

# PEC3\_Análisis de Datos Omicos

Roger\_Massaguer

2023-06-29

## Contents

<b>1 Abstract</b>	<b>2</b>
<b>2 Objetivos</b>	<b>2</b>
<b>3 Procedimiento</b>	<b>2</b>
3.1 Cargar los datos . . . . .	2
3.2 Alineamiento de los “reads” al genoma de referencia . . . . .	2
3.3 Check alignment . . . . .	3
3.4 Visualizar los archivos BAM con un visor de genoma integrado . . . . .	4
3.5 Selección de variables . . . . .	4
<b>4 Referencias</b>	<b>9</b>

## 1 Abstract

Los datos del estudio original se han obtenido a partir de los archivos fastq del link facilitado en el enunciado del problema [https://drive.google.com/drive/folders/1mbBfMRth-VGUqOd\\_1zgPj\\_EM5kde8P4L?usp=sharing](https://drive.google.com/drive/folders/1mbBfMRth-VGUqOd_1zgPj_EM5kde8P4L?usp=sharing).

Estos, han sido extraídos del portal web de la base de datos El Recurso Internacional de Muestras Genómicas (IGSR) que mantiene y comparte los recursos de variación genética humana creados por el Proyecto 1000 Genomas, concretamente de la muestra HG00128, vinculada a un panel de 99 británicos de Inglaterra y Escocia. [https://catalog.coriell.org/0/Sections/Search/Sample\\_Detail.aspx?Ref=HG00128&Product=CC](https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=HG00128&Product=CC)

Para lograr el objetivo del informe se han procesado estos datos (pasos que se explican en el apartado de métodos) con la plataforma Galaxy (servidor europeo, “galaxy.eu”) y se ha implementado con Rmarkdown. Las herramientas que se han utilizado dentro de la plataforma Galaxy son las siguientes: 1- Cargar los datos: Upload data 2- Alineamiento de los “reads” al genoma de referencia: Map with BWA 3- Check alignment: Samtool idxstats 4- Visualizar los archivos BAM con un visor de genoma integrado 5- Selección de variables: FreeBayes y SnpEff

## 2 Objetivos

El estudio se basa en el análisis de datos de lecturas cortas del exoma del cromosoma 22 de un único individuo humano. En el conjunto de datos hay un millón de lecturas de 76 pb, producidas en una Illumina GAIIx a partir de ADN Enriquecido en el exoma. Estos datos se generaron en el marco del proyecto Genomas 1000 genomas. Este informe muestra un análisis para buscar variantes minoritarias (SNVs/INDELS) en datos de exoma del proyecto de los 1000 genomas. El proyecto de los 1000 genomas pretendía encontrar variantes genéticas comunes con frecuencias de al menos el 1% en las poblaciones estudiadas. Se usará la plataforma Galaxy para el análisis de los datos extraídos de individuos declarados sanos. El objetivo de este informe es mostrar los resultados de un análisis de búsqueda de variantes minoritarias (SNVs/INDELS) de exoma para averiguar si existen variantes de determinados genes en individuos sanos.

## 3 Procedimiento

### 3.1 Cargar los datos

Como primer paso se han cargado los datos en la plataforma Galaxy con la herramienta “Upload data” y la opción “Choose local file” para seleccionar el archivo “.fastq” descargado del link aportado en el enunciado. Seleccionando los archivos: “exomeSample9\_2.fq” y “exomeSample9\_1.fq”. Como resultado se obtienen cada una de las secuencias en formato fastq. Dentro del archivo generado se modifica la base de datos por la que usará en el análisis (versión hg19 del genoma humano).

Una vez cargados los datos en el sistema, el primer paso que proceder siempre, en unos datos de ultrasecuenciación, sería hacer un control de calidad, que se puede procesar dentro del mismo programa de Galaxy con la opción FastQ-Quality control. Este paso, finalmente no se va a realizar dadas las instrucciones de la actividad.

### 3.2 Alineamiento de los “reads” al genoma de referencia

Acto seguido se prosigue a realizar el alineamiento de los reads con un genoma de referencia. Existen múltiples programas de alineamiento, por lo que se debe seleccionar aquel que funcione mejor con las características de los reads a trabajar. A parte también se deberá elegir un genoma de referencia, en este caso lo vamos a alinear con una de las últimas versiones del genoma humano: Human genome 19 (hg19). Se va a usar uno

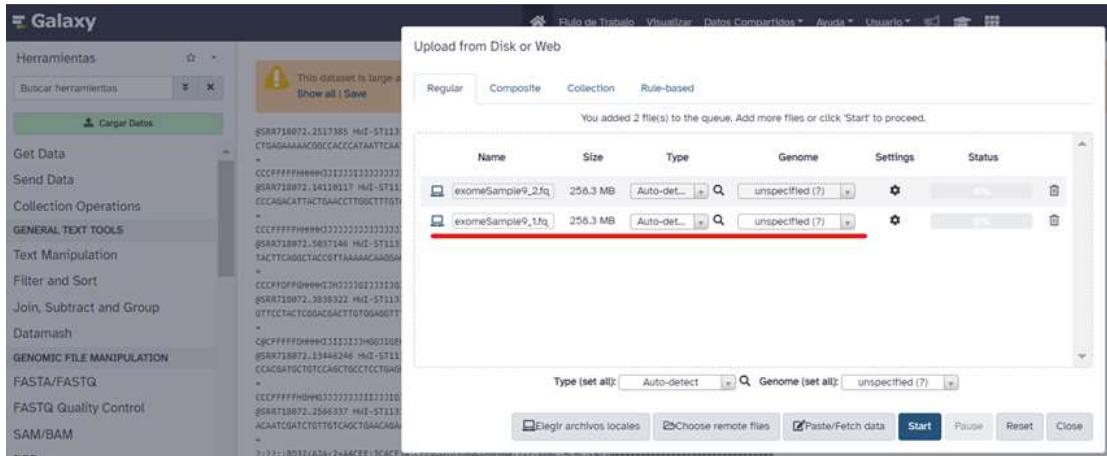


Figure 1: Carga de datos en Galaxy

de los alineadores que se ofrecen en el programa Galaxy: Burrows-Wheeler Aligner (BWA), que se conoce de los más eficientes hasta la fecha. Se ha usado “Map with BWA” en lugar de “BWA-MEM” ya que el primero es para secuencias cortas, como es el caso, mientras que el segundo es para secuencias largas. Una vez cargado el alineamiento, se obtiene una serie de secuencias alineadas: Muestra de las frecuencias de secuencias alineadas por cromosoma por BWA:

SRR718072.16050189	163	chr19	93019	23	101M	=	93149	231	TACTCCATCCTCCCTCGGCCGTGGCATCTGATATTGCTGATATTCAAGACACCAGGTGTCATGTTGCTCT
SRR718072.15624138	163	chr19	93085	29	101M	=	93171	167	TTCATGTTGCTCTGGCCCTGGTAATTCCATCTATCCATGCCCTGCCGTGACAGTGAGGAAGGAGCTGAGACAGCT
SRR718072.2617660	99	chr19	93120	60	101M	=	93147	128	ATCCATGCCCTCGCCGCTGTAACAGTGGAGGAAGGAGCTGCGCTGAACTTGCTGAACTTGCTCTGAGACACGCT
SRR718072.2617660	147	chr19	93147	60	101M	=	93126	-128	GAGGAAGGCCCTGCCTGAACATTGCTGAAACTGCTCTGAGACACGCTTGAGAGAGTCTCTGGATGGCCGTATTAT
SRR718072.16050189	63	chr19	93149	23	101M	=	93019	-231	GGAAGGCCCTGCCTGAACATTGCTGAAACTGCTCTGAGACACGCTTGAGAGAGTCTCTGGATGGCCGTATTATTC
SRR718072.9241347	99	chr19	93162	60	101M	=	93214	153	CCTGAACATTGCTGTAACACTGCTCTGAGACACGCTGTAAGAGTCTCTGGATGGCGTATTATCCCAATGAAGGTGGC
SRR718072.15624138	83	chr19	93171	37	101M	=	93085	-187	TGCTGTAACCTGCTCTGAGACACGCTGTAAGAGTCTCTGGATGGCGTATTATCCCAATGAAGGTGGCTGACATT
SRR718072.2920155	163	chr19	93177	60	101M	=	93289	213	AAACTGCTCTGAGACACGCTGTAAGAGTCTCTGGATGGCGTATTATCCCAATGAAGGTGGCTGACATTAGCCC
SRR718072.12136845	73	chr19	93198	23	101M	=	93198	0	GAAGAGTCTCTGGATGGCGTATTATCCCAATGAAGGTGGCTGACATTAGCCCCGGGGGGATGTCAAGACAGCTGTTACAGA
SRR718072.12136845	133	chr19	93198	6	*	=	93198	0	CGCGGGGGGGCGAGGCTGGGGCTGGGGGGCGGGGGGGCGAGGGGGAGGGAAAAGGGGGTAGGAGCAAGAG
SRR718072.9241347	147	chr19	93214	60	101M	=	93162	-153	GCGTATTATCCCAATGAAGGTGGCTGACATTAGCCCCGGGGGGATGTCAAGACAGCTGTTACAGACAGCTGTTACGTT
SRR718072.6073941	163	chr19	93231	60	101M	=	93314	164	GAAGGTGGCTGACATTAGCCCCGGGGTGGGATGTCAAGACAGCTGTTACGTTGTTGGGAGCCAGTCAGCAAAGTA
SRR718072.14771986	163	chr19	93257	60	101M	=	93310	154	GAGGTGGATGTCAAGACAGCTGTTACGTTGTTGGGAGCCAGTCAGCAAAGTAACGTGCTCTTATCTTGA
SRR718072.2926155	83	chr19	93289	60	101M	=	93177	-213	TTGTTGGGGAGGCCAGTCAGCAAAGTAACGTGCTCTTATCTTGAATGTGAACATTGTTCATCCACCTCCCTCAT
SRR718072.14771986	83	chr19	93310	60	101M	=	93257	-154	AAGTAACGTGCTCTTATCTTGAATGTGAACATTGTTCATCCACCTCCCTCATGGCATGGCACCCCTGAAAATGGCAGC
SRR718072.6073941	83	chr19	93314	60	101M	=	93231	-164	AACGTGTTCTATCTTGAATGTGAACATTGTTCATCCACCTCCCTCATGGCATGGCACCCCTGAAAATGGCAGC

Figure 2: Freq. secuencias

### 3.3 Check alignment

El siguiente paso consiste en revisar la calidad del alineamiento. Normalmente hay una serie de herramientas que son las BAM tool, que nos va a permitir sacar algunas estadísticas para ver y decidir si el resultados del alineamiento es satisfactorio. Este control lo haremos con la herramienta de Galaxy “Samtool idxstats”. Esta función nos va a generar un informe del alineamiento obteniendo una tabla en la que se observa cuantas secuencias aparecen asociadas a cada una de los cromosomas. En este estudio vemos que en el cromosoma

1 aparecen 216467 secuencias (Fig3.).

Col 1	Col 2	Col 3	Col4
chr17	81195216	162679	444
chr17_gi000203_random	37498	0	0
chr17_gi000204_random	81310	44	0
chr17_gi000205_random	174568	79	0
chr17_gi000206_random	41691	4	0
chr18	78077248	28968	102
chr18_gi000207_random	4262	0	0
chr19	59128963	83567	311
chr19_gi000208_random	92689	8	0
chr19_gi000209_random	159169	353	0
chr1	249250621	216467	967
chr1_gi000191_random	106433	96	1
chr1_gi000192_random	547496	623	3
chr20	63625520	49775	151

Figure 3: Resultado de la función Samtoolidxstats

Column Description
1 Reference sequence identifier
2 Reference sequence length
3 Number of mapped reads
4 Number of placed but unmapped reads (typically unmapped partners of mapped reads)

Figure 4: Referencias columnas Figura 3

### 3.4 Visualizar los archivos BAM con un visor de genoma integrado

Además de obtener las estadísticas resultantes del paso anterior, podemos visualizar el alineamiento utilizando un programa de visualización genómica como el Integrated Genome Viewer (IGV).

Integrative Genomics Viewer (IGV) es una herramienta interactiva de alto rendimiento y fácil uso para la exploración visual de datos genómicos. Soporta la integración flexible de todos los tipos comunes de datos genómicos y metadatos, generados por el investigador o disponibles públicamente, cargados desde fuentes locales o en la nube.

Una vez obtenido el alineamiento se procede a realizar un “Pileup”, para ver si las secuencias se distribuyen de forma homogénea entre las distintas regiones. Con esta función se podrá obtener un resumen del numero de secuencias por cada una de las regiones.

Obtenida la tabla con los parámetros entrados por defecto, se puede observar que en una serie de posiciones que resultan indicativo de un buen resultado.

### 3.5 Selección de variables

Para la selección de variables es necesario usar un variant caller, concretamente “FreeBayes”. Esta función, va a generar una tabla que indica en un conjunto de posiciones, cual es la secuencia de referencia y cual es

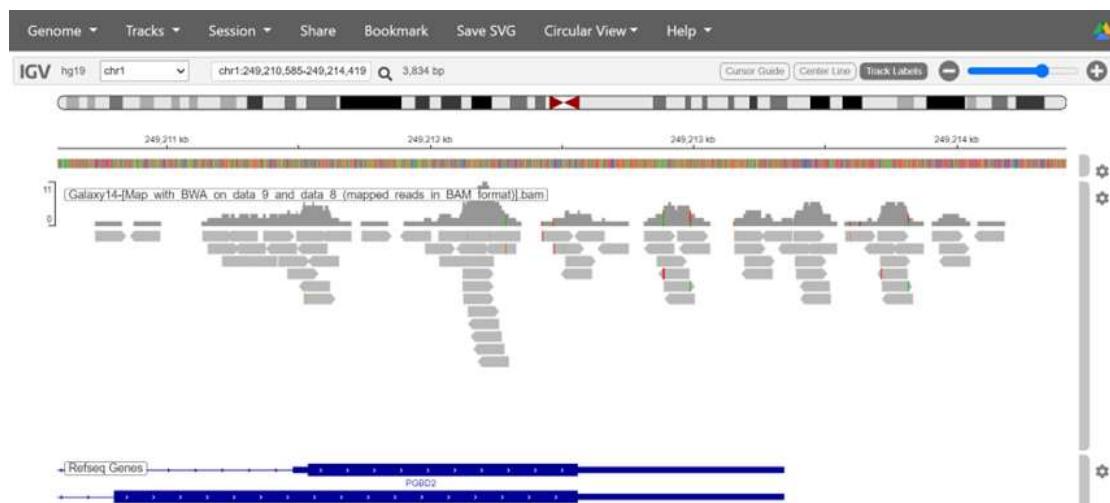


Figure 5: Visualización archivo BAM con IGV\_1

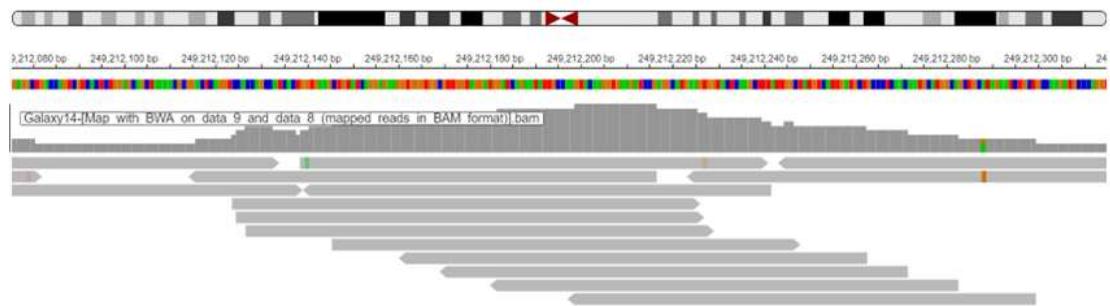


Figure 6: Visualización archivo BAM con IGV\_2

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
chr10	93019	t	T	30	0	23	1	^8,	:
chr10	93020	a	N	0	0	0	1	.	+
chr10	93021	c	C	30	0	23	1	.	1
chr10	93022	t	T	30	0	23	1	.	=
chr10	93023	c	C	30	0	23	1	.	B
chr10	93024	c	C	30	0	23	1	.	B
chr10	93025	t	T	30	0	23	1	.	D
chr10	93026	c	C	30	0	23	1	.	D
chr10	93027	a	A	30	0	23	1	.	D
chr10	93028	t	T	30	0	23	1	.	D
chr10	93029	c	C	30	0	23	1	.	D
chr10	93030	c	C	30	0	23	1	.	D
chr10	93031	t	T	30	0	23	1	.	D
chr10	93032	c	C	30	0	23	1	.	I
chr10	93033	c	C	30	0	23	1	.	I
chr10	93034	t	T	30	0	23	1	.	E
chr10	93035	c	C	30	0	23	1	.	B
chr10	93036	c	C	30	0	23	1	.	>
chr10	93037	t	T	30	0	23	1	.	F
chr10	93038	c	C	30	0	23	1	.	E
chr10	93039	c	C	30	0	23	1	.	B
chr10	93040	t	T	30	0	23	1	.	F
chr10	93041	c	C	30	0	23	1	.	C
chr10	93042	G	G	30	0	23	1	.	8

Figure 7: Pileup de las secuencias del estudio

las variantes que aparecen en cada una de estas posiciones. Estas ultimas van a tener un valor de calidad, permitiendo identificar los scores de aquellas posiciones donde aparecen variantes.

#CHROM	POS ID	REF	ALT	QUAL	FILTER	INFO
chr1	14677 .	G	A	0.0530642 .	AB=0.49	
chr1	14997 .	A	G	66.7439 .	AB=0.75	
chr1	14930 .	AAGG	GAGA,GAGG	65.0854 .	AB=0.26	
chr1	14976 .	G	A	13.0814 .	AB=0.33	
chr1	17452 .	C	T	19.8063 .	AB=0.66	
chr1	762273 .	G	A	63.5283 .	AB=0.46	
chr1	762589 .	GGCC	CGCG	46.5321 .	AB=0.46	
chr1	762601 .	T	C	62.7451 .	AB=0.46	
chr1	762632 .	T	A	49.131 .	AB=0.46	
chr1	857728 .	T	G	59.1574 .	AB=0.46	
chr1	866511 .	CCCCCTCCCTCCCTCCCA	CCCCCTCCCTCCCTCCCTCCCA	43.7256 .	AB=0.46	
chr1	876499 .	A	G	26.4187 .	AB=0.46	
chr1	881627 .	G	A	59.1574 .	AB=0.46	
chr1	883625 .	A	G	82.0049 .	AB=0.46	
chr1	887560 .	A	C	46.5181 .	AB=0.46	
chr1	888639 .	T	C	88.3889 .	AB=0.46	
chr1	888659 .	T	C	160.241 .	AB=0.46	
chr1	892745 .	G	A	96.1753 .	AB=0.46	

Figure 8: Resultado FreeBayes

Para finalizar el estudio, se van a anotar las variantes que se han detectado, por tal de saber cuál es el efecto de estas posibles variantes. Ejecutando el programa SnpEff sobre el resultados obtenido anteriormente del FreeBayes.

Una vez obtenidas todas las variantes candidatas se han anotado en función de sus ubicaciones genómicas y se han predicho los efectos decodificación con la herramienta “SnpEff eff” . Las ubicaciones describen si la variante se encuentra en un intrón, un exón, upstream/downstream, una región de splicing o una región intergénica mientras que los efectos denotan si la variante causa sobre la proteína codificada un cambio de aminoácido sinónimo o no sinónimo, ganancia o pérdida de codón de inicio, ganancia o pérdida de codón de parada o cambio de marco de lectura.

El resultado de esta acción va a ser por un lado un output de un archivo VCF, donde se generar las variantes con más información y un informe con información de cada una de la variantes que se contemplan.

Tras emplear la herramienta “SnpEff eff” se obtienen un archivo con los datos crudos y además un archivo html con diferentes tablas que muestran información sobre las variantes. A continuación se muestra la tabla resumen y el tipo de variantes de forma representativa:

De esta tabla podemos concluir que se obtienen 31.460 SNPs, de los cuales 606 son inserciones, 900 delecciones, no hay variaciones estructurales (Fig.11)

Además nos proporcionará distintas informaciones, des de tipos distintos de cambios que generarán ya sea por el impacto como por el efecto de traducción a proteína.

Col	Field	Description
1	CHROM	Chromosome name
2	POS	1-based position. For an indel, this is the position preceding the indel.
3	ID	Variant identifier (optional). Usually the dbSNP rsID.
4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
5	ALT	Comma delimited list of alternative sequence(s) seen in our reads.
6	QUAL	Phred-scaled probability of all samples being homozygous reference.
7	FILTER	Semicolon delimited list of filters that the variant fails to pass.
8	INFO	Semicolon delimited list of variant information.
9	FORMAT	Colon delimited list of the format of individual genotypes in the following fields.
10+	Sample(s)	Individual genotype information defined by FORMAT.

Figure 9: Referencias columnas Figura 8

Summary	
Genome	hg19
Date	2023-06-28 04:53
SnpEff version	SnpEff 4.3t (build 2017-11-24 10:18), by Pablo Cingolani
Command line arguments	SnpEff -i vcf -o vcf -stats /corral4/main/jobs/051/241/51241358/outputs/galaxy_dataset_6f5212cc-f119-44d0-9770-4b2862a9f8cd.dat hg19 /corral4/main/objects/c/e/b/dataset_cebd718b-a563-4b42-a78b-3a2d9441846a.dat
Warnings	533
Errors	0
Number of lines (input file)	33,592
Number of variants (before filter)	33,668
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	33,666
Number of known variants (i.e. non-empty ID)	0 ( 0%)
Number of multi-allelic VCF entries (i.e. more than two alleles)	56
Number of effects	91,940
Genome total length	3,137,161,265
Genome effective length	3,130,561,488
Variant rate	1 variant every 92,988 bases

Figure 10: Informe sobre las variantes (SnpEff eff)

**Number variants by type**

Type	Total
SNP	31,460
MNP	626
INS	606
DEL	900
MIXED	74
INV	0
DUP	0
BND	0
INTERVAL	0
<b>Total</b>	<b>33,666</b>

Figure 11: Tabla de resultados 1)

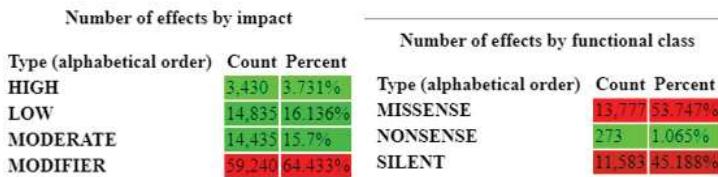


Figure 12: Tabla de resultados 2)

## 4 Referencias

5889a7897b93c9e3d7c687e5736c96cdee468058 @ www.genomamayor.com. (s.d.). [https://www.genomamayor.com/analisis-de-exomas/#:~:text=El servicio de análisis de,a nivel de estudios clínicos.](https://www.genomamayor.com/analisis-de-exomas/#:~:text=El%20servicio%20de%20an%C3%A1lisis%20de,a%20nivel%20de%20estudios%20cl%C3%ADnicos)

SIAF, importancia de los modulos del. (2011). Watch @ Www.Youtube.Com. En The True South Through My Eyes - HK Edgerton.

<https://www.youtube.com/watch?v=mFlITzqRBWY> 9498bc1ea2001c1b4e8a52bc20b9073aa10b99cc @ www.kolabtree.com. (s.d.). <https://www.kolabtree.com/blog/es/guia-paso-por-paso-del-analisis-de-los-datos-del-na/#Alignment>

Index @ Bio-Bwa.Sourceforge.Net. (s.d.). <https://bio-bwa.sourceforge.net/> 1d2e393c9fd56f154ffde8badffc5b261f3cd1c6 @ www.institutobernabeu.com. (s.d.). <https://www.institutobernabeu.com/es/foro/que-es-el-analisis-genetico-de-exoma-tipos-de-analisis-geneticos-de-exomas/>

1191eb6b7e007af8f3b232257d4b351736b72d9f @ software.broadinstitute.org. (s.d.). <https://software.broadinstitute.org/software/igv/> Sánchez, Á., Gonzalo, R., & Ferrer, M. (s.d.). Introduction to Variant Analysis.

f8a11901219e36f391cc54b90730db0bdbd426a2 @ materials.campus.uoc.edu. (s.d.). [https://materials.campus.uoc.edu/cdocent/PID\\_00292294/](https://materials.campus.uoc.edu/cdocent/PID_00292294/) c431ae39cfbd72d13ae531e10469e7921bc7e5e5 @ igv.org. (s.d.). <https://igv.org/app/>

preview @ usegalaxy.org. (s.d.). <https://usegalaxy.org/datasets/f9cad7b01a4721353c0c87580dc4d8e6/> preview