

Preprocesamiento de los datos y reportes de calidad.

Evelin González F.
evefeliu@gmail.com

Organización de las clases

- **Clase 3:** Preprocesamiento de los datos y reportes de calidad.
(Tarea N2)
- **Clase 4:** Llamado de variantes y anotación de variantes patogénicas (Tarea N3)
- **Clase 5:** Introducción a R y visualización de los datos (parte 1)
(Sin Tarea)
- **Clase 6:** Introducción a R y visualización de los datos (parte 2).
(Taller N1)

Replicar - Análisis de datos pacientes con cáncer mama

The screenshot shows a preprint server interface. At the top, logos for medRxiv, CSHL, BMJ, and Yale are displayed, along with navigation links for HOME, SUBMIT, FAQ, BLOG, ALERTS / RSS, RESOURCES, and ABOUT. A search bar and an advanced search link are also present.

The main content area displays a preprint titled "A workflow for clinical profiling of BRCA genes in Chilean breast cancer patients via targeted sequencing". The authors listed are Evelin González, Rodrigo Moreno Salinas, Manuel Muñoz, Soledad Lantadilla Herrera, Mylene Cabrera Morales, Pastor Julian, Waleska Ebner Durrels, Gonzalo Vigueras Stari, Javier Anabalón Ramos, Juan Francisco Miquel, Lilian Jara, Carol Moraga, Alex Di Genova. The DOI provided is <https://doi.org/10.1101/2024.09.25.24314295v1>.

The preprint has been posted on September 27, 2024. Below the post details, there are download options (PDF, Print/Save Options, Author Declarations, Supplementary Material, Data/Code), sharing options (Email, Share, Citation Tools, Get QR code), and social media sharing buttons (Post, Me gusta).

Below the main article, there is a section for COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv, with a subject area in Genetic and Genomic Medicine. A sidebar on the right lists subject areas such as Addiction Medicine, Allergy and Immunology, Anesthesia, Cardiovascular Medicine, and Dentistry and Oral Medicine.

doi:

<https://www.medrxiv.org/content/10.1101/2024.09.25.24314295v1.full-text>

¿Qué es NGS? (Secuenciación de Nueva Generación)

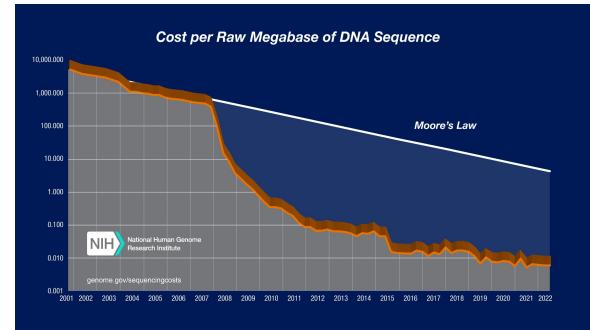
- Tecnología de secuenciación masiva y paralela.
- Permite la secuenciación de millones de fragmentos de ADN simultáneamente.

Ventajas:

- **Escalabilidad:** Alta capacidad de datos.
- **Rapidez:** Secuenciación de genomas completos en días.
- **Precisión:** Alta cobertura y profundidad de lectura

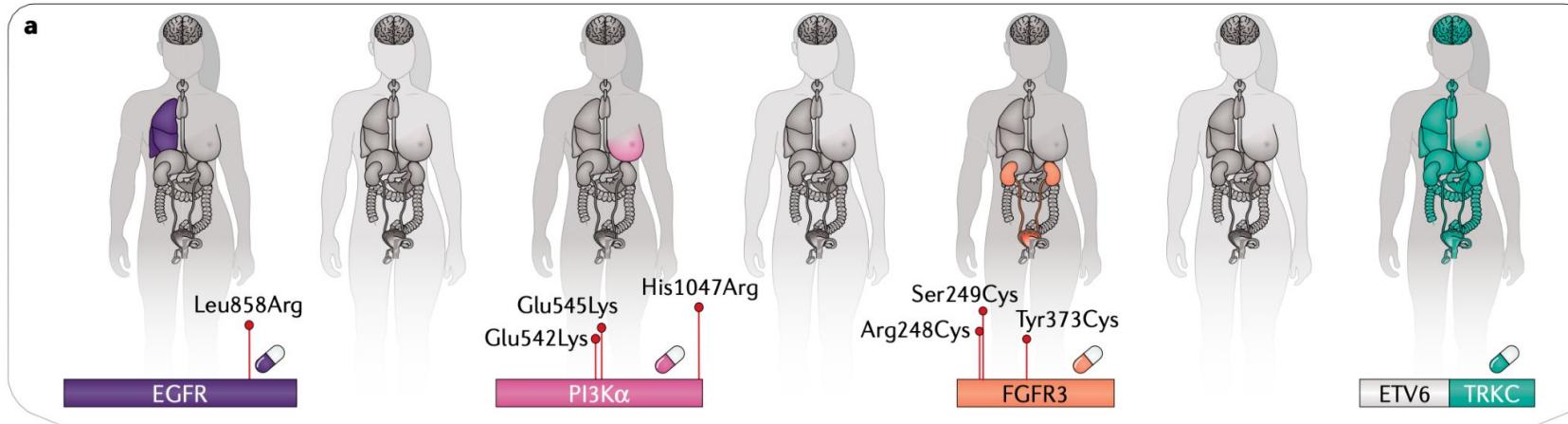
Principales aplicaciones:

- **Diagnóstico genético:** Detección de mutaciones causantes de enfermedades.
- **Investigación de cáncer:** Identificación de mutaciones somáticas y fusiones genéticas.
- **Medicina personalizada:** Terapias dirigidas basadas en el perfil genético.
- **Estudios de metagenómica:** Análisis de microbiomas.



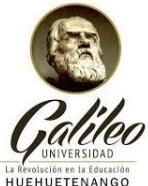


Aplicaciones principales: investigación genética, diagnóstico clínico

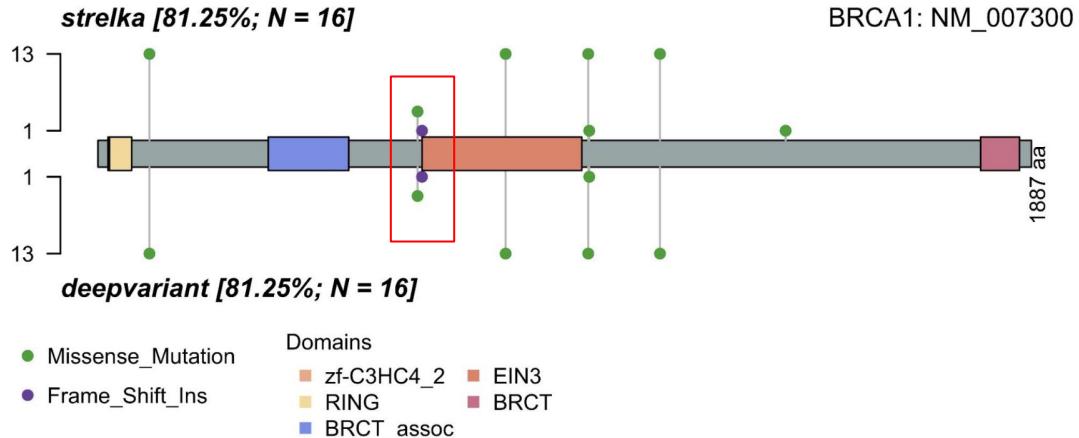


El perfil tumoral puede mejorar la atención del paciente al identificar mutaciones o variantes estructurales que actúan como biomarcadores predictivos de la respuesta a los medicamentos.

Medicina de precisión: Inhibidores de PARP



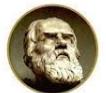
Los inhibidores de PARP, como **olaparib** y **rucaparib**, han mostrado eficacia en pacientes con cánceres asociados a mutaciones en BRCA.



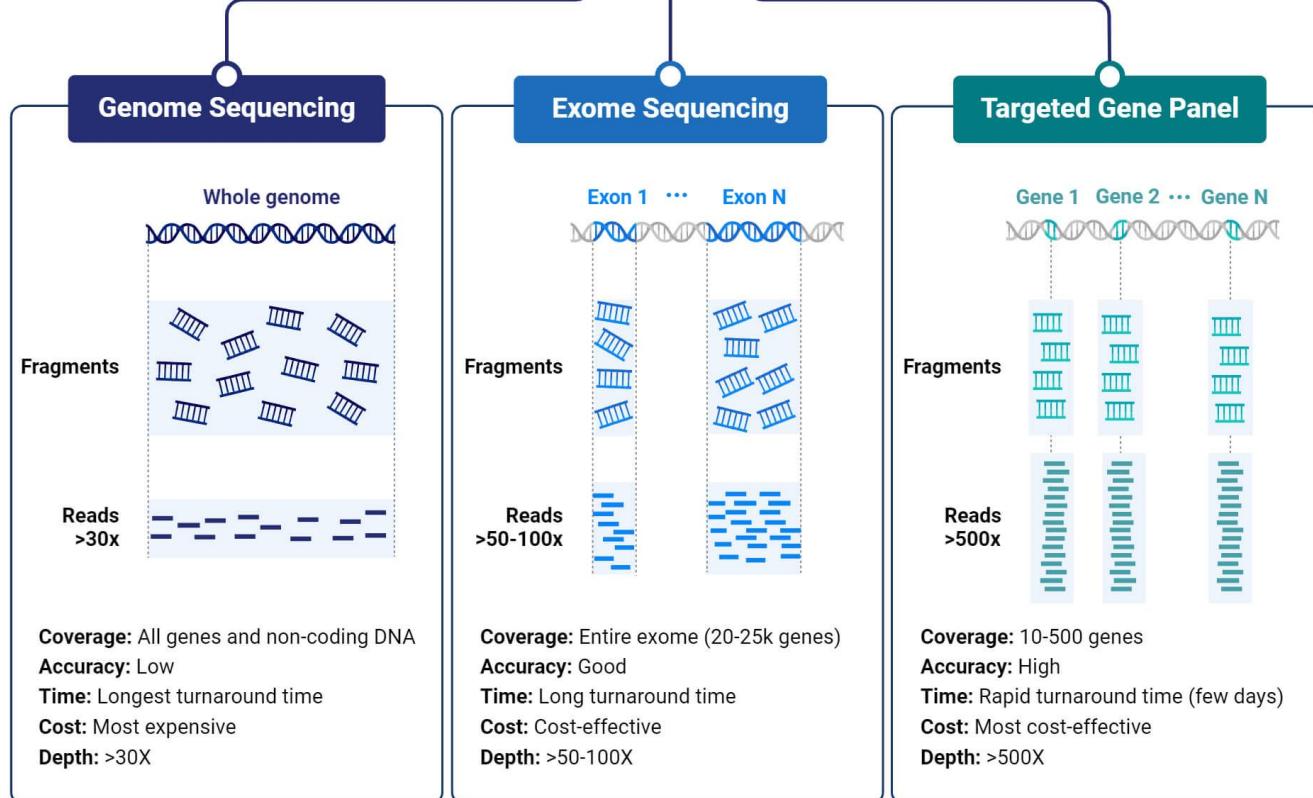
According to the OncoKB database (Chakravarty et al. 2017), this variant is classified as **likely oncogenic** with a **Level 1 evidence** classification, indicating that it is an FDA-recognized biomarker predictive of response to several **FDA-approved drugs**, including PARP inhibitors.

Oncokb		Levels of Evidence	Actionable Genes	Oncology Therapies	CDx	Cancer Genes	API / License	About	News	FAQ	Q	Account		Molecular Skin Karyotype Cancer Center
BRCA1	/ L655Ff*10													
BRCA1 L655Ff*10														
Likely Oncogenic	• Likely Loss-of-function	• - Level 10	- FDA Level 2	⊕										
BRCA1, a tumor suppressor involved in the DNA damage response, is mutated in various cancer types.														
The BRCA1 L655Ff*10 is a truncating mutation in a tumor suppressor gene, and therefore is likely oncogenic.														
Hide mutation effect description	⊕													
The mutation effect description for truncating mutations in BRCA1 is:														
Truncating mutations in BRCA1 can lead to varying C-terminally truncated proteins that result in aberrant protein folding, contributing to loss of BRCA1 protein function. Human breast cancer cell lines that contain a truncating mutation in BRCA1 display elevated levels of aneuploidy and impaired DNA damage response. In addition, truncating mutations have been shown to induce aberrant protein localization, which may impact the interaction of important binding partners (PMID: 20608970). Mouse models of BRCA1 truncating mutations develop cancer, including mammary carcinomas, lymphomas and ovarian carcinomas (PMID: 11358863 , 12483515 , 12947386).														
Select a cancer type														
Therapeutic	FDA-Recognized Content													
A list of the cancer type-specific BRCA1 alterations that may predict response to a targeted drug and the corresponding OncoKB™ level of evidence assigning their level of clinical activity .														
If you notice any mistakes or omissions, please reach out to us														Search ...
Level ▾ Alterations		▴ Level-associated cancer types	⊕		▴ Drugs			▴ Citations		Description				
	Oncogenic Mutations	Ovarian Cancer, Ovary/Fallopian Tube, Peritoneal Serous Carcinoma			Niraparib			3						
	Oncogenic Mutations	Ovarian Cancer, Ovary/Fallopian Tube, Peritoneal Serous Carcinoma			Olaparib			3						
	Oncogenic Mutations	Ovarian Cancer, Ovary/Fallopian Tube, Peritoneal Serous Carcinoma			Olaparib + Bevacizumab			3						
	Oncogenic Mutations	Ovarian Cancer, Ovary/Fallopian Tube, Peritoneal Serous Carcinoma			Rucaparib			4						

https://www.oncokb.org/gene/BRCA1/L655Ffs*10

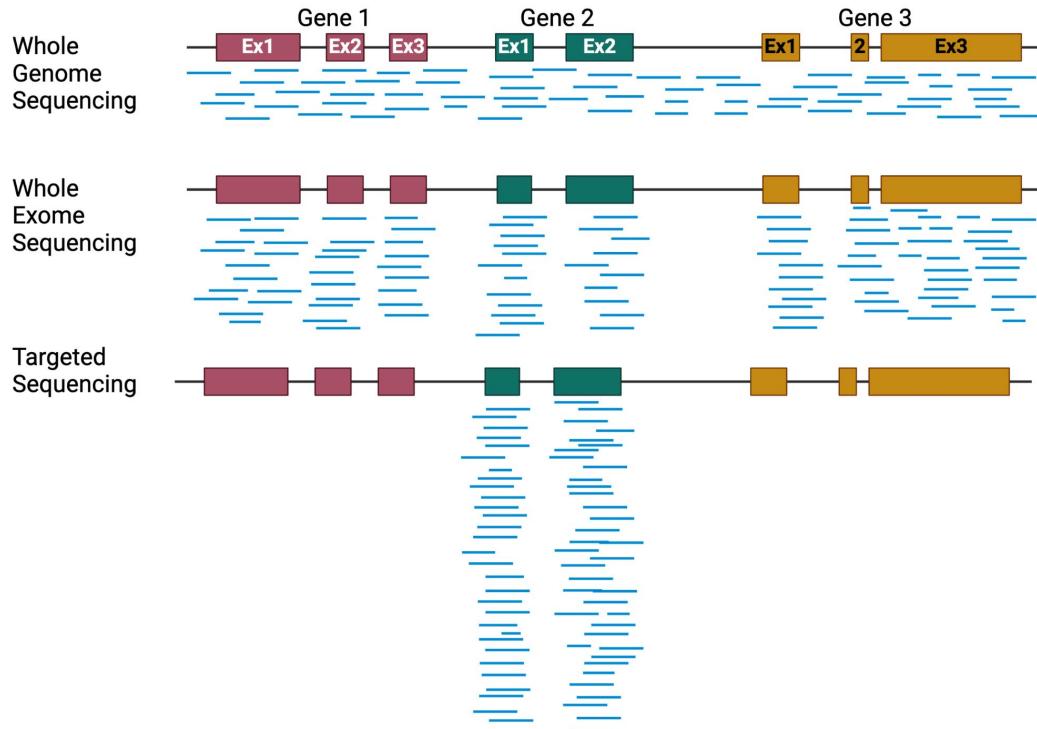


Next Generation Sequencing





Targeted Sequencing



Definición: Técnica que se centra en secuenciar regiones específicas del genoma, previamente seleccionadas.

Objetivo: Obtener información detallada de genes o regiones de interés, como genes asociados a enfermedades, sin necesidad de secuenciar el genoma completo.

Métodos comunes:

Captura por hibridación: Uso de sondas para capturar regiones objetivo.

Amplificación por PCR: Multiplicación de regiones específicas para secuenciar.

Ventajas:

- Mayor profundidad de cobertura en regiones clave.
- Menor costo y tiempo que la secuenciación de genoma completo.

Panel BRCA1/2, regiones codificantes

Products > By type > Sequencing kits > Library preparation kits > AmpliSeq for Illumina BRCA Panel



Sequencing



For Research Use Only



DNA

AmpliSeq for Illumina BRCA Panel
Targeted research panel investigating somatic and germline variants in *BRCA1* and *BRCA2*.

[AmpliSeq for Illumina BRCA Panel Data Sheet](#)
Data sheet | HTML externalFile

5 hours (...)	<1.5 hr	1–100 ng ...
Assay time	Hands-on time	Input quantity

[See full details in the specifications table](#)

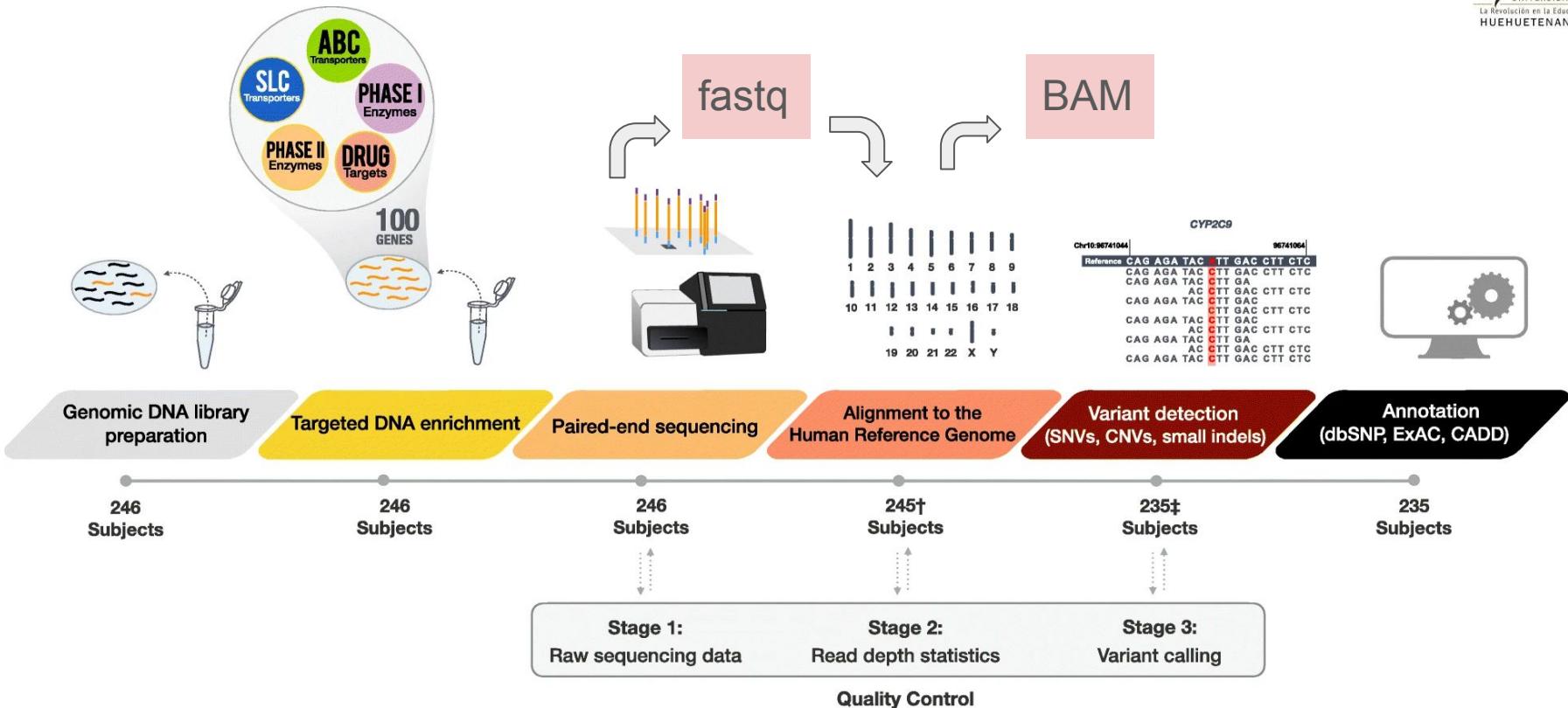
Se secuenciaron 16 pacientes con cáncer de mama.

Utilizando el panel **AmpliSeq for Illumina BRCA**.

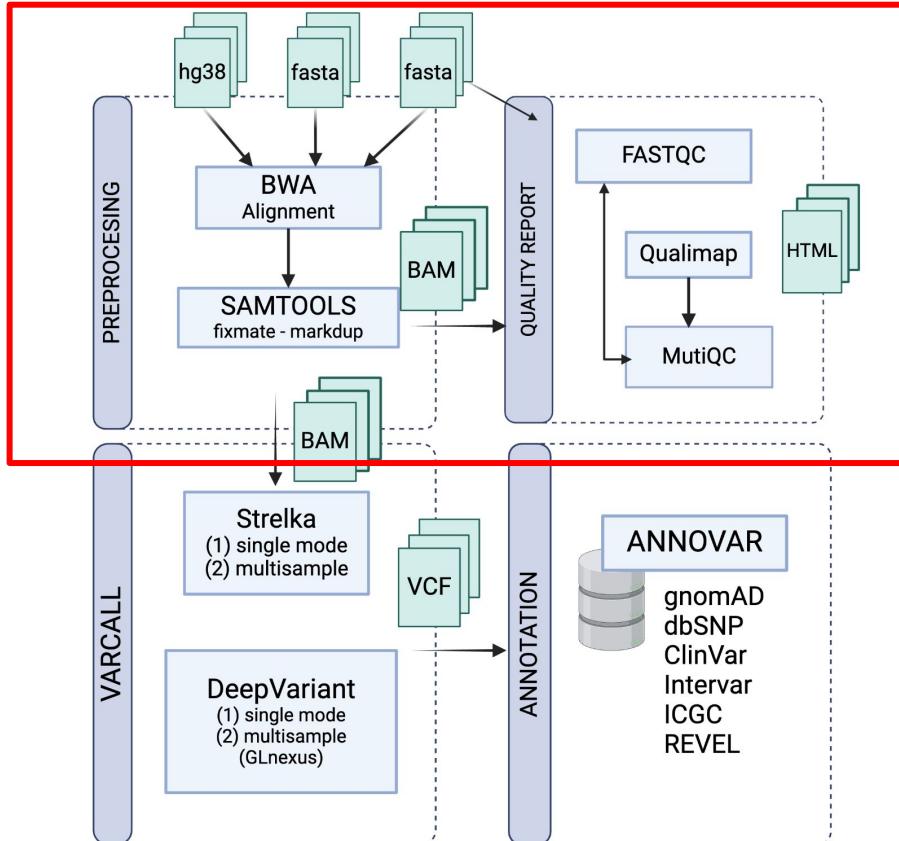
El panel incluye todas las regiones exónicas de los genes **BRCA1** y **BRCA2** y secuencias intrónicas adyacentes.

Cubre un total de **22 Kb**.

Flujo de trabajo “Targeting Sequencing”



Flujo de trabajo bionformático para la detección de variantes en *BRCA1* y *BRCA2*



Por qué es importante revisar las métricas de secuenciación?

Qué tipos de errores podemos detectar. ?

The workflow, including the configurations and tools, is publicly available on the GitHub repository:
<https://github.com/digenoma-lab/BRCA>.

Archivos de entrada y salida

			Approximate File Size (Average Coverage 160×)	
File Type	Full Name	Description	Exome	4800 Genes
FASTQ	Files with consensus assessment of sequence and variation	Raw sequencing data after demultiplexing	50 GB	18 GB
BAM	Binary version of sequence alignment/map	Sequencing data after alignment	16 GB	6 GB
VCF	Variant call file	File containing variants called relative to the reference	9.3 GB	3.5 MB

Abbreviations: GB, gigabytes; MB, megabytes.

Archivo *.fastq

Identifier Sequence ‘+’ sign Quality scores	@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50 TTGCCTGCCTATCATTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT + hhhhhhhhhhghhhhhhhfhhhhfffffe'ee['X]b[d[ed' [Y[^Y
Identifier Sequence ‘+’ sign Quality scores	@SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50 GATTGTATGAAAGTATAACACTAAACTGCAGGTGGATCAGAGTAAGTC + hhhhgfhhcghghggfcffdhfehhhcuhdchhdhaehffffde'bVd



N1_R1.fastq.gz
N1_R2.fastq.gz



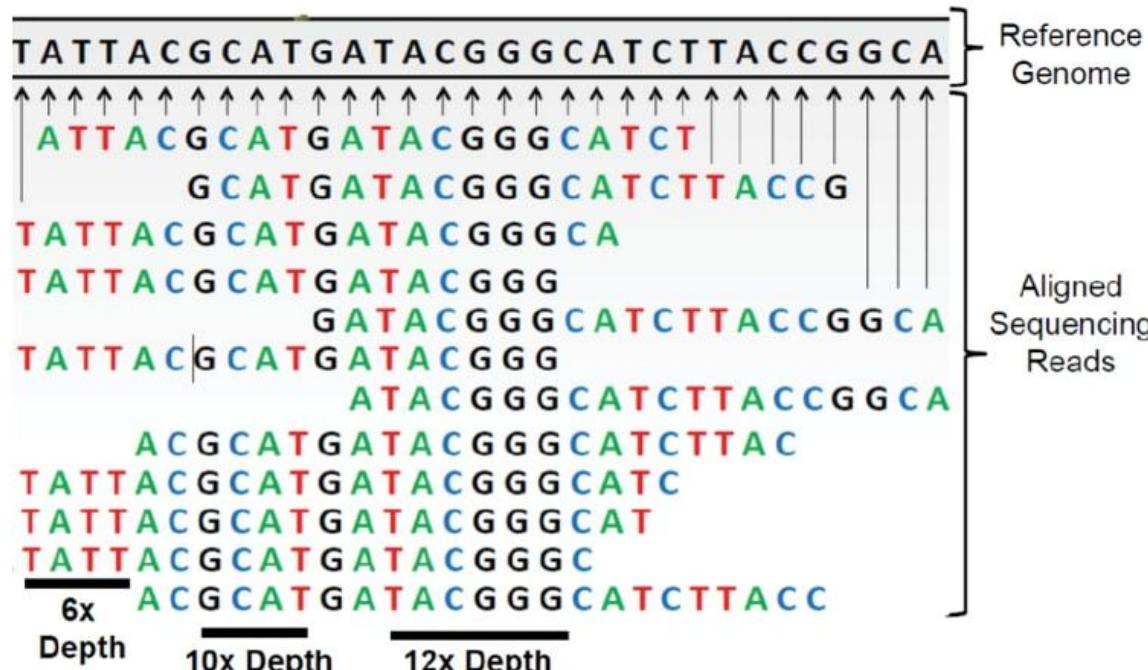
N2_R1.fastq.gz
N2_R2.fastq.gz



N3_R1.fastq.gz
N3_R2.fastq.gz



Archivos *.bam



BAM: Formato binario para el almacenamiento de datos de secuenciación.

La extensión de archivo (.bam) contiene información sobre lecturas de secuencias después de haber sido estas **alineadas contra un genoma de referencia**.

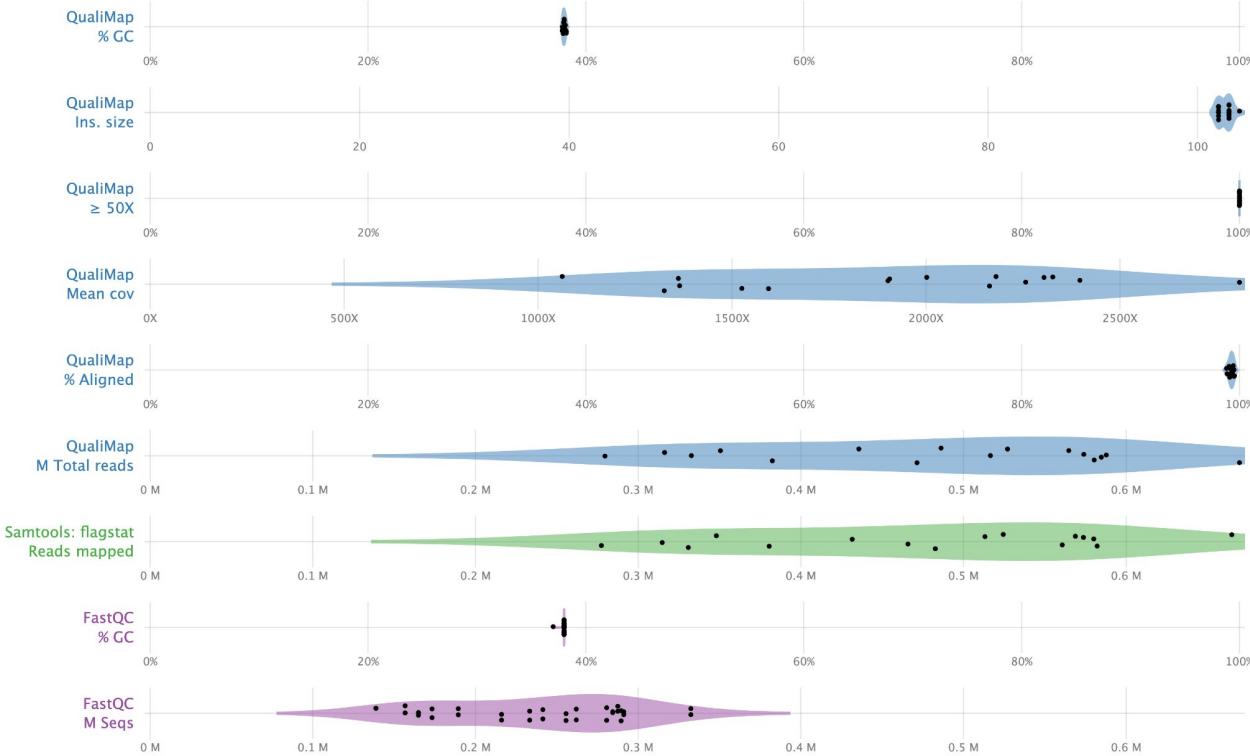
Utilizado para almacenar datos de secuenciación masiva.

Reporte de calidad de secuenciación

QualiMap
*.bam
(input)

FastQC
*.bam
(input)

General Statistics



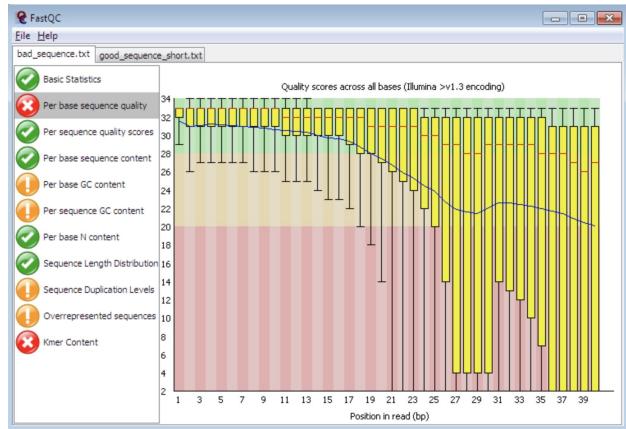
Escala Phred - Calidad de las lecturas secuenciadas

Identifier → @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
 Sequence → TTGCCTGCCTATCATTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
 '+' sign → +
 Quality scores → hhhhhhhhhhhghhhhhhhfhhhhfffffe 'ee['X]b[d[ed' [Y[^Y



Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Reporte de Calidad: FastQC



Permite identificar posibles problemas en las lecturas generadas.

Primer paso en el procesamiento de datos de secuenciación crudos.

Proporciona información sobre la calidad de las lecturas de secuenciación de alto rendimiento (HTS).

- Errores en la identificación de bases.
- Lecturas/bases de baja calidad.
- Contaminación por cebadores/adaptadores

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>



Pair end raw reads

read 1

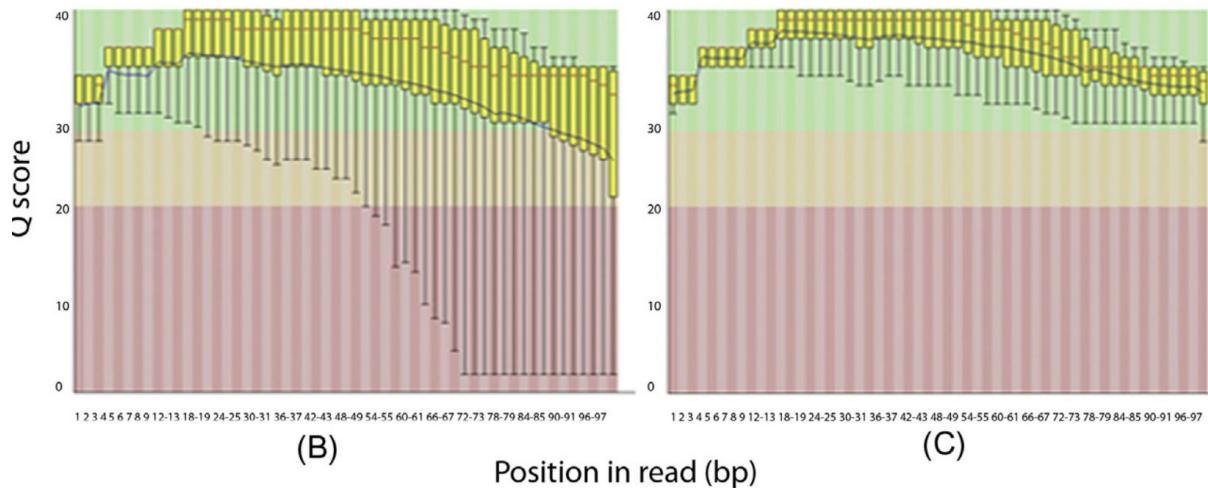
```
@ERR160122.1 HWI-ST478:264:DOMCOACXX:1:1101:1221:2020/1
NTTGATAGCTGGCTGCAAGGAATTCTAGATATACAGTTAAGGATAATGAAAAGAAAACACTGAATACTTGAACA
+
#1=DDDDFFDFHHGGGHIIIGGIGFH@HDEGDGIGGDHHIIJJGGGIIGGGIGEIFIJJHIIIEHGJJJJH
```

read 2

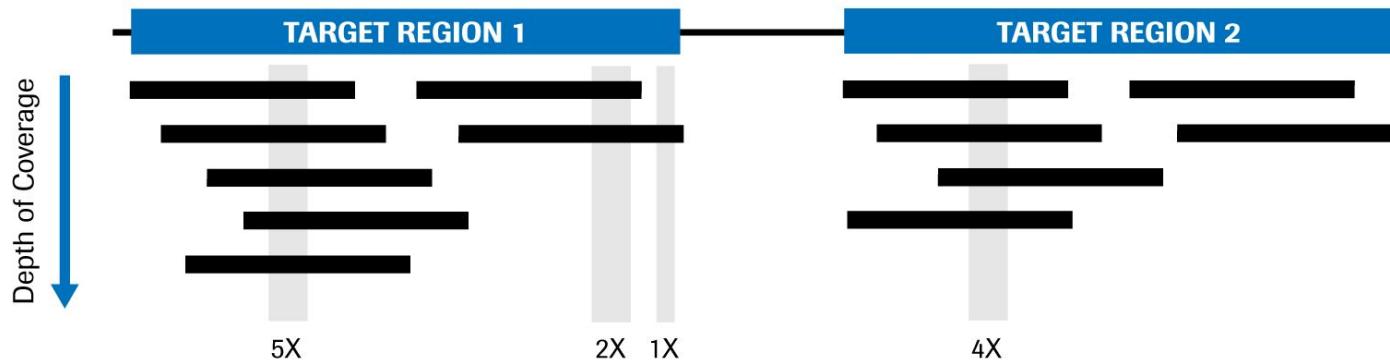
```
@ERR160122.1 HWI-ST478:264:DOMCOACXX:1:1101:1221:2020/2
CTTCATAAGTATGAATCATCTTACCAATTACTTCATTCTCTGTAAATTCTCCTTCCATTGGAAATCTG
+
@@CFFFFFFHCFHIGGGEDDHIGGHJEBEHECFHIIEIIIJGIJJJFDGHIIJDGIJIIJJJJJJJIJGH
```

(A)

Quality scores across all bases (Sanger/ Illumina 1.9 encoding)



Promedio de cobertura en regiones “On-target”

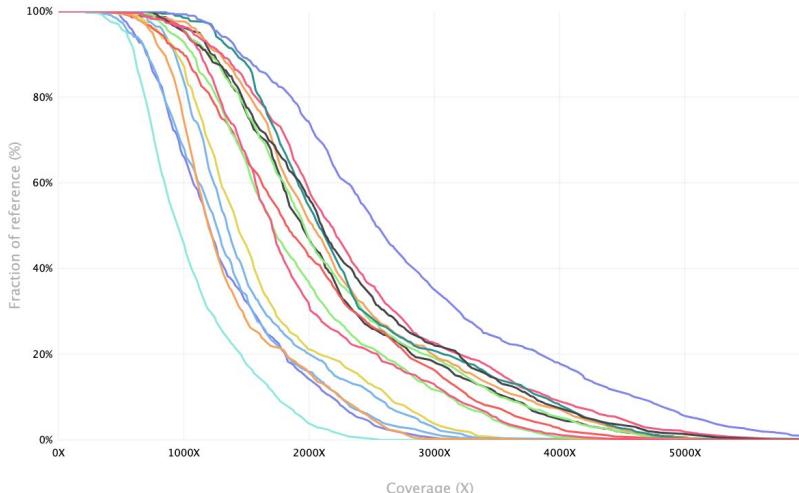


Cual es la cobertura mínima que necesitamos para detectar una variante en BRCA1/2. ? Linea germinal
 Cuales con los promedios de cobertura que se obtuvieron por en los pacientes de Cáncer de mama.

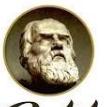
Promedio de cobertura en pacientes con cáncer de mama

Sample Name	Mean cov ▾
VM.md	2 808.1 X
LV.md	2 396.8 X
ML.md	2 326.9 X
PV.md	2 304.2 X
JM.md	2 257.0 X
JC.md	2 180.6 X
DC.md	2 163.8 X
NS.md	2 002.0 X
PZ.md	1 906.2 X
YA.md	1 902.0 X
MC.md	1 594.1 X
AL.md	1 525.3 X
RQ.md	1 364.9 X
PP.md	1 361.7 X
KV.md	1 325.3 X

QualiMap: BamQC: Genome fraction covered by at least X reads

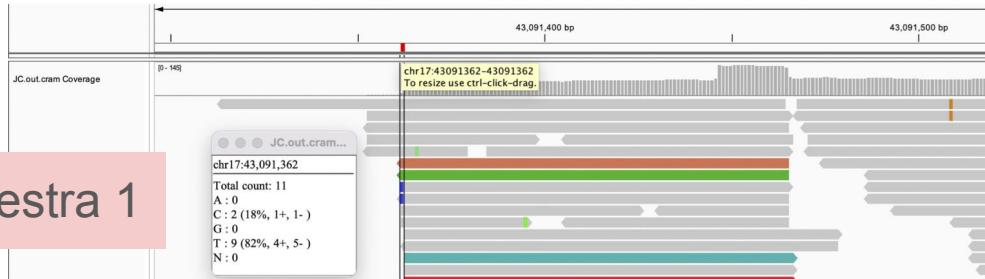


Created with MultiQC

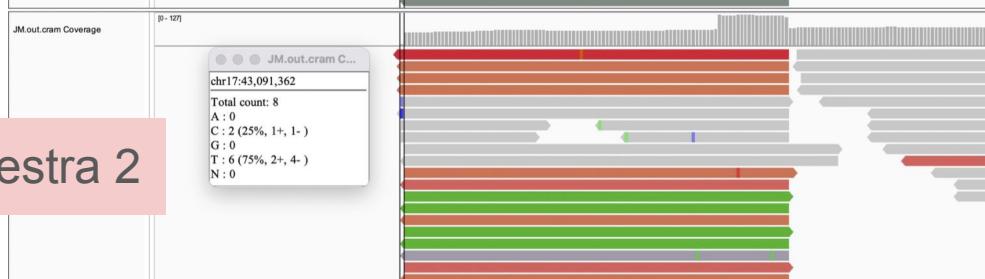


Ejemplo: Variante r p.E1390G BRCA1

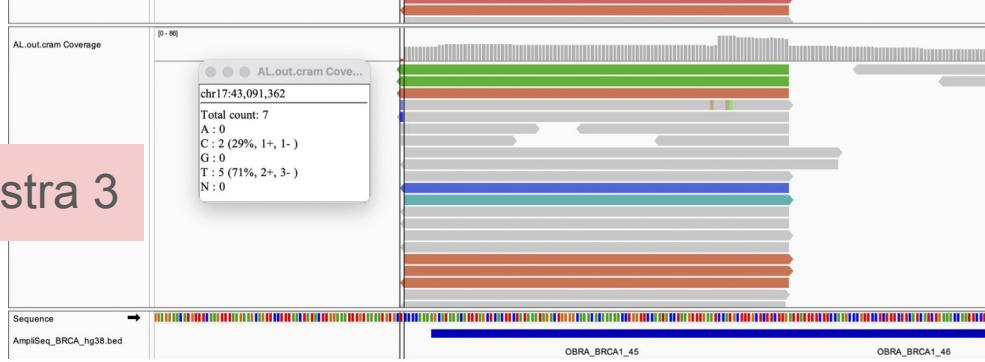
Muestra 1



Muestra 2



Muestra 3



Inspección visual de las lecturas alineadas al genoma de referencia de tres pacientes.

Cobertura en regiones On-targets



- Cual es el porcentaje de las lecturas iniciales que se encuentran en las regiones blanco?
- Que indica un bajo % de lecturas On-targets?
- Como se puede observar esta métrica en un reporte de calidad ?

Uniformidad de la cobertura de las regiones “on-target”

Ideal Uniformity



desired coverage

targeted regions

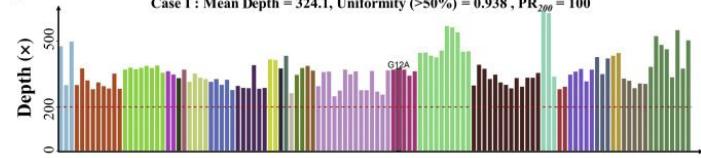
- On-target rate = 100%
- All target regions have exactly the desired coverage
- Fold-80 = 1

Observed Uniformity

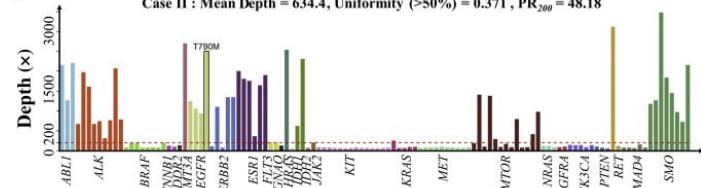


- On-target rate < 100%
- Target regions captured at various levels, and reads mapping to off-target regions
- Fold-80 > 1

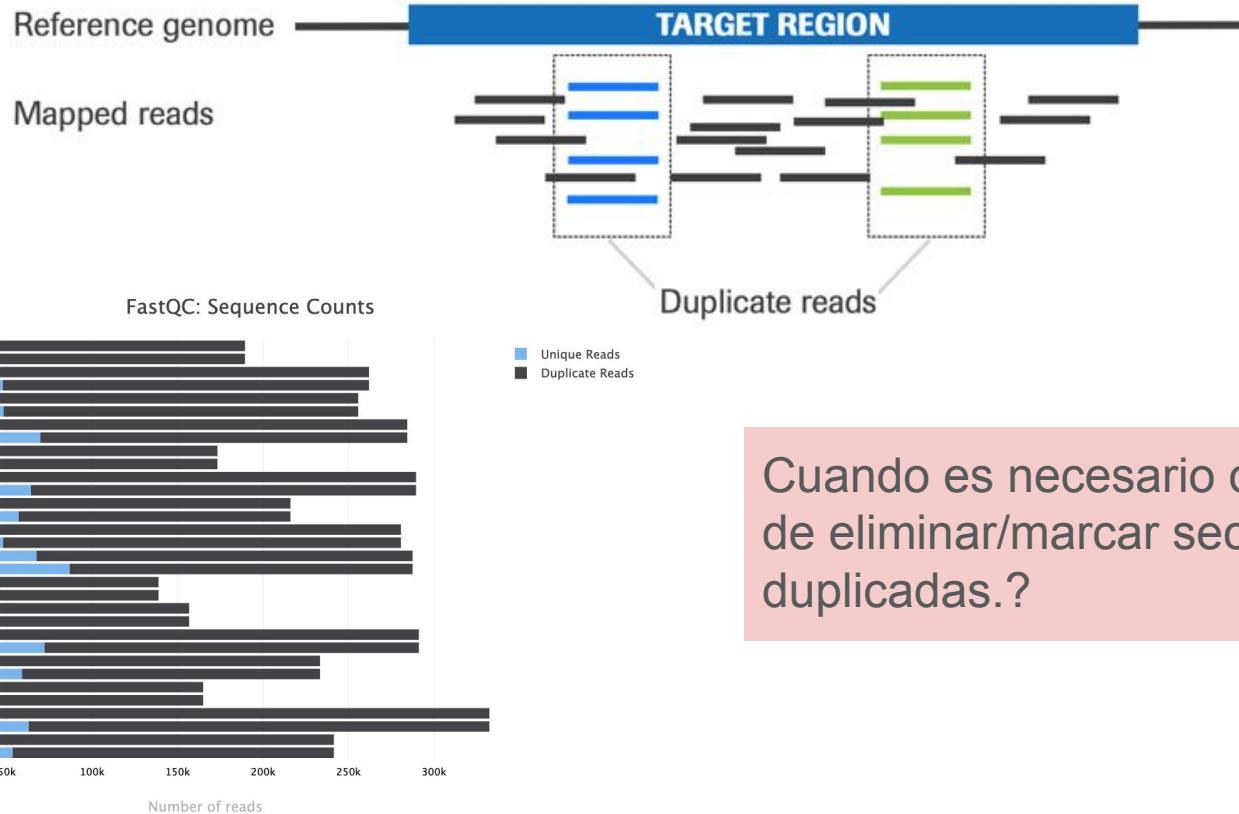
A



B



Número de lecturas Duplicadas



Cuando es necesario omitir el paso de eliminar/marcar secuencias duplicadas.?

Reporte HTML: MultiQC

multqc
v1.22.3

[General Stats](#)

[QualiMap](#)

Coverage histogram

Cumulative genome coverage

Insert size histogram

GC content distribution

[Samtools](#)

[FastQC](#)

Sequence Counts

Sequence Quality Histograms

Per Sequence Quality Scores

Per Base Sequence Content

Per Sequence GC Content

Per Base N Content

Sequence Length Distribution

Sequence Duplication Levels

Overrepresented sequences by sample

Top overrepresented sequences

Adapter Content

Status Checks

[Software Versions](#)

multqc

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2024-06-26, 13:51 -04 based on data in: /mnt/beegfs/home/efeliu/work2024/080524_nextflow_BRCA/Runs_nextflow_BRCA/RUN1

Welcome! Not sure where to start?

[Watch a tutorial video](#) (6:06)

don't show again 

General Statistics

Sample Name	% GC	Ins. size	≥ 30X	Median cov	Mean cov	% Aligned	Reads mapped	% Dups	% GC	M Seqs
AL							0.4 M			
AL.md				1525.3 X	99.5 %					
AL.qualimap	38 %	102	100.0 %	1344 X						
AL_S9.R1							87.3 %	38 %	0.2 M	
AL_S9.R2							79.8 %	38 %	0.2 M	
DC						0.5 M				
DC.md				2163.8 X	99.5 %					
DC.qualimap	38 %	102	100.0 %	1942 X						
DC_S14.R1							88.8 %	38 %	0.3 M	
DC_S14.R2							81.8 %	38 %	0.3 M	
JC						0.5 M				
JC.md				2180.6 X	99.4 %					
JC.qualimap	38 %	103	100.0 %	1956 X						
JC_S13.R1							91.6 %	38 %	0.3 M	
JC_S13.R2							81.1 %	38 %	0.3 M	

QualiMap