

# 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

**medRxiv**  
THE PREPRINT SERVER FOR HEALTH SCIENCES

 **BMJ** Yale

HOME | SUBMIT | FAQ | BLOG | ALERTS / RSS | RESOURCES | ABOUT

Advanced Search

 Follow this preprint

Previous


Next


**A workflow for clinical profiling of BRCA genes in Chilean breast cancer patients via targeted sequencing**


Evelin González, Rodrigo Moreno Salinas, Manuel Muñoz, Soledad Lantadilla Herrera, Mylene Cabrera Morales, Pastor Jullian, Waleska Ebner Durreles, Gonzalo Viguera Stari, Javier Anabalón Ramos, Juan Francisco Miquel, Lilian Jara, Carol Moraga, Alex Di Genova


doi: <https://doi.org/10.1101/2024.09.25.24314295>

**This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.**


 0


 0

 0

 0

 0

 0


 11

Abstract

Full Text

Info/History

Metrics

 Preview PDF

**Abstract**

**Background** Breast cancer (BC) is the leading cause of cancer-related deaths among women globally and in Chile. Mutations in the tumor-suppressor genes BRCA1 and BRCA2 significantly increase the risk of developing cancer, with the probability rising by more than 50%. Identifying pathogenic variants in BRCA1 and BRCA2 is crucial for both diagnosis and treatment. Targeted panels, which focus on medically relevant subsets of genes, have become essential tools in precision oncology. Beyond technical and human resource factors, standardized bioinformatics workflows are essential for the accurate interpretation of results. We developed a robust bioinformatics pipeline, implemented with Nextflow, to process sequencing data from targeted panels to identify germline variants.

Download PDF

Print/Save Options

Author Declarations

Supplementary Material


Data/Code


Email

Share

Citation Tools

Get QR code

 Post

 Me gusta 0

**COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv**

**Subject Area**

**Genetic and Genomic Medicine**

**Subject Areas**

**All Articles**

- Addiction Medicine
- Allergy and Immunology
- Anesthesia
- Cardiovascular Medicine
- 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)

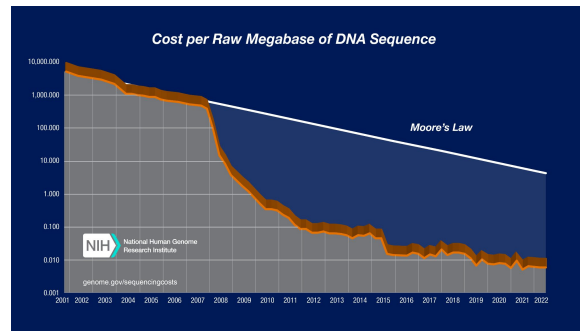
- Tecnología de secuenciación masiva y paralela.
- Permite la secuenciación de millones de fragmentos de ADN simultáneamente.

### Ventajas:

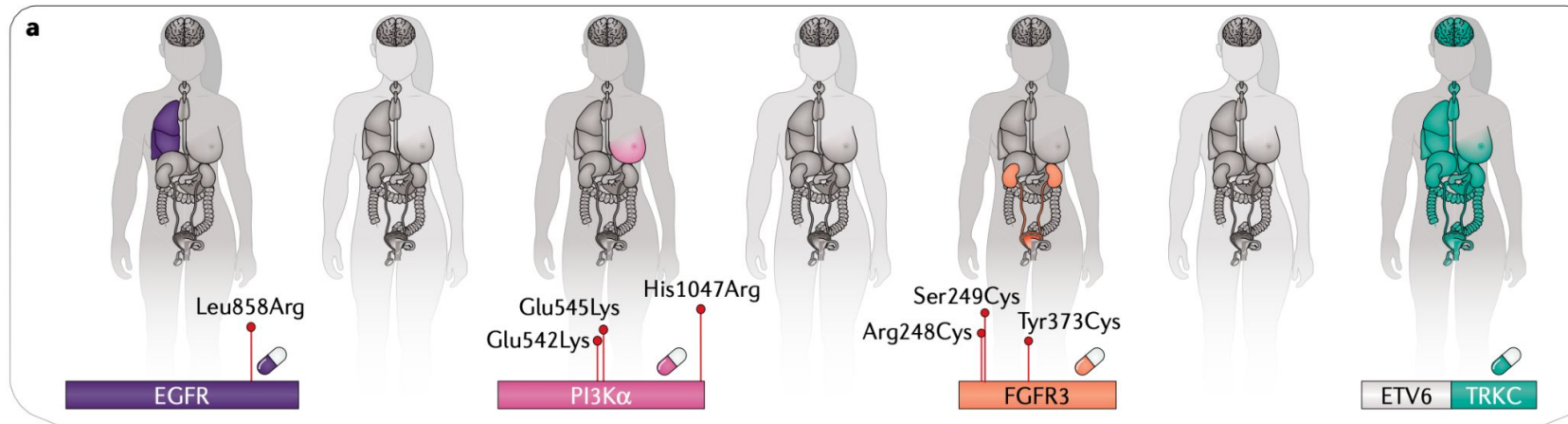
- **Escalabilidad:** Alta capacidad de datos.
- **Rapidez:** Secuenciación de genomas completos en días.
- **Precisión:** Alta cobertura y profundidad de lectura

### 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.



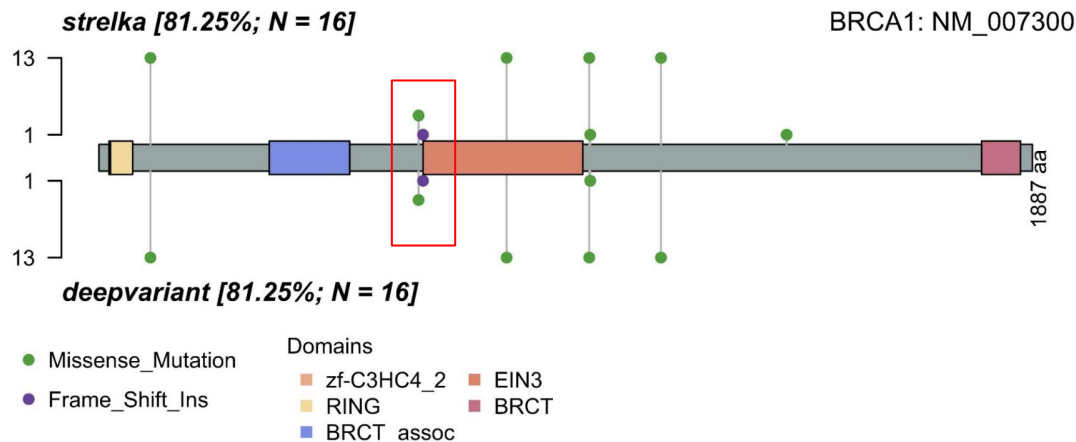
# Aplicaciones principales: investigación genética, diagnóstico clínico



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 CDIs Cancer Genes API / License About News FAQ Account

BRCA1 / L655Ffs\*10

**BRCA1 L655Ffs\*10**

**Likely Oncogenic** • **Likely Loss-of-function** • **Level 1** • **FDA Level 2**

BRCA1, a tumor suppressor involved in the DNA damage response, is mutated in various cancer types.

The BRCA1 L655Ffs\*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 results 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**

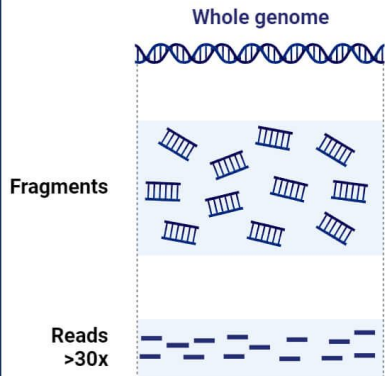
A list of the cancer type-specific BRCA1 alterations that may predict response to a targeted drug and the corresponding OncoKB® level of evidence assigning their level of **clinical actionability**.  
If you notice any mistakes or omissions, please reach out to us.

| Level | Alterations         | Level-associated cancer types                                     | Drugs                  | Citations | Description |
|-------|---------------------|---|------------------------|-----------|-------------|
| 1     | Oncogenic Mutations | Ovarian Cancer, Ovary/Fallopian Tube, Peritoneal Serous Carcinoma | Niraparib              | 3         | #           |
| 1     | Oncogenic Mutations | Ovarian Cancer, Ovary/Fallopian Tube, Peritoneal Serous Carcinoma | Olaparib               | 3         | #           |
| 1     | Oncogenic Mutations | Ovarian Cancer, Ovary/Fallopian Tube, Peritoneal Serous Carcinoma | Olaparib + Bevacizumab | 3         | #           |
| 1     | Oncogenic Mutations | Ovarian Cancer, Ovary/Fallopian Tube, Peritoneal Serous Carcinoma | Rucaparib              | 4         | #           |

[https://www.oncokb.org/gene/BRCA1/L655Ffs\\*10](https://www.oncokb.org/gene/BRCA1/L655Ffs*10)

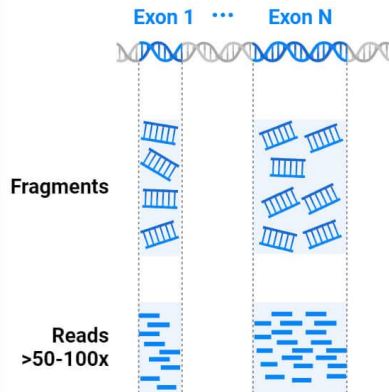
# Next Generation Sequencing

## Genome Sequencing



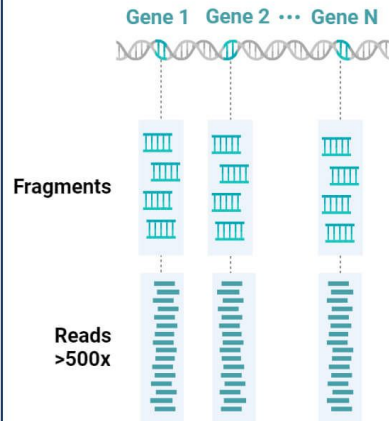
**Coverage:** All genes and non-coding DNA  
**Accuracy:** Low  
**Time:** Longest turnaround time  
**Cost:** Most expensive  
**Depth:** >30X

## Exome Sequencing



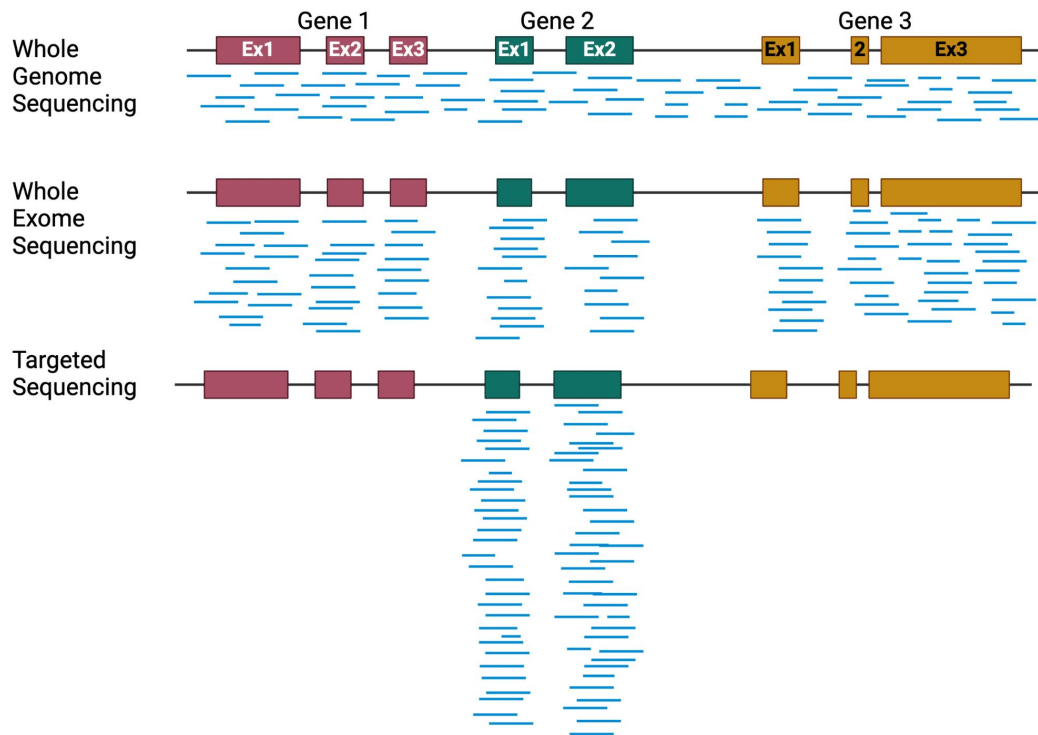
**Coverage:** Entire exome (20-25k genes)  
**Accuracy:** Good  
**Time:** Long turnaround time  
**Cost:** Cost-effective  
**Depth:** >50-100X

## Targeted Gene Panel



**Coverage:** 10-500 genes  
**Accuracy:** High  
**Time:** Rapid turnaround time (few days)  
**Cost:** Most cost-effective  
**Depth:** >500X

# 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 (...)

Assay time



<1.5 hr

Hands-on time



1–100 ng ...

Input quantity

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

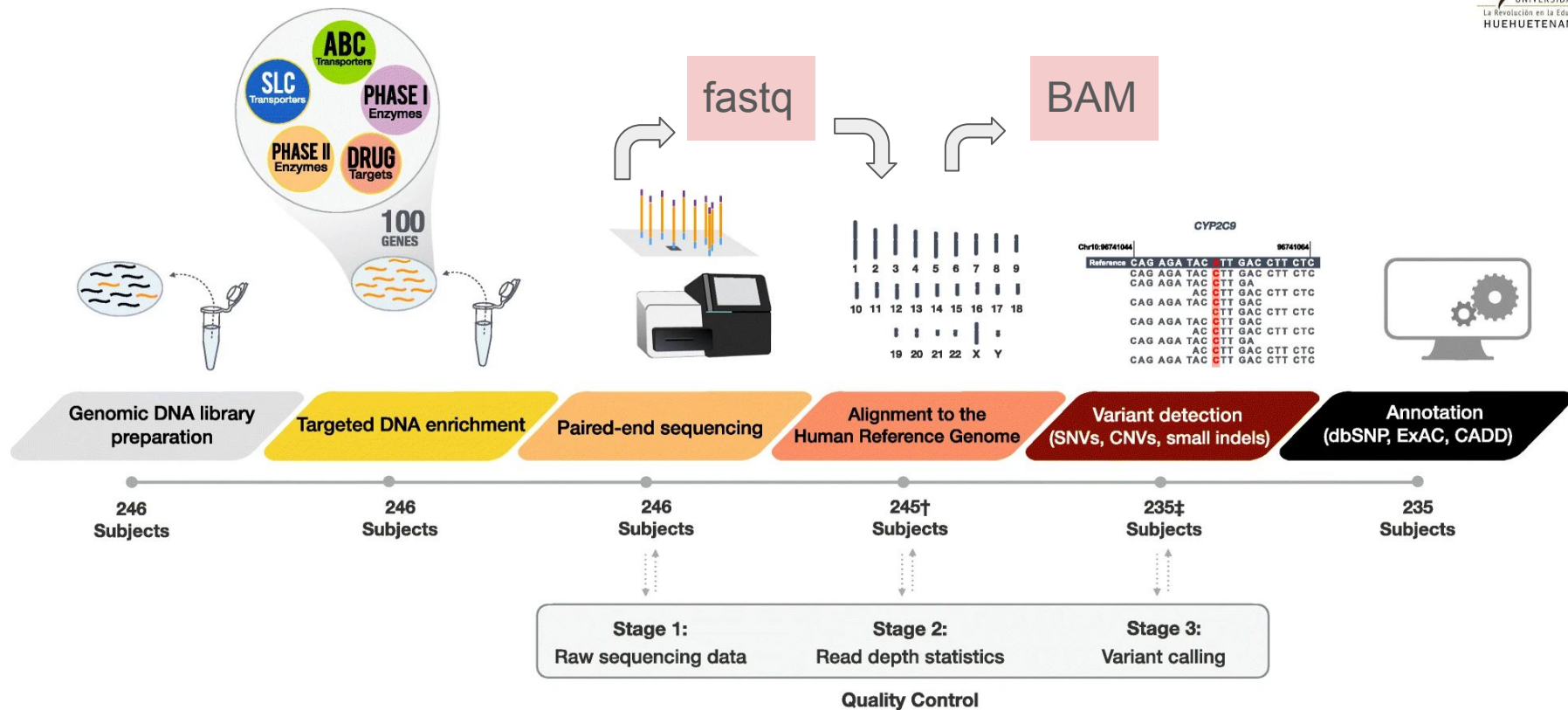
Se secuenciaron 16 pacientes con cáncer de mama.

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

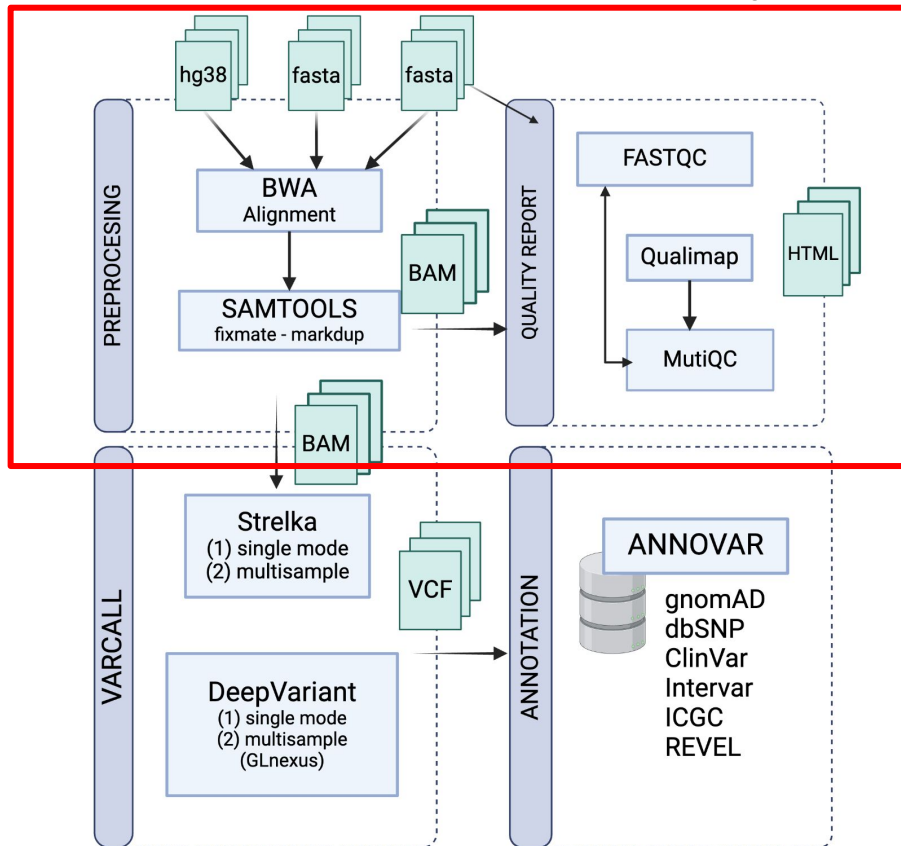
El panel incluye todas las regiones exónicas de los genes **BRCA1** y **BRCA2** y secuencias intrónicas adyacentes.

Cubre un total de **22 Kb**.

# Flujo de trabajo “Targeting Sequencing”



# Flujo de trabajo bioinformático 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

| File Type | Full Name   | Description   | Approximate File Size (Average Coverage 160×) |            |
|-----------|---|---|---|------------|
|           |   |   | 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     | ● @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50 |
| Sequence       | ● TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT                 |
| '+' sign       | ● +  |
| Quality scores | ● hhhhhhhhhghghghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y                 |
| Identifier     | ● @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50 |
| Sequence       | ● GATTTGTATGAAAGTATACAACATAAACTGCAGGTGGATCAGAGTAAGTC                 |
| '+' sign       | ● +  |
| Quality scores | ● hhhhghfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd                |



N1\_R1.fastq.gz  
N1\_R2.fastq.gz

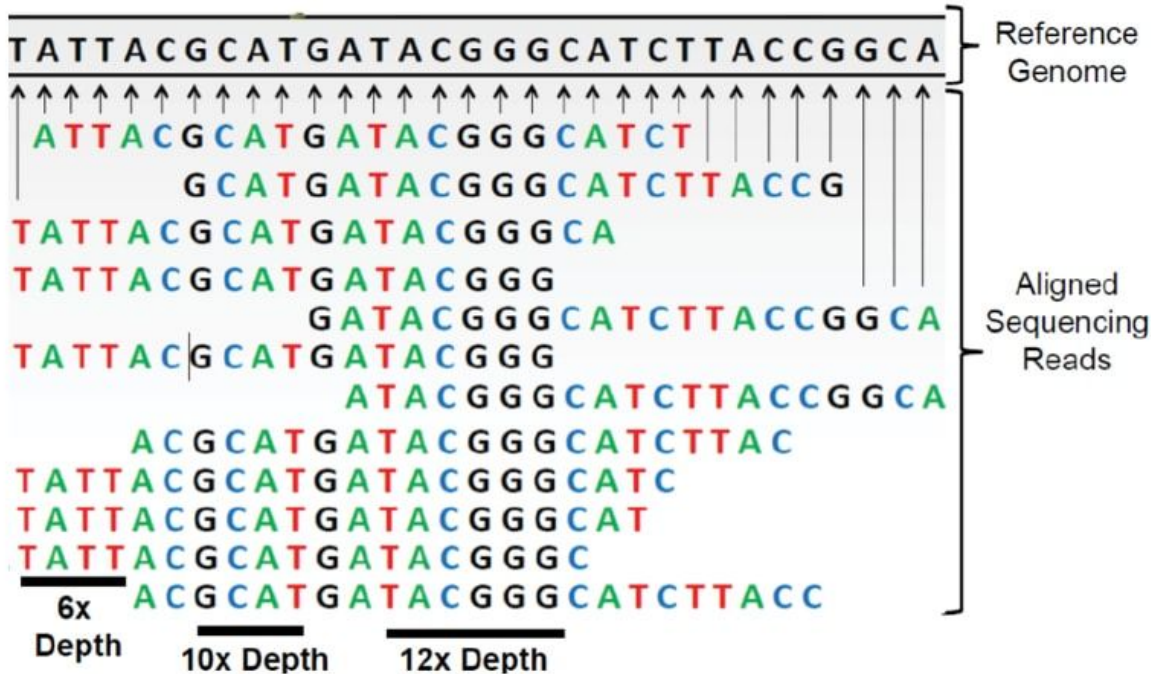


N2\_R1.fastq.gz  
N2\_R2.fastq.gz



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**.

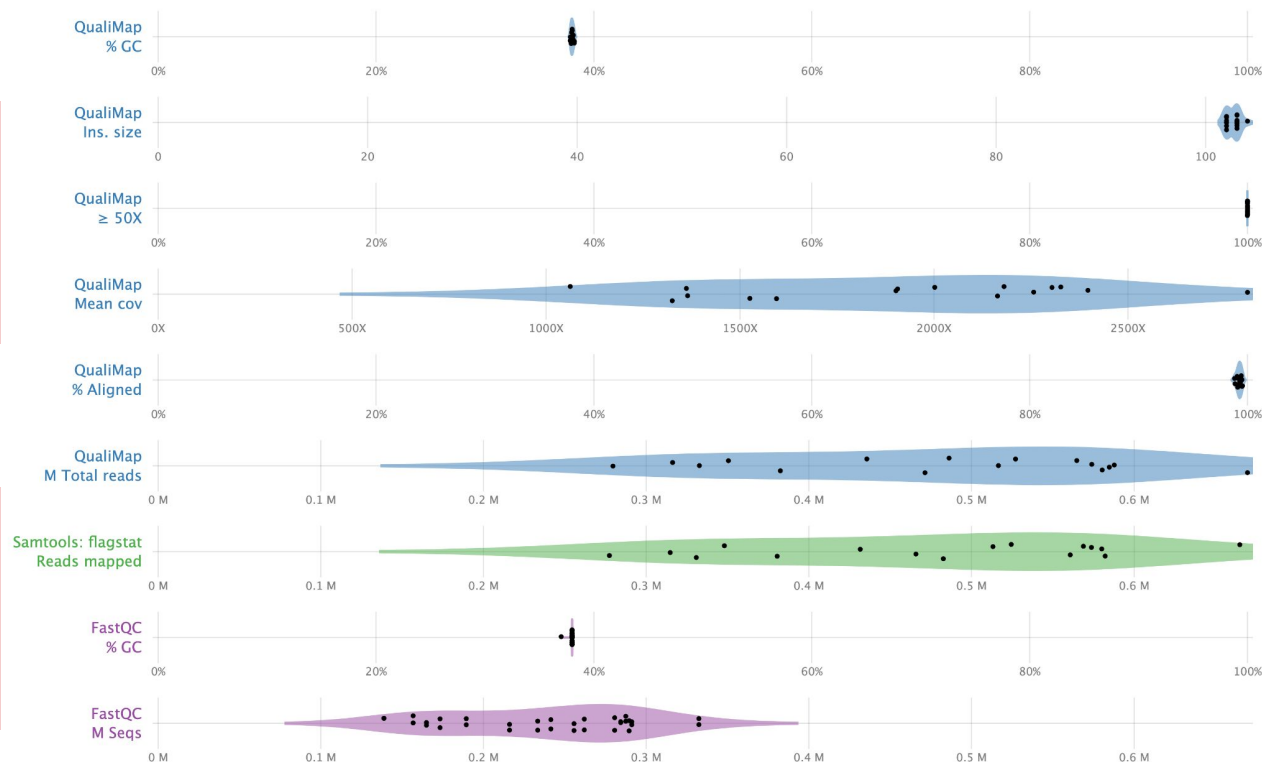
Utilizado para almacenar datos de secuenciación masiva.

# Reporte de calidad de secuenciación

QualiMap  
\*.bam  
(input)

FastQC  
\*.bam  
(input)

## General Statistics



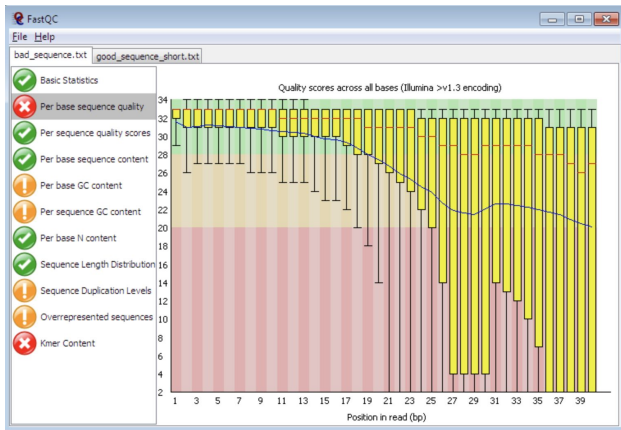
# Escala Phred - Calidad de las lecturas secuenciadas

Identifier —● @SRR566546.970 HWUSI-EAS1673\_11067\_FC7070M:4:1:2299:1109 length=50  
 Sequence —● TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT  
 '+' sign —● +  
 Quality scores —● hhhhhhhhhhhghhghhhhhfhhhhhfffffe'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

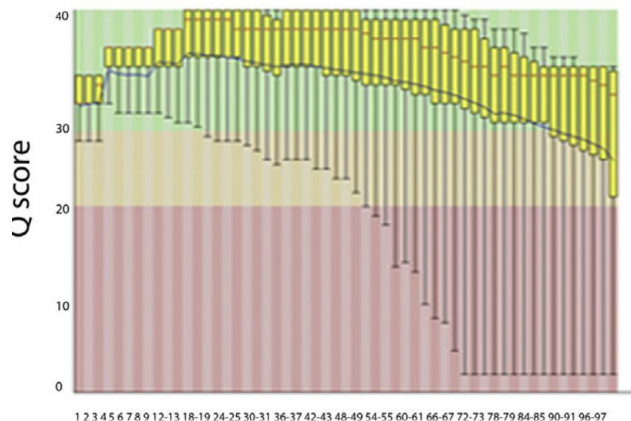
```
@ERR160122.1 HWI-ST478:264:D0MCOACXX:1:1101:1221:2020/1
NTTGATAGCTGGCTGCAAGGAATTTCTAGATATACAGTTAAGGATAAATGAAAAGAAAACACTGAATACTTTGAACA
+
#1=DDDDFFDFHHGGGHHIIIGGIGFHH@HDEGDGIGGDHHIIJGGGIIGIGGIGEIFIIJHHHIEHEHGJJJJJH
```

read 2

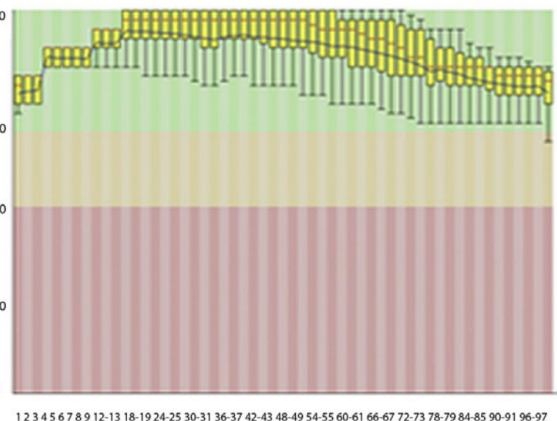
```
@ERR160122.1 HWI-ST478:264:D0MCOACXX:1:1101:1221:2020/2
CTTTCATAAGTATGAATCATCTTTACCAATTTACTTTTCATTCTCTTGTTTTAATTCTTCCTTCCATTGGAAATCTG
+
@@CFFFFFFFHCFFHIGGGGEDDHIGGHJEJEHECFHHIEIIIIJGIJJJFDGHHIIJDGIJIIIIJJJJIIIIJJJIJGH
```

(A)

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



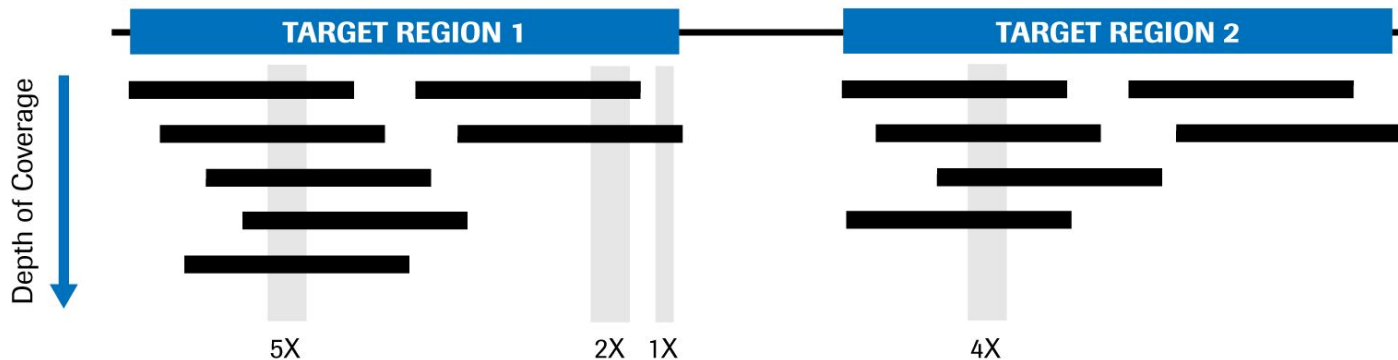
(B)



(C)

Position in read (bp)

# Promedio de cobertura en regiones “On-target”

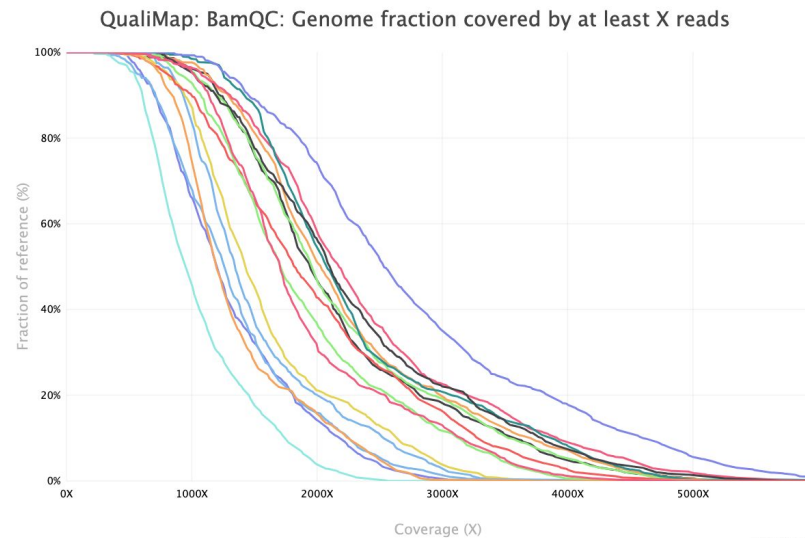


Cual es la cobertura mínima que necesitamos para detectar una variant en BRCA1/2. ? Linea germinal

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

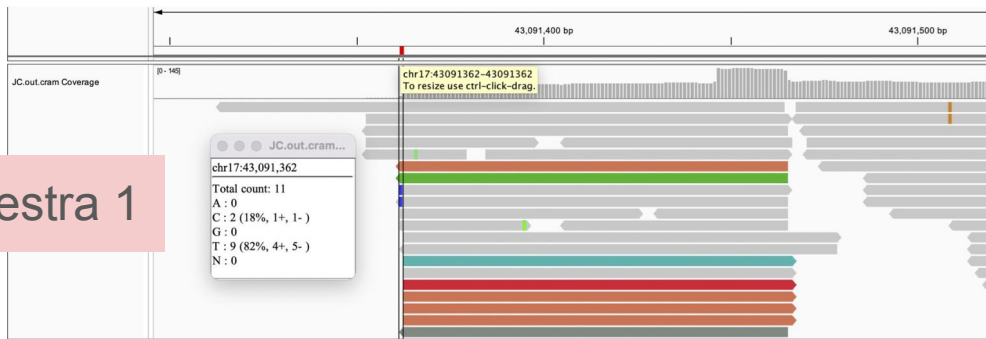
| Sample Name | Mean cov ▾ |
|-------------|------------|
| VM.md       | 2 808.1 X  |
| LV.md       | 2 396.8 X  |
| ML.md       | 2 326.9 X  |
| PV.md       | 2 304.2 X  |
| JM.md       | 2 257.0 X  |
| JC.md       | 2 180.6 X  |
| DC.md       | 2 163.8 X  |
| NS.md       | 2 002.0 X  |
| PZ.md       | 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  |



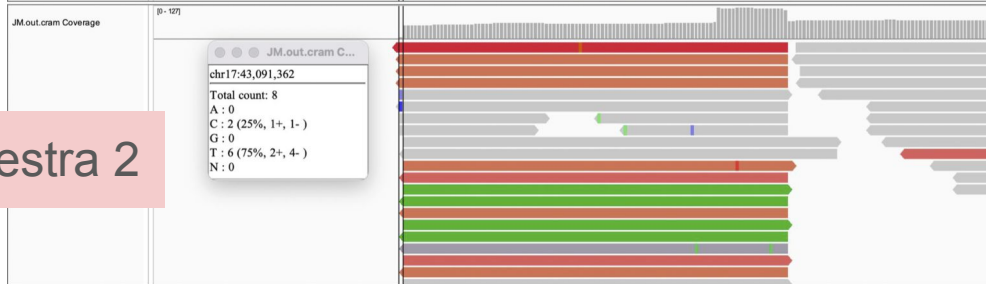
# Ejemplo: Variante r p.E1390G *BRCA1*

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

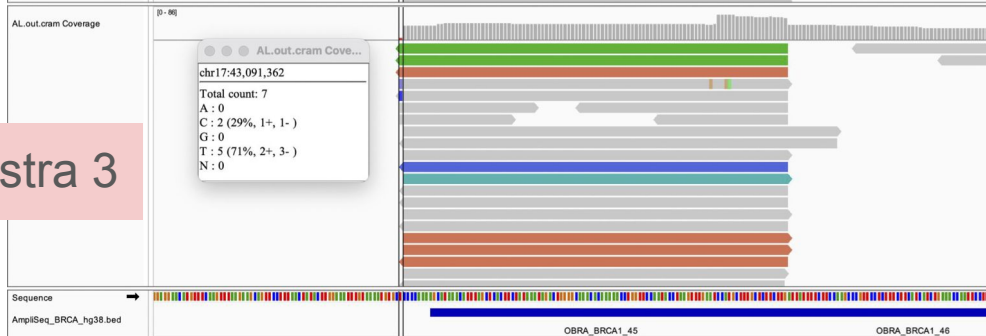
Muestra 1



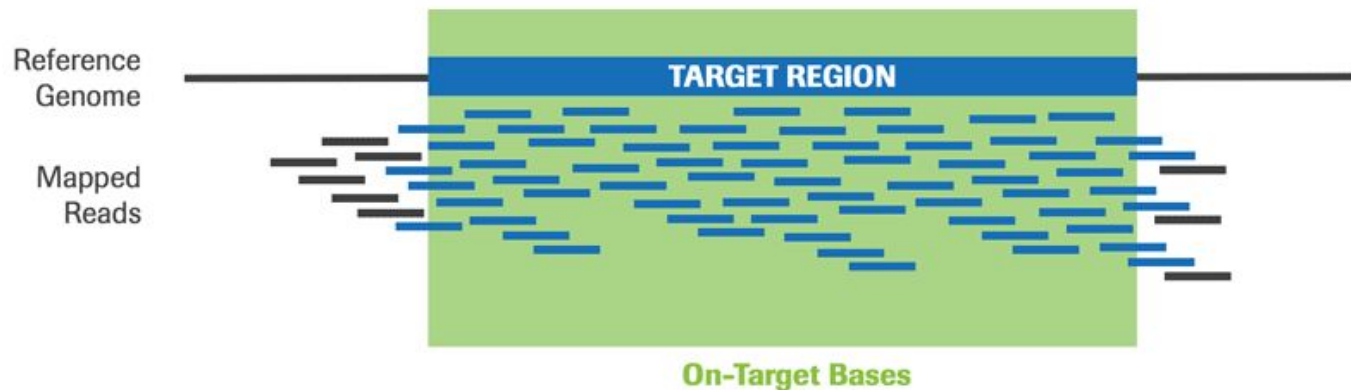
Muestra 2



Muestra 3



# 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

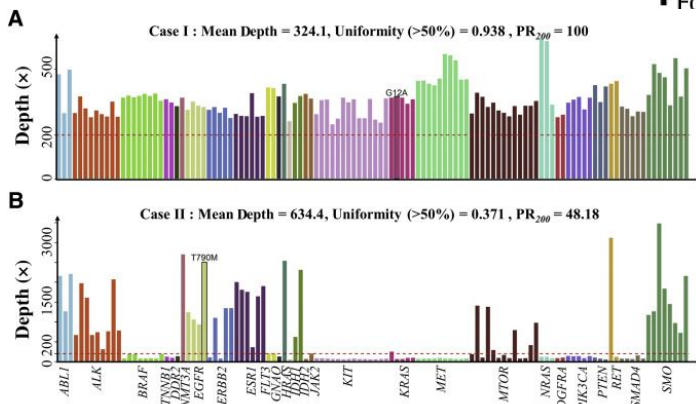


- On-target rate = 100%
- All target regions have exactly the desired coverage
- Fold-80 = 1

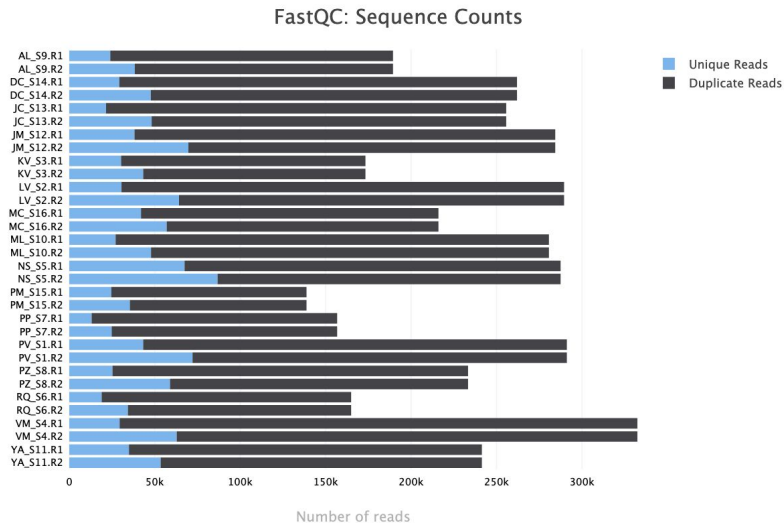
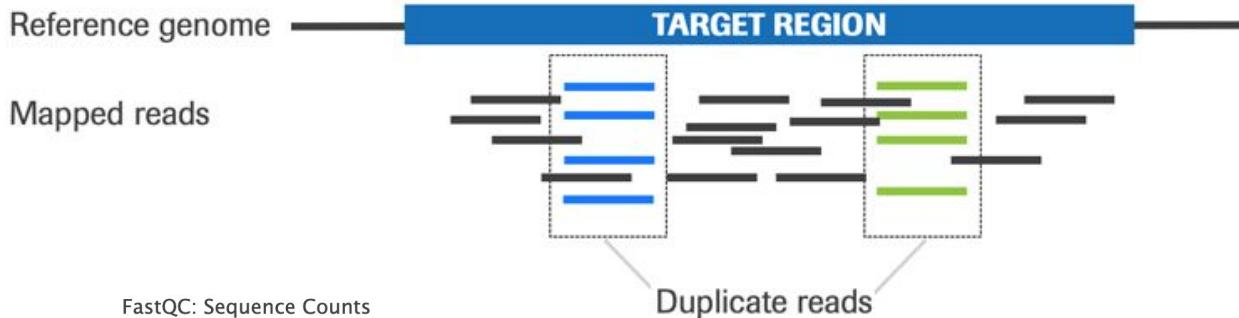
## Observed Uniformity



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

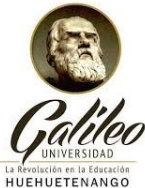



# Número de lecturas Duplicadas



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

# Reporte HTML: MultiQC



  
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



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

 Copy table

 Configure columns

 Scatter plot

 Violin plot

Showing <sup>80</sup>/<sub>100</sub> rows and <sup>10</sup>/<sub>22</sub> columns.

[Export as CSV](#)

| Sample Name | % GC | Ins. size | ≥ 30X   | Median cov | Mean cov  | % Aligned | Reads mapped | % Dups | % GC | M Seqs |
|-------------|------|-----------|---------|------------|-----------|-----------|--------------|--------|------|--------|
| AL          |      |           |         |            |           |           | 0.4 M        |        |      |        |
| AL.md       |      |           |         |            | 1 525.3 X | 99.5 %    |              |        |      |        |
| AL.qualimap | 38 % | 102       | 100.0 % | 1 344 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       |      |           |         |            | 2 163.8 X | 99.5 %    |              |        |      |        |
| DC.qualimap | 38 % | 102       | 100.0 % | 1 942 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       |      |           |         |            | 2 180.6 X | 99.4 %    |              |        |      |        |
| JC.qualimap | 38 % | 103       | 100.0 % | 1 956 X    |           |           |              |        |      |        |
| JC_S13.R1   |      |           |         |            |           |           |              | 91.6 % | 38 % | 0.3 M  |
| JC_S13.R2   |      |           |         |            |           |           |              | 81.1 % | 38 % | 0.3 M  |

## QualiMap