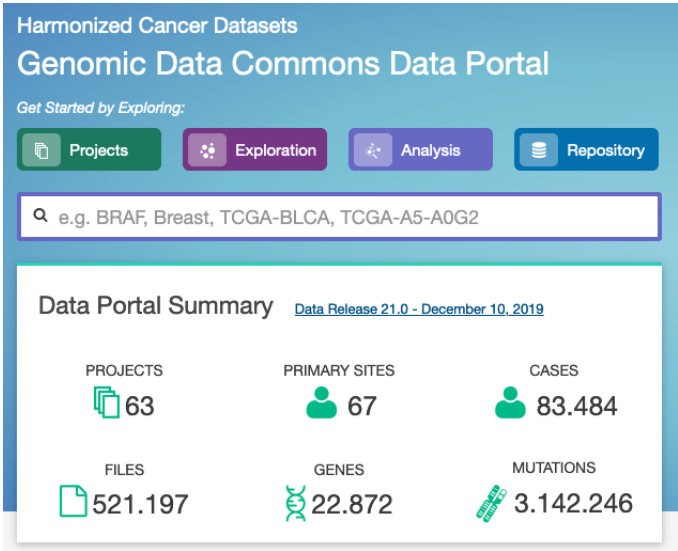


Figura 4: Barra de búsqueda y estadísticas de los datos que contiene GDC.

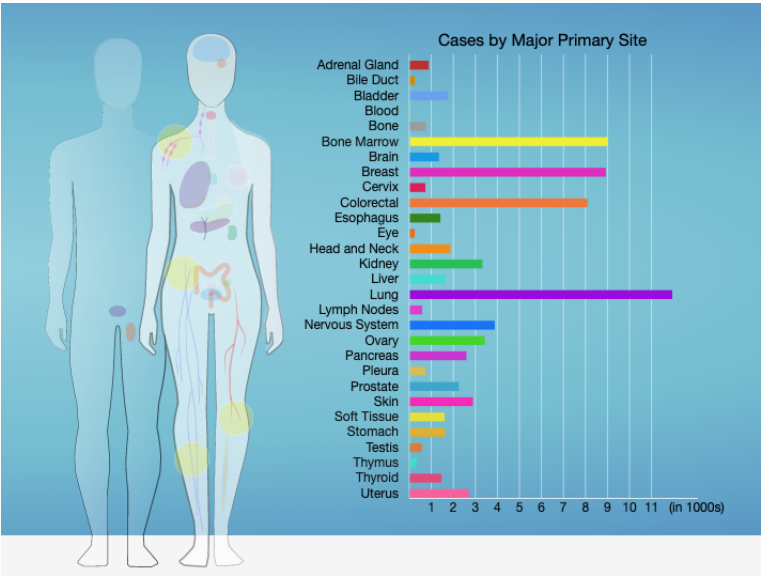


Figura 5: Visualización de todos los tipos de cáncer que contiene GDC y la cantidad de datos que se tienen de cada uno.

4. Seleccionar, explorar y descargar datos

Como ejemplo de como seleccionar, explorar y descargar datos de los pacientes vamos a usar el cáncer de pulmón. Para ello en la parte derecha de la pantalla de inicio (la que se muestra en la Figura 5) seleccionamos **Lung**.

Antes de comenzar con las distintas acciones que se pueden realizar sobre los pacientes y datos es conveniente explicar algunas reglas que siguen los datos de GDC. Los pacientes vienen identificados por un **Case ID**. Este es el identificador único del paciente (un ejemplo sería TCGA-18-3409). Cuando descarguemos los datos también descargaremos un fichero llamado **Sample Sheet** que contiene la información de los mismos. En el mismo se encuentra una columna importante, llamada **Sample ID**. Este columna contiene un string compuesto por Case ID y tres letras. El significado de las mismas es el siguiente:

- **01A**: El tejido es de tipo tumoral, es decir, la clase de ese dato es Primary Tumor.
- **01Z**: El tejido de la muestra es de metástasis.
- **11A**: El tejido de la muestra es de tejido sano, es decir, la clase del dato es Solid Tissue Normal.

De esta forma podemos saber rápidamente a qué clase pertenece ese dato de ese paciente (tumoral, metastásico o sano). Un ejemplo de un Sample Sheet se mostrará más adelante.

4.1. Selección y filtrado de los datos

Al haber seleccionado Lung anteriormente se nos llevará a una nueva pantalla que se puede observar en la Figura 6. Como se puede observar en la figura, tenemos una tabla donde para cada paciente figura la siguiente información: a qué proyecto pertenece, dónde se encuentra el tumor, el género, y los ficheros con los distintos datos (genómicos, imágenes, y otros) que se tiene del paciente. Podemos ordenarlos de mayor a menor o realizar distintas acciones con el menú de acciones superior, que se puede observar en la Figura 7.

A la izquierda de esta pantalla (figura 6) podemos observar el menú que nos permite filtrar todos los datos de la plataforma. En la parte superior del menú podemos observar como hay cuatro pestañas. Cada una de ellas nos proporcionará filtros sobre distintos aspectos. La pestaña que se encuentra abierta es **Cases**. En ella podemos observar distintos filtros que nos permitirán ajustar más los datos que queremos mostrar:

- **Primary Site:** Este primero ya se encuentra marcado y corresponde al cáncer que hemos seleccionado.
- **Program:** En este caso nos muestra los programas que contienen datos de este tipo de cáncer. Para la realización de este guión seleccionaremos TCGA (2) ya que es el programa que contiene datos tanto de RNA-Seq como de Tissue Slide.
- **Disease Type:** Nos permite seleccionar los distintos tipos de cáncer que existen dentro del cáncer global.
- **Experimental Strategy:** Permite seleccionar el tipo de datos que queremos.
- **Sample Type:** El tipo de muestras que queremos seleccionar. En nuestro caso seleccionaremos **Primary Tumor y Solid Tissue Normal**.

Una vez que ya hemos seleccionado todos estos datos en la columna, podemos observar que nos quedan 1090 casos (este número podría variar en el futuro). Ahora, si quisiésemos, podríamos filtrar por otro tipo de variables, por ejemplo clínicas. Si en la columna de la izquierda pulsamos sobre **Clinical** podemos observar como podríamos realizar otro tipo de filtrados por variables tales como demográficas, de diagnóstico, de tratamientos o de exposición.

Es de notar que se puede obtener aún información más extensa sobre los datos que ya hemos filtrado, tales como los genes más representativos dentro del filtrado, las mutaciones o las variantes que se producen. Para ello, en la parte superior de la tabla, podemos observar cuatro pestañas distintas:

- **Cases:** En ella encontramos la información de los pacientes descrita anteriormente en la tabla.
- **Genes:** En ella podemos encontrar distinta información de como se ven afectados los genes para la enfermedad que tengamos seleccionada. Se puede observar en la Figura 8. En la parte superior podemos observar dos gráficas distintas. La gráfica de la izquierda muestra los genes más frecuentemente mutados para esta enfermedad, donde el eje Y es el porcentaje de los casos que contiene GDC donde ese gen está mutado y el eje X el nombre del gen. La gráfica de la derecha representa el survival rate, donde para cada paciente se observa la supervivencia.
- **Mutations:** En ella podemos encontrar información sobre las distintas mutaciones que se han encontrado. Se puede observar en la Figura 9. En la parte superior podemos observar otra vez la gráfica de survival rate y en la parte inferior tenemos una tabla con la información de las mutaciones.

- **OncoGrid:** En ella podemos observar información sobre las mutaciones y los cambios en el copy number. En ella podemos observar de forma visual usando el color azul y rojo en los casos en los que se ha perdido una base y en otros casos en los que se ha ganado. Se puede observar en la Figura 10.

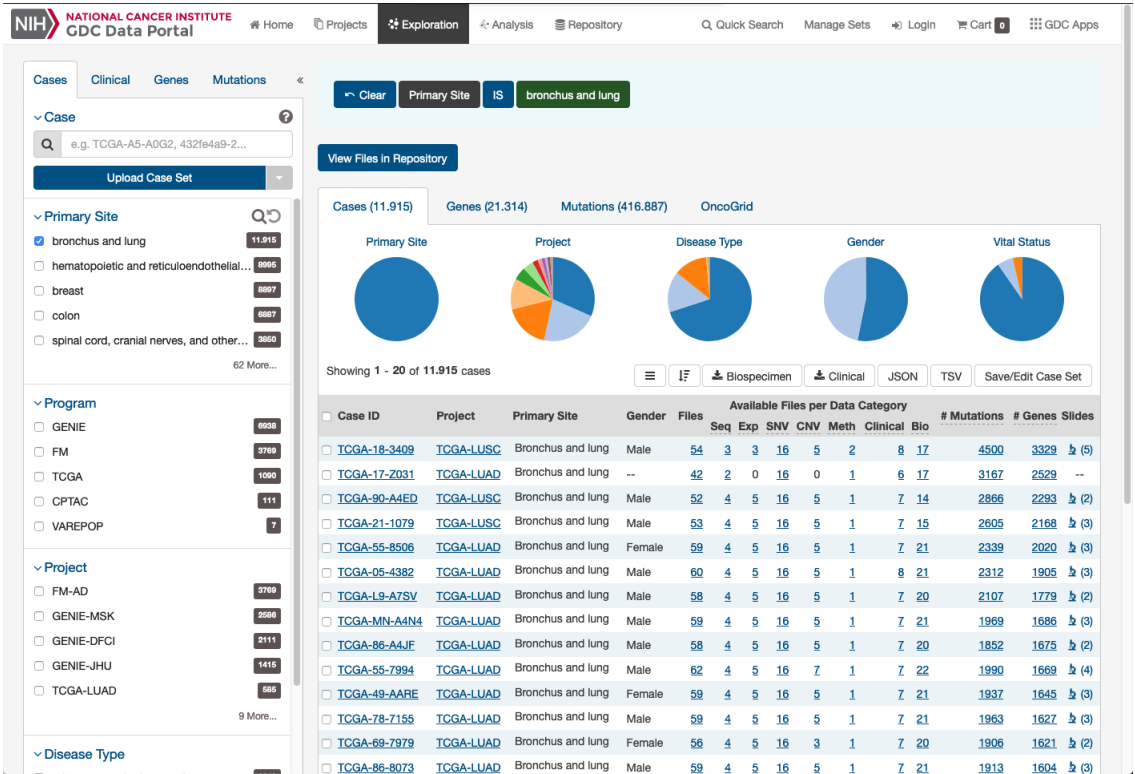


Figura 6: Todos los pacientes de Lung Cancer que se encuentran en GDC.

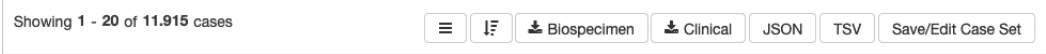


Figura 7: Acciones que se pueden realizar sobre la tabla de pacientes.

Guión I: GDC Portal

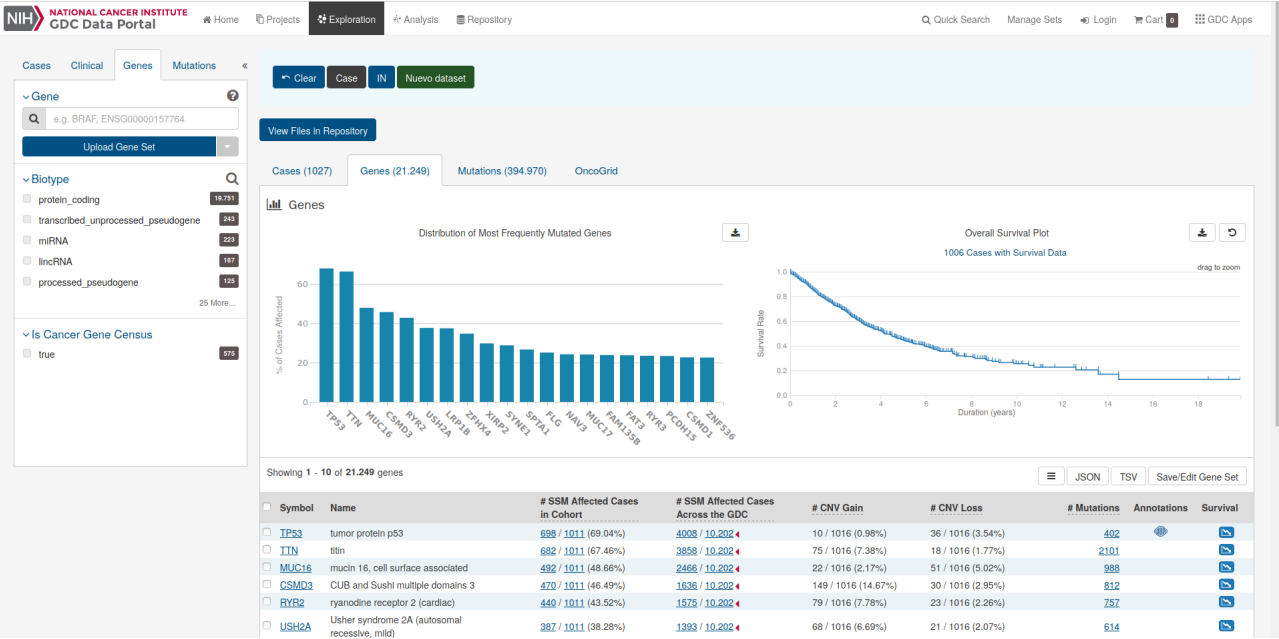


Figura 8: Pestaña de genes con los datos de los mismos, histograma de distribución y una survival plot.

Guión I: GDC Portal

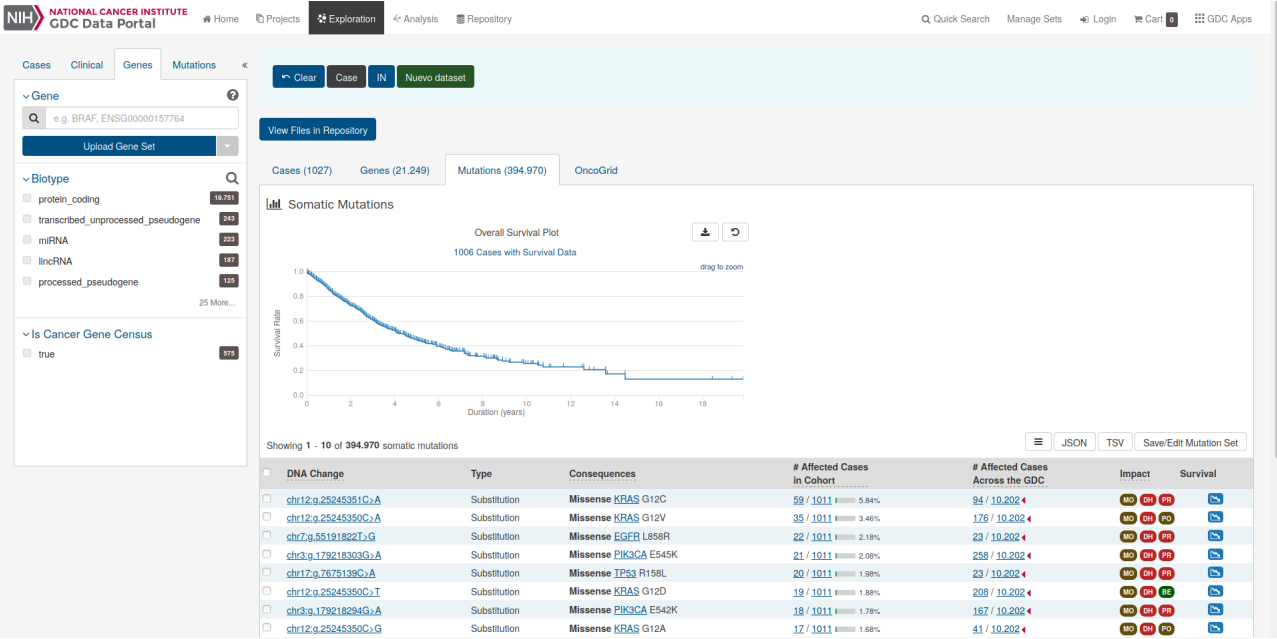


Figura 9: Pestaña de mutaciones con los datos de los mismos y gráfica de survival plot.

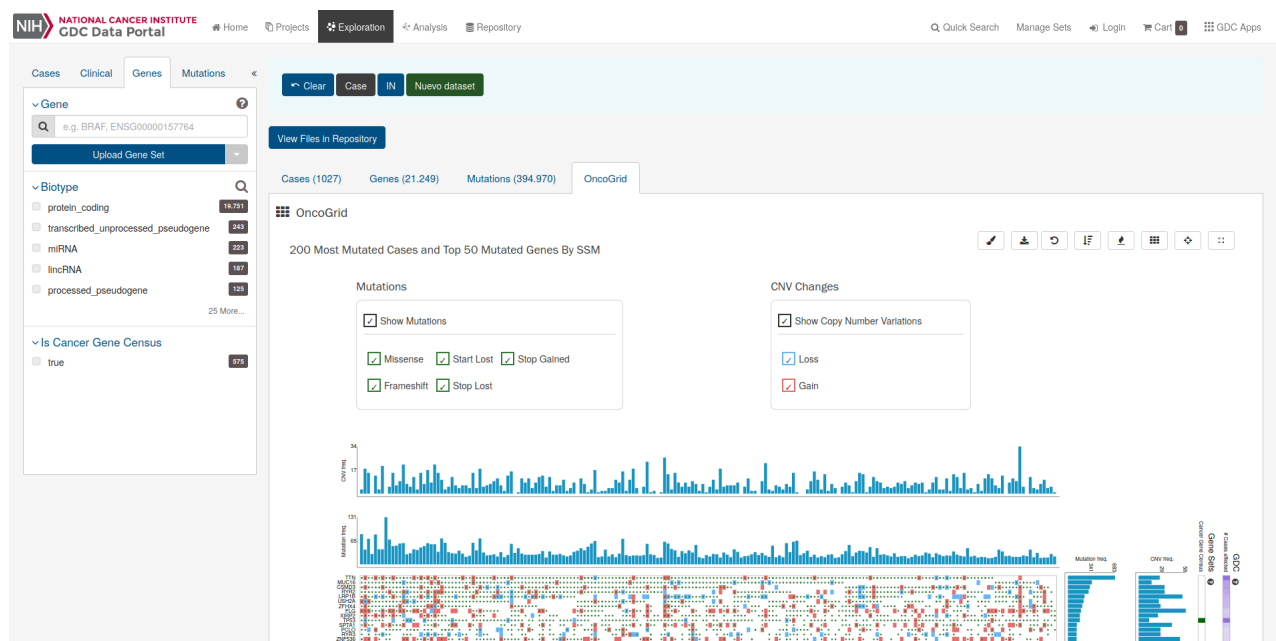


Figura 10: Pestaña de las variaciones que se producen en los genes, con el copy number. En ella observamos una gráfica donde se indica donde se ha ganado una base o se ha perdido.

4.2. Guardado de los datos como un set

Una vez que hemos realizado el filtrado y selección de los datos, guardar esta selección es de gran utilidad para no perderla y para poder hacer un análisis con las propias herramientas que nos proporciona GDC. Para ello, en el menú de la Figura 7, pulsamos sobre **Save/Edit Case Set** (parte derecha de la tabla) y a continuación en **Save as new case set**. Con esto, se nos abrirá un pop-up donde podremos seleccionar el nombre del set y el se visualizará el número de casos que tenemos seleccionados, como se puede observar en la Figura 11. Una vez rellenado el campo pulsaremos en save.

Una vez guardado podremos visualizarlo desde la barra de herramientas superior de la pantalla pulsando en **Manage Sets**, que se observa en la Figura 3.

4.3. Añadir datos al carrito y descarga

Una vez que accedemos a **Manage Sets** nos encontramos con los distintos sets que hayamos creado. Para poder ver los archivos del set que hemos creado pulsamos en

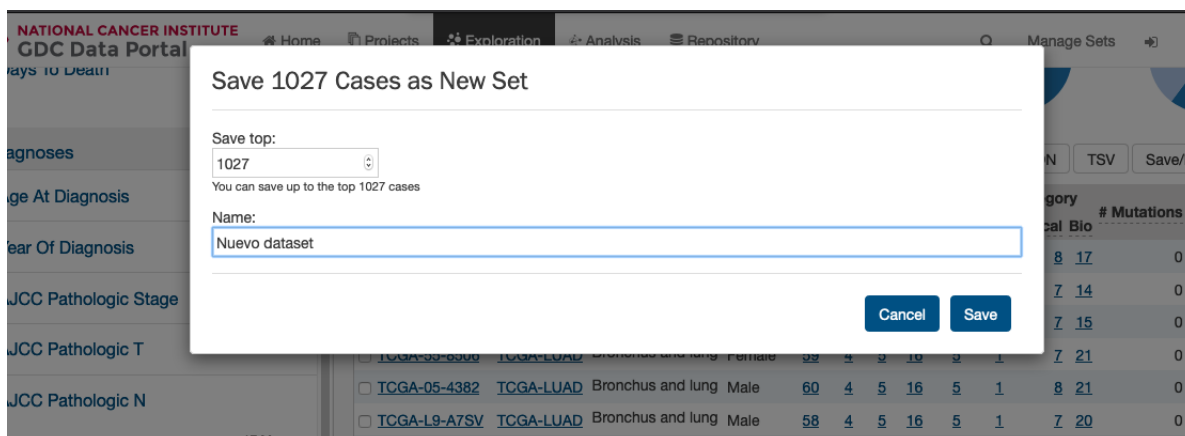


Figura 11: Pestaña para el guardado del nuevo set.

el último icono que aparece en la fila del set (**View files in repository**, icono a la derecha en la Figura 12).

Esto nos llevará a una pantalla muy similar a las que ya hemos visto, como se puede observar en la Figura 13. Podemos volver a filtrar por distintas variables, tanto de tipos de fichero como propias de los pacientes. En este caso nos vamos a centrar en la variable de **Experimental Strategy**, que se puede observar en la Figura 14. Para los siguientes guiones que realizaremos vamos a trabajar con dos tipos de datos distintos. Uno es RNA-Seq y los otros son Tissue Slides. Es por ello que vamos a filtrar dentro de nuestro set los datos que nos interesa descargar.

Comencemos con los datos de RNA-Seq. Para ello, seleccionamos RNA-Seq en la sección de **Experimental Strategy**. Una vez seleccionado se nos abrirá una nueva sección justo debajo que se llama **Workflow Type**. Los datos que nosotros utilizaremos para el análisis son los ficheros **count**, por ello seleccionaremos la opción **HTSeq-Counts**. Las opciones deberían quedar seleccionadas como se observa en la Figura 15.

Una vez que los tengamos, seleccionamos el icono del carrito que se encuentra justo a la derecha en la tabla y seleccionamos añadir todos (**Add all files to Cart**). Para ver el carrito seleccionamos en la parte superior de la pantalla el icono/pestaña del mismo nombre (**Cart**), lo que nos llevará a una nueva interfaz que se puede observar en la Figura 17. En este caso para poder descargar los datos simplemente pulsamos en el icono de **Download** y podemos descargar tanto el **Manifest**, que contiene la información del carrito (que puede ser usado también para la descarga), como los ficheros en sí. Además, es muy importante que descarguemos el **Sample**

Sheet pulsando en el icono correspondiente, ya que este contiene la información de a qué paciente corresponde cada fichero, así como la clase a la que pertenece el mismo.

Una vez ya descargados podemos limpiar el carrito y pasar a las imágenes. Para ello volvemos a nuestro set siguiendo los pasos anteriormente citados y en **Experimental Strategy** seleccionamos Tissue Slide, como se puede observar en la Figura 16. Volvemos a añadir todos los ficheros al carrito y volvemos a acceder al mismo. En este caso las imágenes son mucho más pesadas que los ficheros count. En esta situación es cuando la descarga haciendo uso del fichero **Manifest** y la herramienta **GDC Data Transfer Tool** es de gran utilidad. Justo en la página del carrito, en la esquina superior derecha tenemos un enlace con una explicación de cómo instalar y utilizar la herramienta para la descarga de los archivos. También debemos descargar el Sample Sheet para poder identificar a qué paciente corresponde cada imagen, aunque en el nombre del fichero si nos aparece en el caso de las imágenes.

Un ejemplo de como quedaría un Sample Sheet se puede observar en la Figura 18.

| <input type="checkbox"/> | Entity Type | Name | # Items | |
|--------------------------|-------------|---|----------------------|---|
| <input type="checkbox"/> | Cases | Nuevo dataset  | 1027 |   |

Figura 12: El último icono que aparece en la barra es el que debemos pulsar para ver los archivos del set. El anterior nos baja en fichero TSV la información de los pacientes del set.

Guión I: GDC Portal

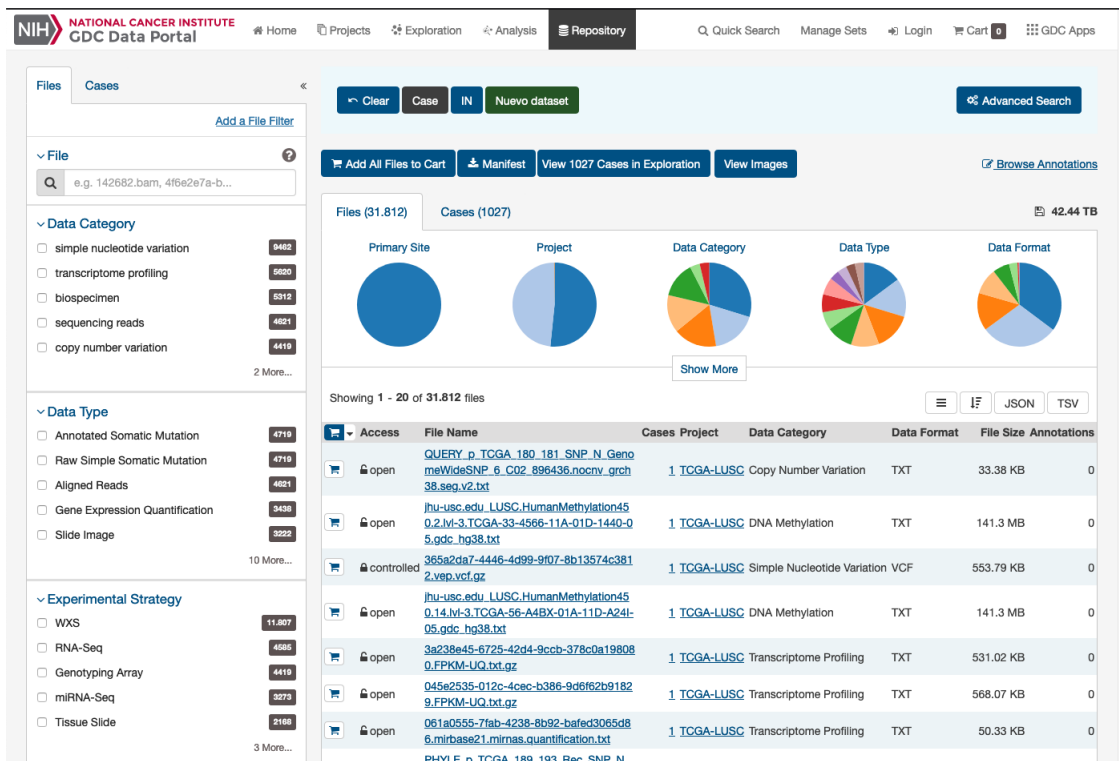


Figura 13: Pantalla de inicio de los datos que componen el set.

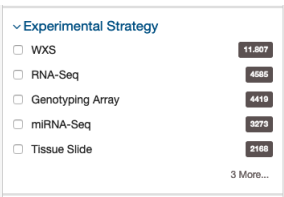


Figura 14: Distintos tipos de datos que podemos filtrar.

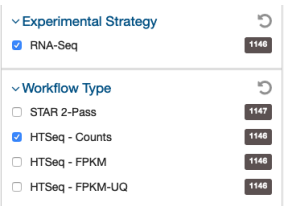


Figura 15: Para descargar los datos de RNA-Seq seleccionamos RNA-Seq en **Experimental Strategy** y luego en **Workflow Type** seleccionamos HTSeq-Counts.



Figura 16: Para descargar los datos que corresponden a las imágenes seleccionamos Tissue Slide en **Experimental Strategy**.

NIH
NATIONAL CANCER INSTITUTE
GDC Data Portal

HomeProjectsExplorationAnalysisRepository

Quick SearchManage SetsLoginCart 1146GDC Apps

FILES
1146

CASES
1017

FILE SIZE
290.93 MB

File Counts by Project

| Project | Cases (n=1017) | Files (n=1146) | File Size (1-290.93 MB) |
|-----------|----------------|----------------|-------------------------|
| TCGA-LUAD | 515 | 594 | 150.4 MB |
| TCGA-LUSC | 501 | 551 | 140.28 MB |
| TCGA-MESO | 1 | 1 | 253.31 KB |

File Counts by Authorization Level

| Level | Files (n=1146) | File Size (1-290.93 MB) |
|------------|----------------|-------------------------|
| Authorized | 1146 | 290.93 MB |

How to download files in my Cart?

Download Manifest:
Download a manifest for use with the [GDC Data Transfer Tool](#). The GDC Data Transfer Tool is recommended for transferring large volumes of data.

Download Cart:
Download Files in your Cart directly from the Web Browser.

BiospecimenClinicalSample SheetMetadataDownloadRemove From Cart

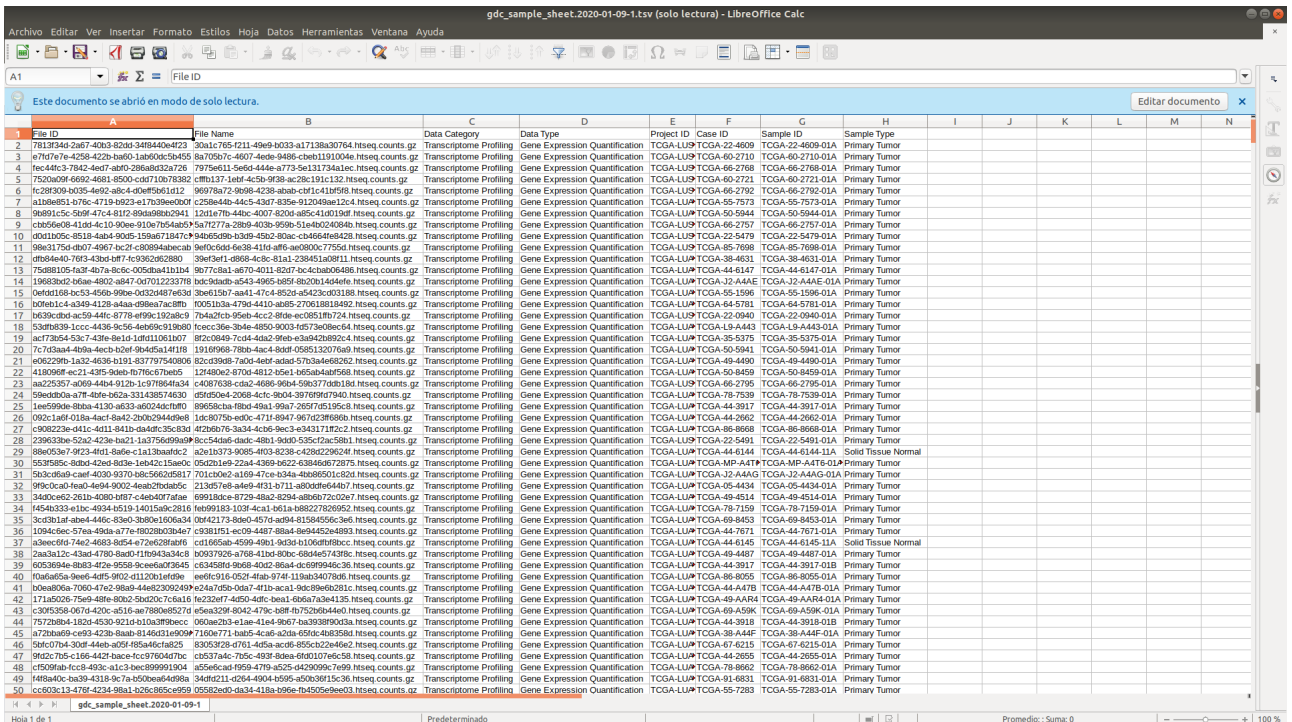
Cart Items

Showing 1 - 20 of 1146 files

| Access | File Name | Cases | Project | Data Category | Data Format | File Size | Annotations |
|--------|--|-------|-----------|-------------------------|-------------|-----------|-------------|
| open | 30a1c765-f211-49e9-b033-a17138a30764.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 257.03 KB | 0 |
| open | 8a705b7c-4607-4ede-9486-cbeb1191004e.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 255.8 KB | 1 |
| open | 7975e611-5e6d-444e-a773-5e131734a1ec.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 256.82 KB | 0 |
| open | cfff137-1ebf-4c5b-9f38-ac28c191c132.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 258.37 KB | 1 |
| open | 96978a72-9b98-4238-abab-cbf1c41bf5f8.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 254.41 KB | 0 |
| open | c258e44b-44c5-43d7-835e-912049ae12c4.htseq_counts.gz | 1 | TCGA-LUAD | Transcriptome Profiling | TXT | 252.63 KB | 1 |
| open | 12d1e7fb-44bc-4007-820d-a85c41d019df.htseq_counts.gz | 1 | TCGA-LUAD | Transcriptome Profiling | TXT | 249.25 KB | 0 |
| open | 5a7f277a-28b9-403b-959b-51e4b024084b.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 255.2 KB | 0 |
| open | 94b65d9b-b3d9-45b2-80ac-cb4664fe8428.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 252.76 KB | 0 |
| open | 9ef0c6dd-6e38-41fd-aff6-ae0800c7755d.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 258 KB | 0 |
| open | 39ef3ef1-d868-4c8c-81a1-238451a08f11.htseq_counts.gz | 1 | TCGA-LUAD | Transcriptome Profiling | TXT | 246.71 KB | 1 |
| open | 9b77c8a1-a670-4011-82d7-bc4cbab06486.htseq_counts.gz | 1 | TCGA-LUAD | Transcriptome Profiling | TXT | 255.06 KB | 2 |
| open | bdc9dad9-a543-4965-b85f-8b20b14d4efe.htseq_counts.gz | 1 | TCGA-LUAD | Transcriptome Profiling | TXT | 255.24 KB | 0 |
| open | 3be615b7-aa41-47c4-852d-a5423cd03188.htseq_counts.gz | 1 | TCGA-LUAD | Transcriptome Profiling | TXT | 250.58 KB | 0 |
| open | f0051b3a-479d-4410-ab85-270618818492.htseq_counts.gz | 1 | TCGA-LUAD | Transcriptome Profiling | TXT | 246.69 KB | 0 |
| open | 7b4a2fcb-95eb-4cc2-8fde-ac0851fb724.htseq_counts.gz | 1 | TCGA-LUSC | Transcriptome Profiling | TXT | 255.02 KB | 0 |

Figura 17: Pantalla del carrito. Como podemos observar se encuentran todos los ficheros que hemos añadido junto con información de los mismos.

Guión I: GDC Portal



| File ID | File Name | Data Category | Data Type | Project ID | Case ID | Sample ID | Sample Type |
|--|--|---------------|-------------------------|------------|--------------|------------------|---------------------|
| 7813f340-2ae7-40b3-82bd-3495440e4223 | 30a1c705-1211-49e9-b033-a17138a30764 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-22-4609 | TCGA-22-4609-01A | Primary Tumor |
| 67617e7a-4258-422b-4a6b-1a900a3c9459 | 8a70591c-4607-4d4b-9489-4a6b1191004e | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-60-2710 | TCGA-60-2710-01A | Primary Tumor |
| 4ec4f4c3-7842-4eaf-a8f0-286a8d3a2a76 | 7975e611-56d4-444e-a773-5e131734a1e1 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-66-2768 | TCGA-66-2768-01A | Primary Tumor |
| 7520a09f-6692-4681-8500-c0a710b76882 | c7fb137f-1e4f-4c5b-9f38-ac28c191c132 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-60-2721 | TCGA-60-2721-01A | Primary Tumor |
| 6c28300a-4025-4e92-a8c4-c0e9f6a1e112 | 9697b72c-8698-423b-a9ab-c9141d9f50 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-66-2792 | TCGA-66-2792-01A | Primary Tumor |
| a1a0e851-b76c-4719-b823-e13703ee0c0f | c25984d6-44c5-43d7-835e-912049ae12c4 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-55-7573 | TCGA-55-7573-01A | Primary Tumor |
| 9a8921c5-509f-417c-8122-89a4980e2941 | 12d1e77b-4d0c-4007-820d-a85c-41d019df | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-50-5944 | TCGA-50-5944-01A | Primary Tumor |
| c8b56e08-416d-4c10-90ee-9106-403a26c2880 | 5a17277a-2868-403b-959b-51e46024084b | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-66-2757 | TCGA-66-2757-01A | Primary Tumor |
| 0d0110c5-851b-4ab4-90c5-159af711a7c7 | 94d65e8b-404b-43b2-80ac-c4d6d46e4281 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-22-5479 | TCGA-22-5479-01A | Primary Tumor |
| 08a1375d-d807-4967-bc2f-c8089a4baccab | 9e0f5c6d-6e38-411d-a8fe-ae0800c-7755d1 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-85-7698 | TCGA-85-7698-01A | Primary Tumor |
| 12fb84e40-7813-43bd-b87f-c19362a62880 | 39a13e1f-b868-4c3c-81a1-238451a08f11 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-38-4631 | TCGA-38-4631-01A | Primary Tumor |
| 75d80105-423f-4a7a-8dc0-005b6a411b1a | 9b77c8a1-a670-4011-62d7-bc4cbab0a648b | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-6147 | TCGA-44-6147-01A | Primary Tumor |
| 196830c5-b6ae-4802-4847-00701223378f | bdc3da1b-4543-4965-b85f-8b20b14d4fe1 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-32-44AE | TCGA-32-44AE-01A | Primary Tumor |
| 0e0af169-bc33-456b-990e-0d32d8e763d3 | 3be01507-aad1-47c4-852d-a5423cd33188 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-55-1596 | TCGA-55-1596-01A | Primary Tumor |
| 1d0f61c4-a349-412b-84aa-d89ea7ac8ff1 | 70051d3a-479d-4410-a885-270618818492 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-64-5781 | TCGA-64-5781-01A | Primary Tumor |
| 453c8d4cf-ac59-4446-8773-ef9fc1302a03 | 7b4a2c09-996f-4c22-88ae-ec08518b724 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-22-0940 | TCGA-22-0940-01A | Primary Tumor |
| 53d7a833-1cc9-4436-9c5b-4e69e9191860 | 1ccc3c6-3b4e-4850-9003-fd573d0ec64 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-L9-AA43 | TCGA-L9-AA43-01A | Primary Tumor |
| 1a7f7854-53e7-43be-ba1e-10b110110107 | 802c0949-7c3d-40a2-99e6-e3a942b692c4 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-35-5375 | TCGA-35-5375-01A | Primary Tumor |
| 7c173d4a-40ba-4ecb-12d4-9a4da5141119 | 19109968-78b0-ba4c-40d8-09851320716f | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-50-5941 | TCGA-50-5941-01A | Primary Tumor |
| 1a0c2290-1a32-4636-8131-837797540806 | 82c33908-7a0d-4aef-adad-57b3a4e68262 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-49-4490 | TCGA-49-4490-01A | Primary Tumor |
| 1401908f-e231-4395-9a0b-b70c47b6e05 | 12f40a02-7f04-4812-05e1-b65ba4a7568 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-50-8459 | TCGA-50-8459-01A | Primary Tumor |
| 1a022537f-a099-404a-912b-1c97864f6a34 | c4037c3b-c24c-4698-99d4-59b3770d018f | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-66-2795 | TCGA-66-2795-01A | Primary Tumor |
| 19eac8b0a-a7ff-4dfe-b62a-331438574830 | d5f405d4-2068-4cfc-9b04-39769d7940 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-78-7539 | TCGA-78-7539-01A | Primary Tumor |
| 1ee1599a-8baa-4130-a633-a802a04cf0f0 | 89658c8a-bfbd-49d1-99a7-265f7d5195c8 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-3917 | TCGA-44-3917-01A | Primary Tumor |
| 2602c1a0-018a-4ac1-6842-2b0c2944d9e8 | 1dc607b0-e0dc-4711-8947-967c239f80b | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-2662 | TCGA-44-2662-01A | Primary Tumor |
| 1902222c-d41c-4411-941b-84dc4dc35c93 | 4229a07b-3a3a-4c2e-bc63-e343171f7c2 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-86-8668 | TCGA-86-8668-01A | Primary Tumor |
| 239633ba-52a2-432e-ba2f-1a3756d958a9 | 8cc54a65-dadc-48b1-9d05-535c2ac58b1 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-22-5491 | TCGA-22-5491-01A | Primary Tumor |
| 88e653e1-f923-4a1f-8a6e-c1a13baafdc2 | a2e16373-8085-4f03-8238-c428d29624f | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-6144 | TCGA-44-6144-11A | Solid Tissue Normal |
| 5538585-8d8d-42ed-82e8-1ed42c15a0e2 | 05c2b199-22e4-4369-b022-63846b672875 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-MP-A17F | TCGA-MP-A17F-01A | Solid Tissue Normal |
| 53c3a08a-canf-4030-9370-b6c56020e817 | 701c0a02-6169-474e-b34a-4b68501c82d1 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-32-AAAG | TCGA-32-AAAG-01A | Primary Tumor |
| 39f0c0a0-4e40-4e94-9002-4ea22b0da5fc | 213c57e8-a4e9-4f31-b711-a800df644b7 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-05-4434 | TCGA-05-4434-01A | Primary Tumor |
| 340dc0c2-2610-4080-0b87-c4e40a778aae | 699130ce-8729-48a2-8294-a8b6b72c0287 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-49-4514 | TCGA-49-4514-01A | Primary Tumor |
| 1543a333-e13c-4934-1619-1a015a9c7916 | 98f91163-102f-4c41-161a-b8927769662 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-78-7159 | TCGA-78-7159-01A | Primary Tumor |
| 3c0d81a1-2b4d-438c-3b0d01606a34 | 0642173-8ae0-457d-a9f4-815845563e8 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-69-8453 | TCGA-69-8453-01A | Primary Tumor |
| 109a6cee-57ea-49da-a77e-9028b03ab4e7 | 19361f5f-ec09-4a87-88a4-8e94452e4893 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-7671 | TCGA-44-7671-01A | Primary Tumor |
| 13e0e0f1-74e2-4693-8a5a-e726c28f8aef | c31665a8-4599-49b1-9d04-1106a0f8fbc | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-6145 | TCGA-44-6145-11A | Solid Tissue Normal |
| 2ba3a12c-43ad-78d0-8a07-119d943a34c8 | 005932f8-7f68-41b8-80bc-684e547439c | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-49-4487 | TCGA-49-4487-01A | Primary Tumor |
| 6053694e-8b83-472e-955b-9ceea0f3645 | c534589f-8068-40a2-89a4-6c699946c36 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-3917 | TCGA-44-3917-01B | Primary Tumor |
| 10a6a5e9-9eaf-a0f5-9002-41120a1e0fde | ee0cf916-057f-4697-974f-113a83a070b18 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-86-8625 | TCGA-86-8625-01A | Primary Tumor |
| 0ba0a00a-7060-4760-98a8-44e4c2002499 | c2a7a7d9-05a7-471b-acd1-9e39b6b2031c | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-AA7B | TCGA-44-AA7B-01A | Primary Tumor |
| 171a502d-75e0-480e-800e-5bdc20c7c6a16 | 1e232f27-4650-4dcf-bea1-6b6a7a3e4135 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-49-AA9A | TCGA-49-AA9A-01A | Primary Tumor |
| 13073594-067d-420c-451e-a67889b5527b | 95ea32f8-8042-479c-b8f7-7529b640460 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-69-A59K | TCGA-69-A59K-01A | Primary Tumor |
| 73728b4a-1d2d-433a-921d-e10a3b9bccc | 060a0a33-e1ae-41e4-9b67-ba39389f03d4 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-3918 | TCGA-44-3918-01B | Primary Tumor |
| 47a2b0d9-c093-423b-8ba8-8146c31e0909 | 710e7711-ba05-4c4b-a2da-65dc4b8358d | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-38-AA4F | TCGA-38-AA4F-01A | Primary Tumor |
| 46c07b7a-30af-4a6b-a05f-85844cfa825 | 83053f28-d761-4d5a-ac0e-85c5c22e40e2 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-67-4215 | TCGA-67-4215-01A | Primary Tumor |
| 16d27b5c-c186-4a29-ba6c-8c97804c7b6 | c937374b-7525-493f-849b-68037d7c5c58 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-44-2655 | TCGA-44-2655-01A | Primary Tumor |
| 48c50fab-f0c8-493c-a1c3-bce89991304 | a556e6ad-9599-4799-a525-4c29099c7e99 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-78-8662 | TCGA-78-8662-01A | Primary Tumor |
| 14f8a40c-ba3c-4318-9c7a-b50bea04098a | 34d012d1-0264-9044-b595-a50b36f15c36 | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-91-6831 | TCGA-91-6831-01A | Primary Tumor |
| cc0c3c13-478f-4234-98a1-b26c9c85e959 | 05502c9d-4d34-419a-b96e-b5405e6e03c | htseq counts | Transcriptome Profiling | TCGA-LU9 | TCGA-55-7283 | TCGA-55-7283-01A | Primary Tumor |

Figura 18: Ejemplo de Shample Sheet descargado con datos de RNA-Seq.

5. Análisis de datos dentro de GDC Portal

Una vez que ya hemos creado nuestro set(ver figura 12), podemos realizar un análisis de los datos dentro de la propia plataforma. Para ello en la barra superior pulsamos en **Analysis**, opción que se observa en la Figura 2.

Una vez hemos pulsado, podemos observar la pantalla de inicio con tres opciones distintas, como se puede observar en la Figura 19:

- **Set Operations:** Aplicar operaciones tales como intersección o unión entre distintos sets que hayamos creado.
- **Cohort Comparison:** En este caso comparáramos variables entre dos sets distintos, utilizando uno como control, y ver las diferencias con otro set.
- **Clinical Data Analysis:** Muestra varias gráficas y podemos realizar un análisis de un set.

Cada una de las opciones tiene una opción de **Demo** donde podemos navegar por las distintas funciones que hay dentro de cada opción. En nuestro caso, con nuestro set

creado, vamos a profundizar en las opciones que se nos presentan dentro de **Clinical Data Analysis**. Para ello pulsamos en su **Select**, y a continuación seleccionamos nuestro set y pulsamos en **Run**.

Esto nos llevará a una nueva interfaz que se puede observar en la Figura 20. A la izquierda podemos observar distintos filtros opciones que se pueden activar o desactivar dentro de las categorías de: Demographic, Diagnosis, Treatment y Exposure. Cada una que activemos añadirá una nueva gráfica en el panel de la derecha. Si desactivamos, desaparecerá del panel. Cada gráfica en sí puede descargarse los datos, la gráfica en sí o actualizar la misma, realizar zoom dentro de la gráfica y cambiar de tipo de gráfica dentro de una misma variable. Un ejemplo se puede observar en la Figura 21, donde pasamos de ver un histograma (que se observa en la Figura 20) a la gráfica del survival rate.

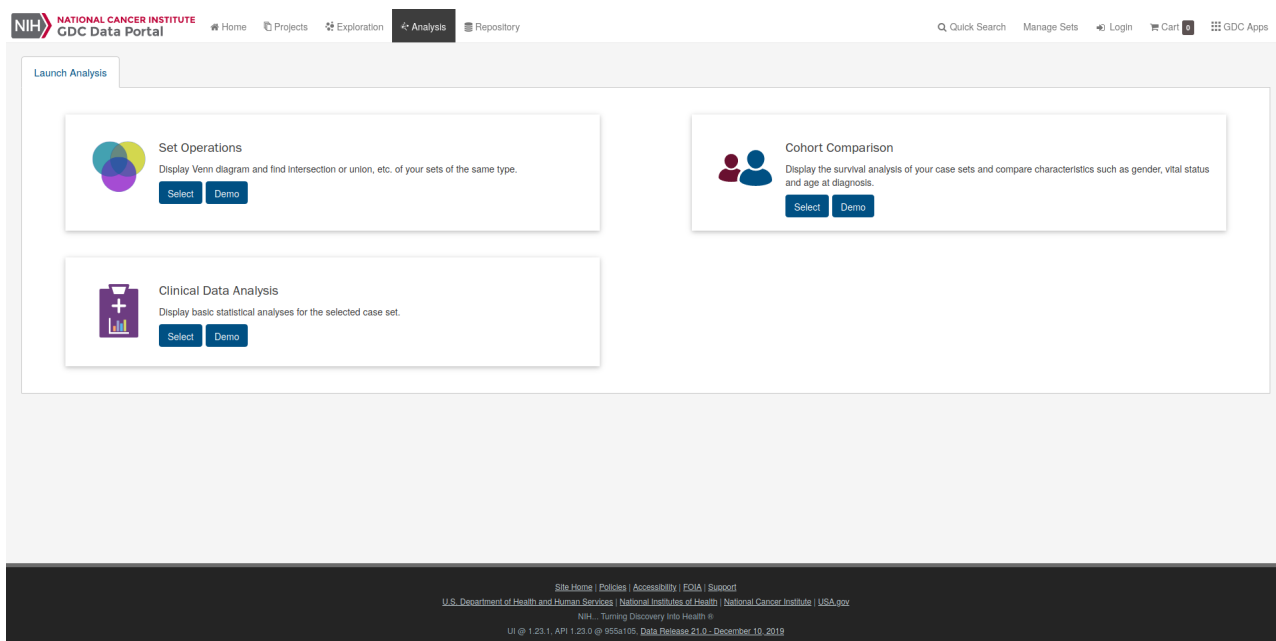


Figura 19: Pantalla de inicio de Analysis, con las distintas opciones.

Guión I: GDC Portal

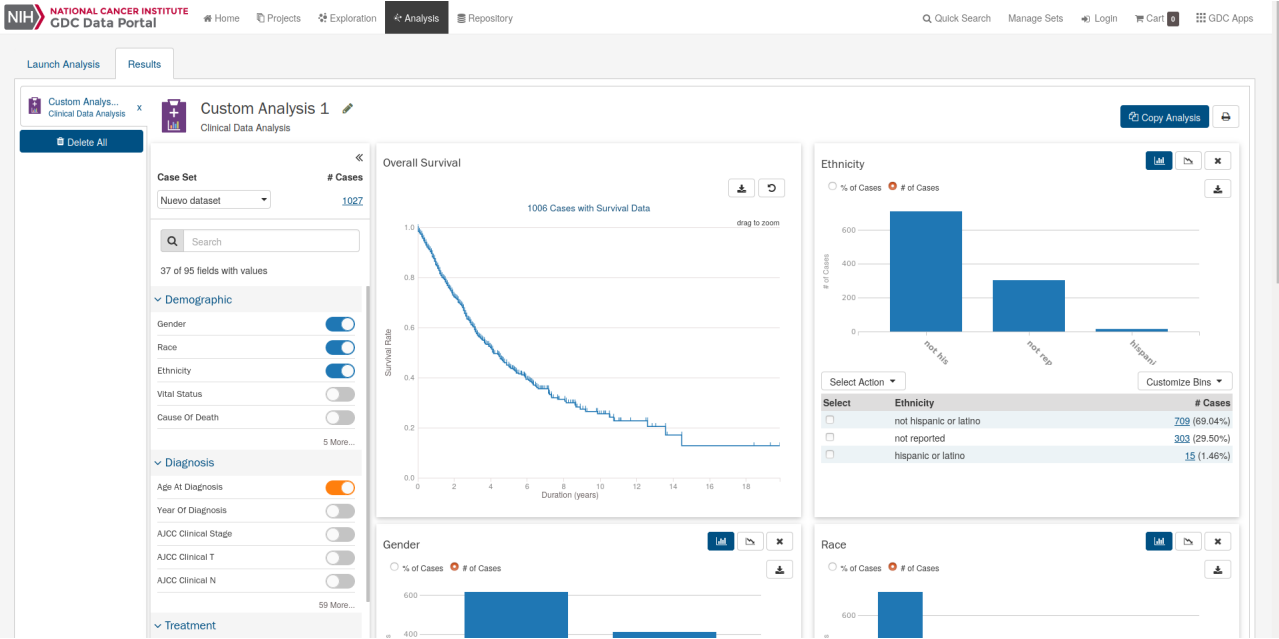


Figura 20: Pantalla análisis clínico donde podemos encontrar distintas opciones a la izquierda y las gráficas a la derecha.

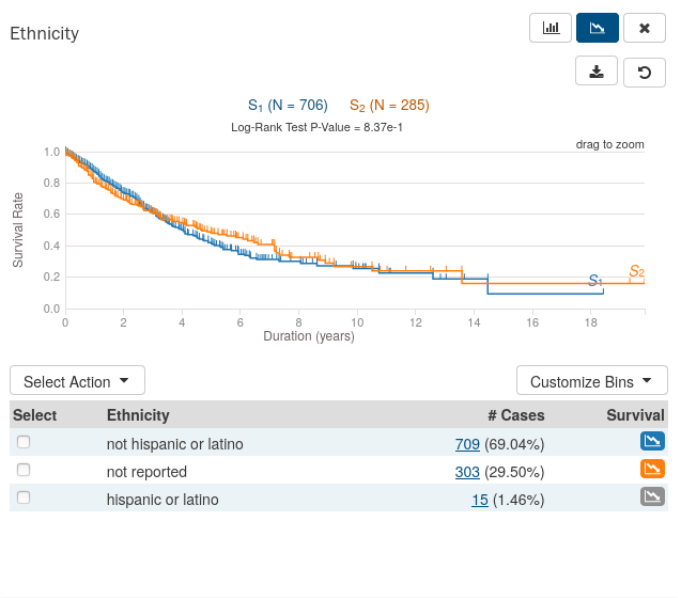


Figura 21: Ejemplo de gráfica donde podemos encontrar distintas opciones en la parte superior e inferior.

Referencias

- [1] Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., Staudt, L. M. (2016). Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12), 1109-1112.
- [2] TCGA Research Network. <https://www.cancer.gov/tcga>