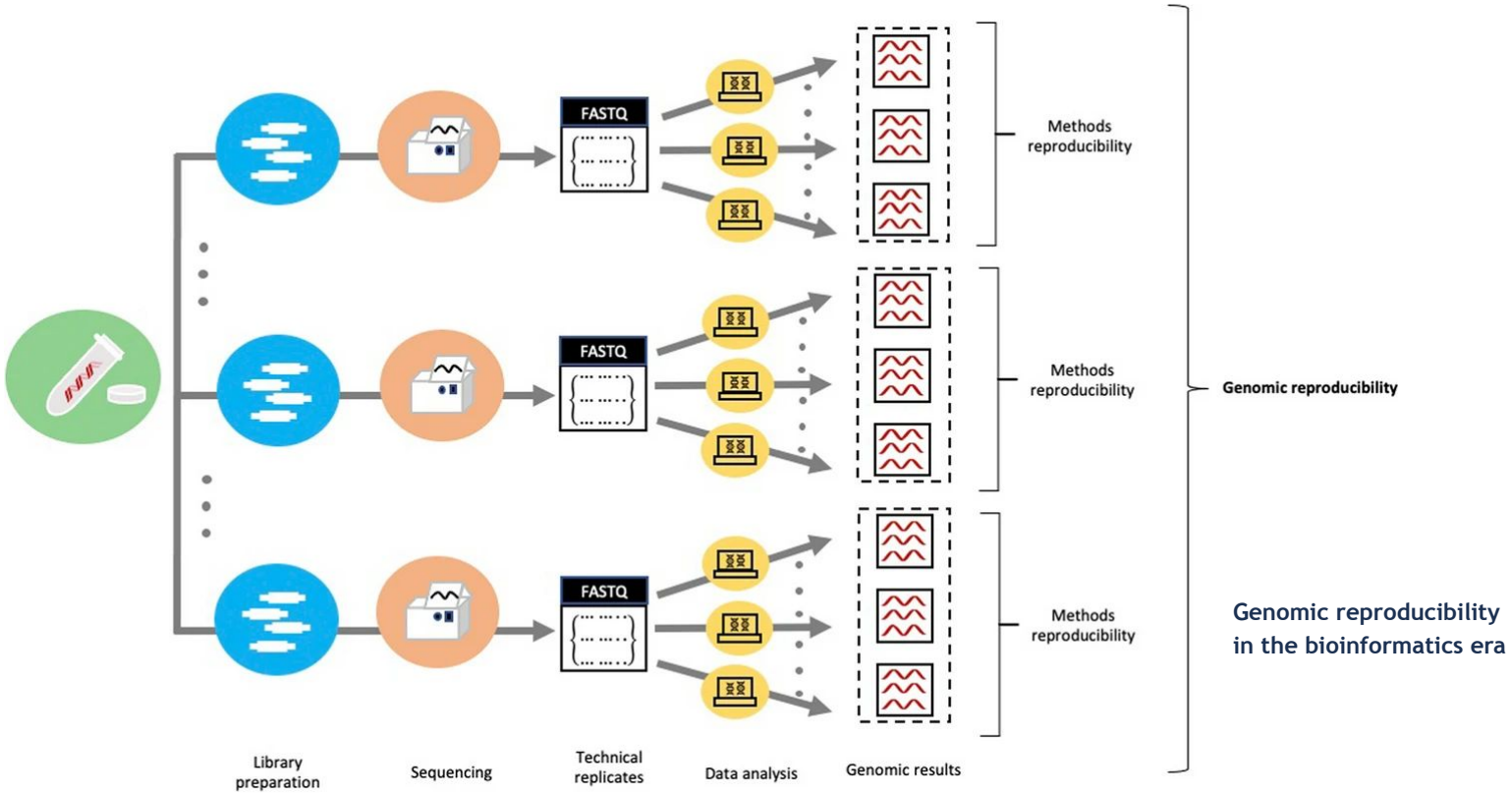
Netflow y automatización de pipeline de análisis, uso de pipeline y reproducibilidad.

Evelin González F.
evelyn.gonzalez@uoh.cl

Organización de las clases: Parte 1

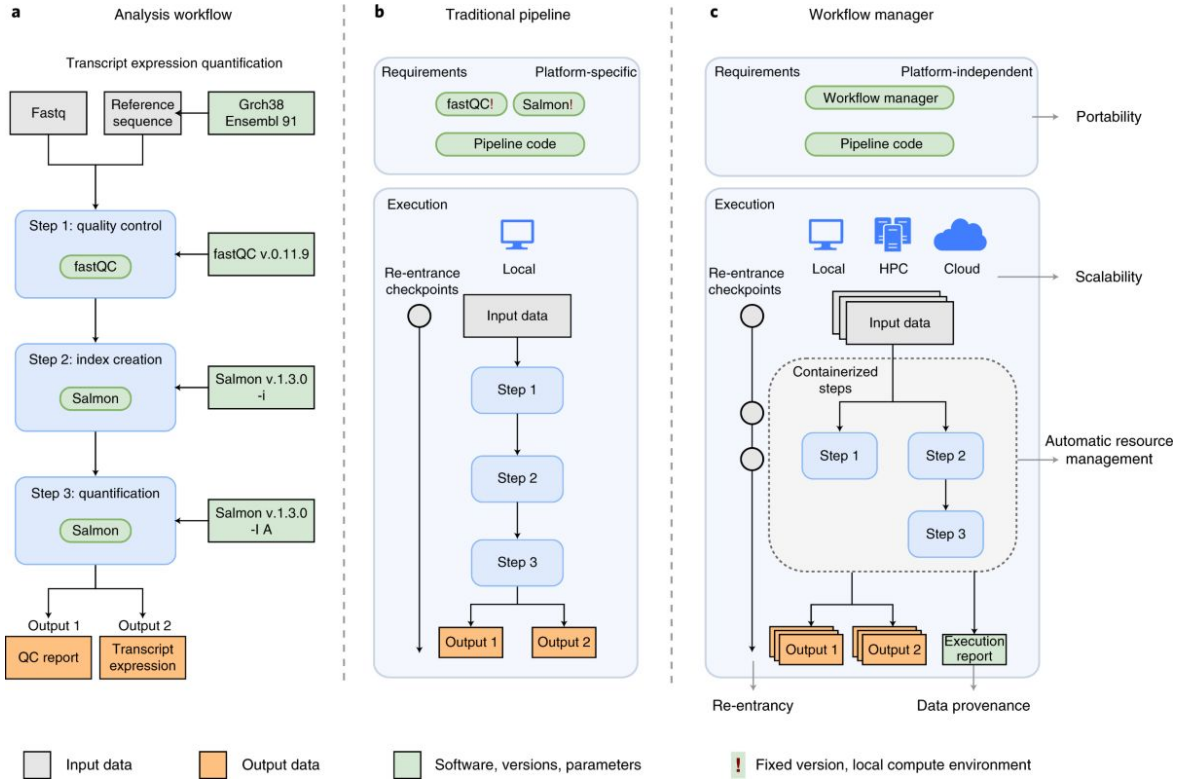
Clase 1: Preprocesamiento de los datos y reportes de calidad.
Clase 2: Llamado de variantes y visualización con IGV
Clase 3: Anotación de variantes (uso de bases de datos y clasificación de variantes patogénicas).
Clase 4: Netflow y automatización de pipeline de análisis, uso de pipeline y reproducibilidad. Taller Práctico.

Reproducibilidad genómica en bioinformática



Genomic reproducibility in the bioinformatics era

Flujos de trabajos



NEXTFLOW

Un motor de flujos de trabajo para la automatización y escalabilidad de análisis bioinformáticos.

Principales características:

Integración con **contenedores** (Docker, Singularity).

Escalabilidad en **HPC** y la nube (AWS, Azure, Google Cloud).

Reproducibilidad y portabilidad de pipelines.

Beneficios en bioinformática:

- Optimización de análisis complejos.
- Trazabilidad y manejo de datos masivos.
- Reducción del tiempo de ejecución mediante paralelización.

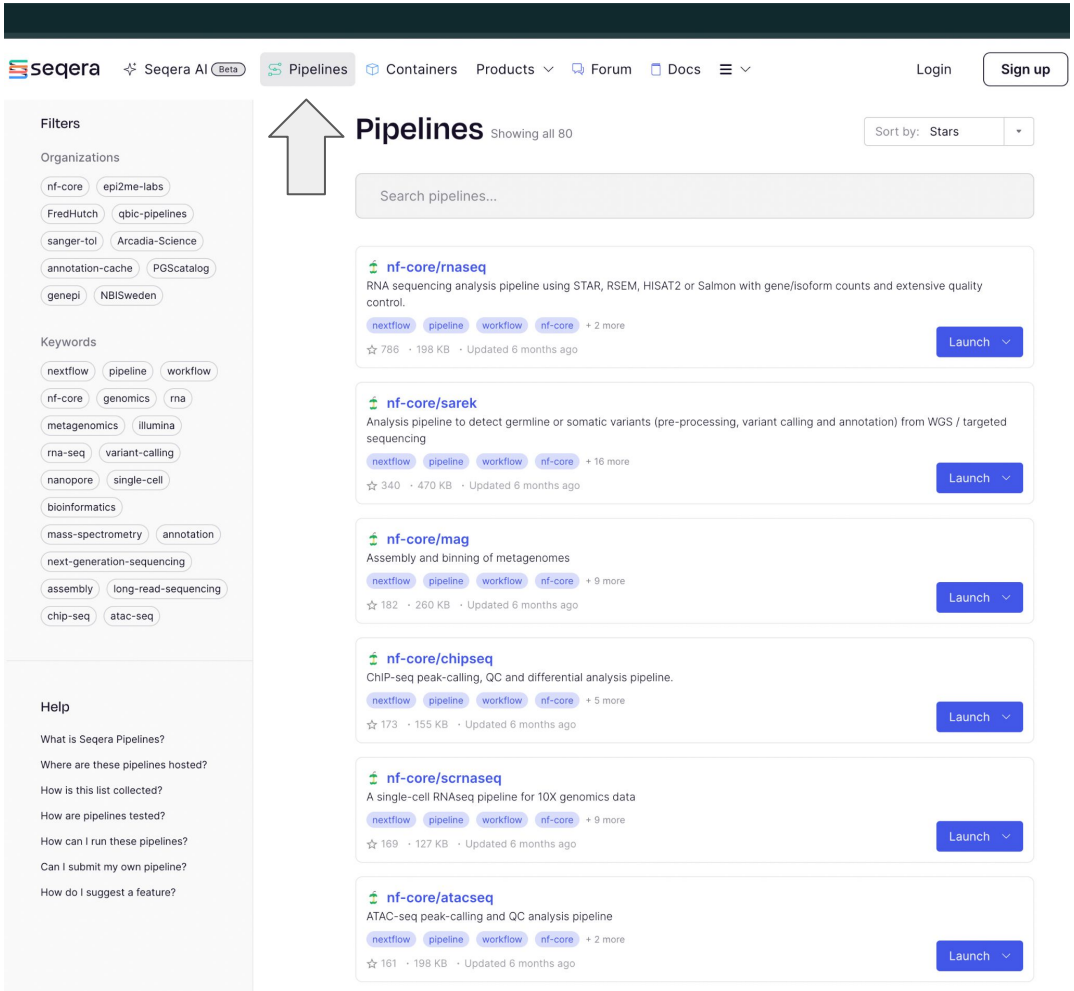
Gestores de flujos de trabajo

Table 1 | Overview of workflow managers for bioinformatics (top, editable version; bottom, image version)

Tool	Class	Ease of use[a]	Expressiveness[b]	Portability[c]	Scalability[d]	Learning resources[e]	Pipeline initiatives[f]
Galaxy	Graphical	●●●	●○○	●●●	●●●	●●●	●●○
KNIME	Graphical	●●●	●○○	○○○	●●◐	●●●	●●○
Nextflow	DSL	●●○	●●●	●●●	●●●	●●●	●●●
Snakemake	DSL	●●○	●●●	●●◐	●●●	●●○	●●●
GenPipes	DSL	●●○	●●●	●●○	●●○	●●○	●●○
bPipe	DSL	●●○	●●●	●●○	●●◐	●●○	●○○
Pachyderm	DSL	●●○	●●●	●○○	●●○	●●●	○○○
SciPipe	Library	●●○	●●●	○○○	○○○	●●○	○○○
Luigi	Library	●●○	●●●	●○○	●●◐	●●○	○○○
Cromwell + WDL	Execution + workflow specification	●○○	●●○	●●●	●●◐	●●○	●●○
cwltool + CWL	Execution + workflow specification	●○○	●●○	●●◐	○○○	●●●	●●○
Toil + CWL/ WDL/Python	Execution + workflow specification	●○○	●●●	●◐○	●●●	●●○	●●○

Please refer to Supplementary Table 1 for details. This information is based on online documentation and manuscripts and may not be reflective of the current state of the projects. Scores for Galaxy are based on the graphical user interface. [a]Ease of use: graphical interface with execution environment (score of 3), programming interface with in-built execution environment (score of 2), separated development and execution environment (score of 1). [b]Expressiveness: based on an existing programming language (3) or a new language or restricted vocabulary (2), primary interaction with graphical user interface (1). [c]Portability: integration with three or more container and package manager platforms (3), two platforms are supported (2), one platform is supported (1). [d]Scalability: considers cloud support, scheduler and orchestration tool integration, and executor support. Please refer to Supplementary Table 1. [e]Learning resources: official tutorials, forums, and events (3), tutorials and forums (2), tutorials or forums (1). [f]Pipelines initiatives: community and curated (3), community or curated (2), not community or curated (1).

Pipelines de análisis



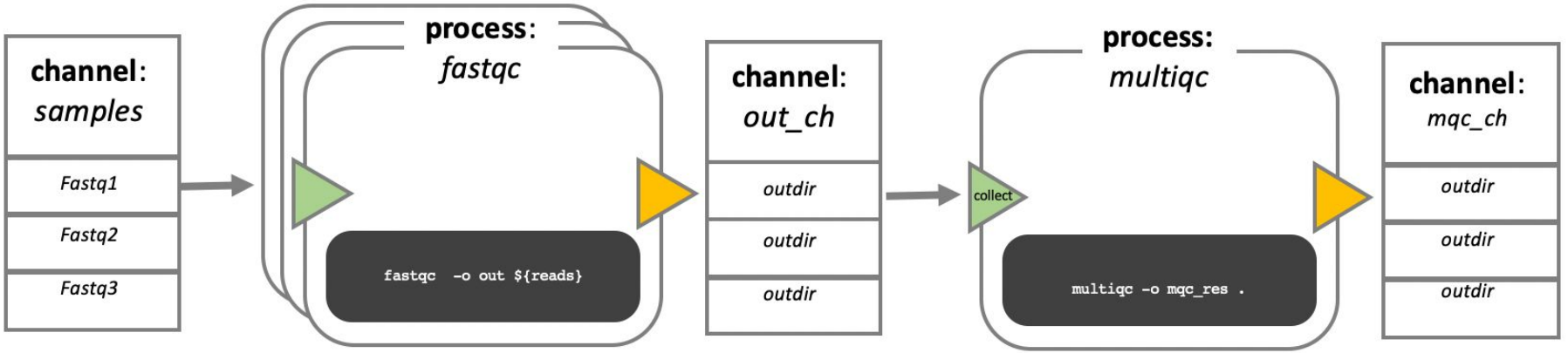
Seqera es una compañía que desarrolla soluciones para la gestión y escalabilidad de flujos de trabajo bioinformáticos, siendo la creadora de **Nextflow**.

https://seqera.io/pipelines/

¿Qué es Nextflow?

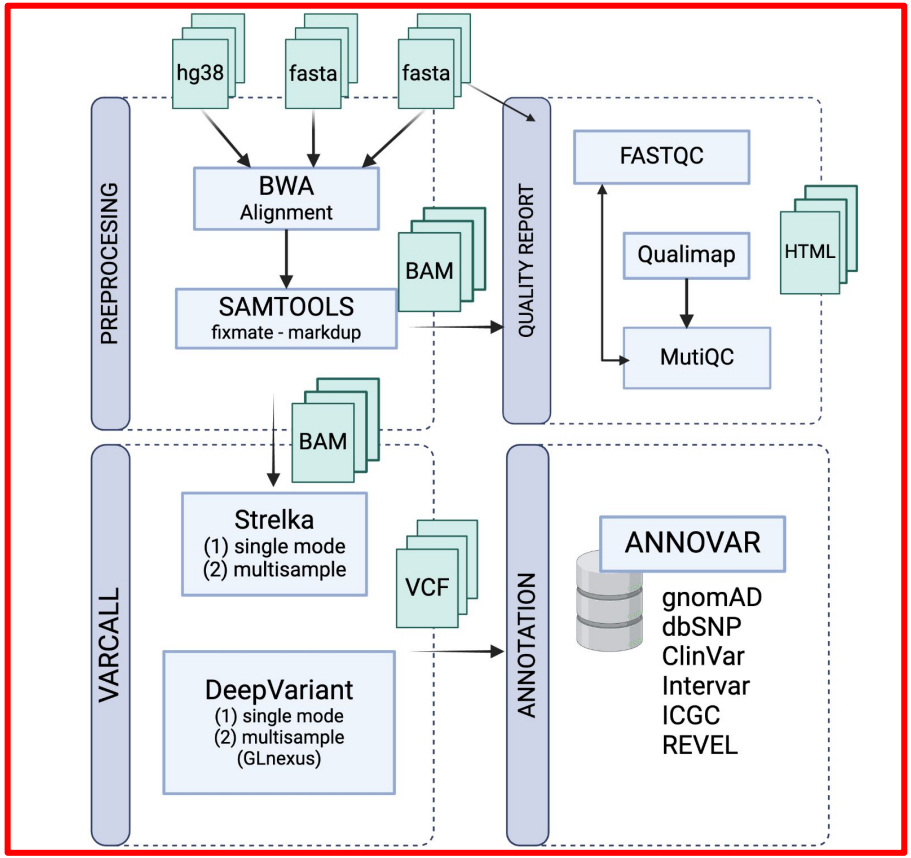
- Herramienta para automatizar flujos de trabajo en bioinformática.
- Diseñada para procesar grandes volúmenes de datos.
- Compatible con múltiples entornos: local, clústeres HPC, nube.
- Lenguaje declarativo sencillo, basado en scripts.
- Facilita la reproducibilidad y el control de versiones.
- Ejecuta tareas en paralelo de manera eficiente.

Procesos en Nextflow



Un **proceso** en Nextflow es la **unidad básica de ejecución** dentro de un pipeline. Representa un paso específico de análisis que se ejecuta de manera independiente, con entradas definidas, un script que realiza la tarea y salidas generadas.

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

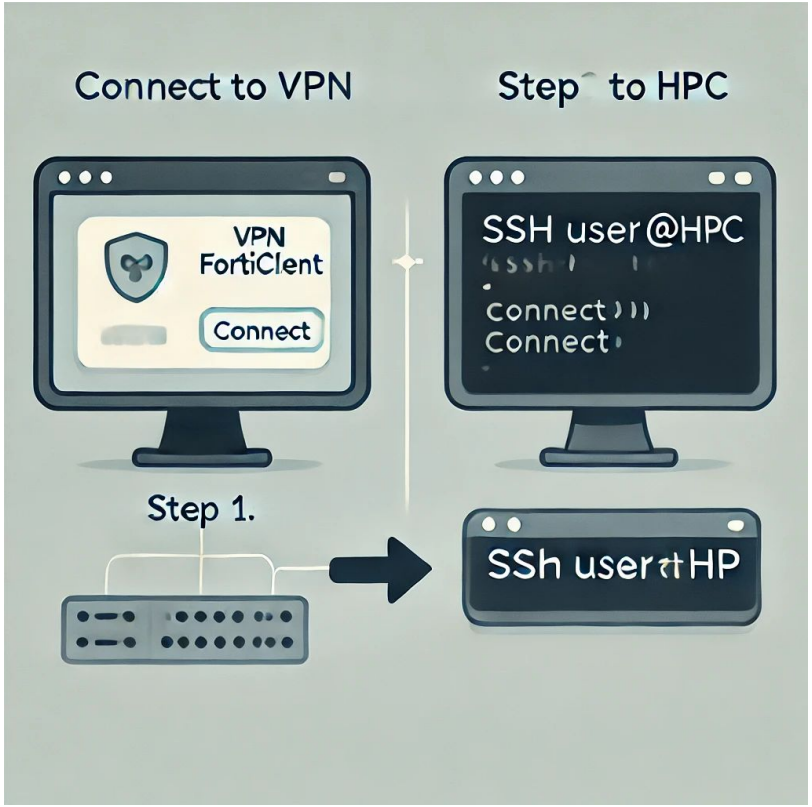


Ejecución de todo el flujo de trabajo de manera automática en un solo paso.

Interpretación de resultados y procesamiento de datos

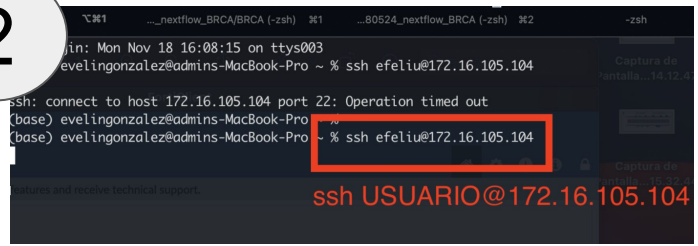
The workflow, including the configurations and tools, is publicly available on the GitHub repository: https://github.com/digenoma-lab/BRCA.

Primer Paso :Conexión a cluster HPC



1

2



Ejecutar pipeline en Nextflow

```
nextflow run main.nf \
-c nextflow.config \
-profile kutral \
-params-file ../params-brca.yml \
--csv ../readsHRR_1-4.csv \
-resume
```

Ejecutar pipeline en Nextflow de BRCA1

```
(brca12) [efeliu@host04 BRCA]$ nextflow run main.nf -c nextflow.config -profile kutral -params-file ../params-brca.yml --csv ../readsHRR_1-4.csv -resum
e
curl: (6) Could not resolve host: www.nextflow.io
N E X T F L O W  ~  version 23.10.1
Launching `main.nf` [cranky_euclid] DSL2 - revision: 4b9c1c4390
[a4/2c84e0] process > PRINT_VERSIONS    [100%] 1 of 1, cached: 1 
[-        ] process > FASTQC            [  0%] 0 of 3
[-        ] process > BWAMEM            [  0%] 0 of 3
[-        ] process > MERGEB            -
[-        ] process > SAMTOOLS          -
[-        ] process > QUALIMAP          -
[-        ] process > B2C               -
[-        ] process > STRELKA_ONESAMPLE -
[-        ] process > STRELKA_POOL      -
[-        ] process > BCFTOOLS_FILTER   -
[-        ] process > BF                -
[-        ] process > ANNOVAR_SP        -
[-        ] process > ANNOVAR_SS        -
```

Ejecutar pipeline en Nextflow

```
nextflow run main.nf \
-c nextflow.config \
-profile kutral \
-params-file ../params-brca.yml \
--csv ../readsHRR_1-4.csv \
-resume
```

nextfow run main.nf

BODY: Inputs Workflow

```
workflow {

    // merge bam files
    MERGEB(greads)

    //ONT aligment
    MINIMAP2_ALIGN(MERGEB.out.mbam
                   ch_ref_fasta,
                   true,
                   'bai',
                   false,
                   false)

    //nanoplot - nanostat
    NANOPLOT(MINIMAP2_ALIGN.out.bam)


    //germline varcall to log reads, ONT
    CLAIR3(MINIMAP2_ALIGN.out.bam,
           MINIMAP2_ALIGN.out.index,
           ch_ref_fasta,
           ch_ref_index)

    //
    LONGPHASE(CLAIR3.out.vcf,
              MINIMAP2_ALIGN.out.bam,
              MINIMAP2_ALIGN.out.index,
              ch_ref_fasta,
              ch_ref_index)
    //
    METHYLARTIST_SEG(LONGPHASE.out.phasebam,
                     LONGPHASE.out.phaseindex,
                     params.bed,
                     ch_ref_fasta,
                     ch_ref_index)
    //DSS
    METHYLARTIST_WGMETH(LONGPHASE.out.phasebam,
                     LONGPHASE.out.phaseindex,
                     ch_ref_fasta,
                     ch_ref_index)

    }
```

HEAD: MÓDULOS

```
#!/usr/bin/env nextflow
nextflow.enable.dsl = 2

//loading scripts from modules
include {MERGEB} from './modules/mergeb'
include {MINIMAP2_ALIGN} from './modules/minimap2'
include {CLAIR3} from './modules/clair3'
include {NANOPLOT} from './modules/nanoplot'
include {LONGPHASE} from './modules/longphase'
include {METHYLARTIST_SEG} from './modules/methylartist'
include {METHYLARTIST_WGMETH} from './modules/methylartist_wgmeth'

process PRINT_VERSIONS {
    publishDir "$params.outdir/software", mode: "copy"

    output:
    path("versions.txt")
    """
    echo "samtools: 1.19" > versions.txt
    echo "minimap2: 2.28-r1209" >> versions.txt
    echo "clair3: 1.0.10" >> versions.txt
    echo "nanoplot: 1.41.6" > versions.txt
    echo "longphase: 1.5.1" >> versions.txt
    echo "methylartist: 1.3.0" >> versions.txt
    """
}
```

Ejecutar pipeline en Nextflow

```
nextflow run main.nf \
-c nextflow.config \
-profile kutral \
-params-file ../params-brca.yml \
--csv ../readsHRR_1-4.csv \
-resume
```

-c nextflow.config -profile kutral

```
process {
    withName: 'PRINT_VERSIONS()'{
    cpus = 1
    memory = '1 GB'
   }
    withName: 'MERGEB' {
        cpus = 10
        memory = '20 GB'
   }
    withName: 'MINIMAP2_ALIGN' {
        cpus = 20
    memory = '40 GB'
    }
}
```

```
profiles {

kutral {
        singularity.enabled    = true
        singularity.autoMounts = true
        docker.enabled         = false
        podman.enabled         = false
        shifter.enabled        = false
        charliecloud.enabled   = false
        executor.queueSize = 8
    process.executor = 'slurm'
    process.queue = 'uohhm'
    }
}
```

```
params {
debug = false


}
```

`-c nextflow.config`:
Especifica un archivo de configuración adicional, en este caso `nextflow.config`. Este archivo puede contener configuraciones como:

- Recursos (CPU, memoria).
- Parámetros globales del pipeline.
- Configuraciones específicas para perfiles y entornos.

Ejecutar pipeline en Nextflow

```
nextflow run main.nf \
-c nextflow.config \
-profile kutral \
-params-file ../params-brca.yml \
--csv ../readsHRR_1-4.csv \
-resume
```

-params-file [ARCHIVO_PARAMETROS.YML]

120924_Nextflow_meth > ! methont-param-ont.yml

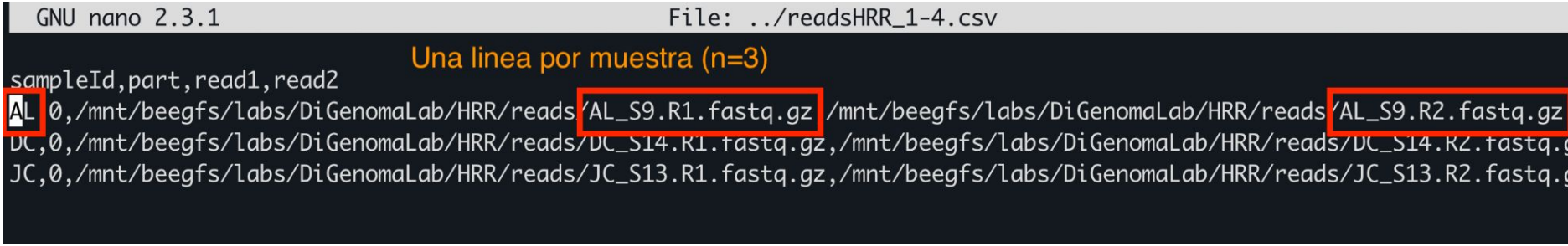
```yaml
ref: /mnt/beegfs/labs/DiGenomaLab/Prunus_dulcis/references/rosaceae/GCA_021292205.2_OSU_Pdul_2.5_genomic.fa
minimap2: "-ax  map-ont"
outdir: ont_result
aln_only: true
bed: /mnt/beegfs/home/efeliu/work2024/120924_Nextflow_meth/protein_coding_prunusDulcis.bed

```

`-params-file`:
Permite cargar un archivo de parámetros en formato YAML para configurar el pipeline. Este archivo contiene valores específicos para las variables que se utilizan en el pipeline, facilitando la personalización sin modificar el código.

Ejecutar pipeline en Nextflow

```
nextflow run main.nf \
-c nextflow.config \
-profile kutral \
-params-file ../params-brca.yml \
--csv ../readsHRR_1-4.csv \
-resume
```

–csv [ARCHIVO_ENTRADA]

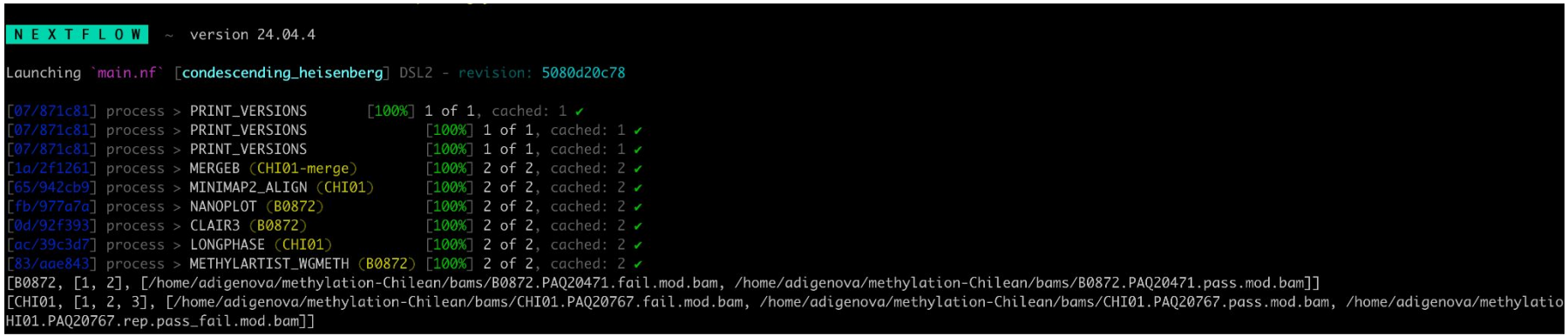


Archivo de entrada

Ejecutar pipeline en Nextflow

```
nextflow run main.nf \
-c nextflow.config \
-profile kutral \
-params-file ../params-brca.yml \
--csv ../readsHRR_1-4.csv \
-resume
```

-resume



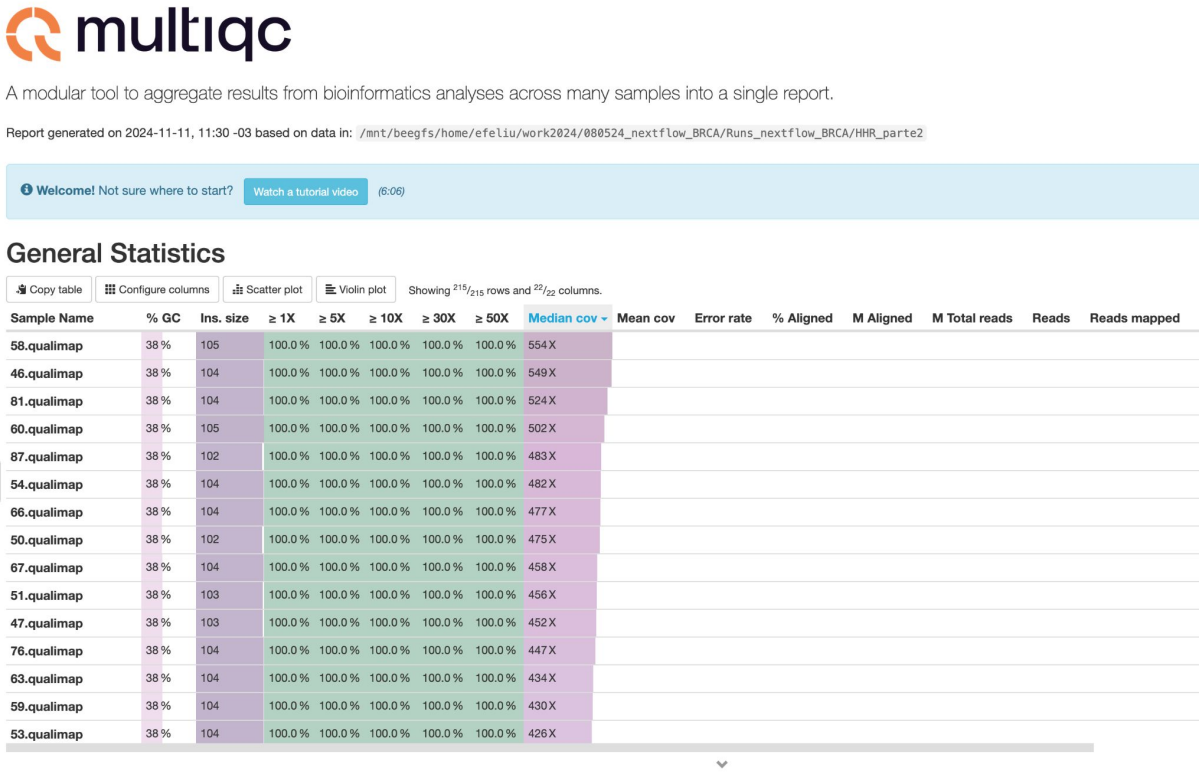
`-resume`:

Indica que Nextflow debe **reutilizar los resultados previos** de ejecuciones anteriores. Esto es útil si la ejecución se interrumpe o si se desean realizar pequeños ajustes sin reanalizar todos los pasos.Nextflow verifica los pasos ya completados y reanuda desde donde se quedó, ahorrando tiempo y recursos.

Archivos de salida: BRCA pipeline

- Reporte de calidad: archivo HTML con parámetros de calidad. MultiQC.
- Archivos BAM (uno para cada muestra)
- Archivos VCF (uno para cada muestra)
- Anotación de variantes - ANNOVAR (uno para cada muestra)

QC, datos de calidad



multiqc

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

Report generated on 2024-11-11, 11:30 -03 based on data in: `/mnt/beegfs/home/efeliu/work2024/080524_nextflow_BRCA/Runs_nextflow_BRCA/HHR_parte2`

Welcome! Not sure where to start? Watch a tutorial video *(6:06)*

General Statistics

Copy table | Configure columns | Scatter plot | Violin plot | Showing 215/215 rows and 22/22 columns.

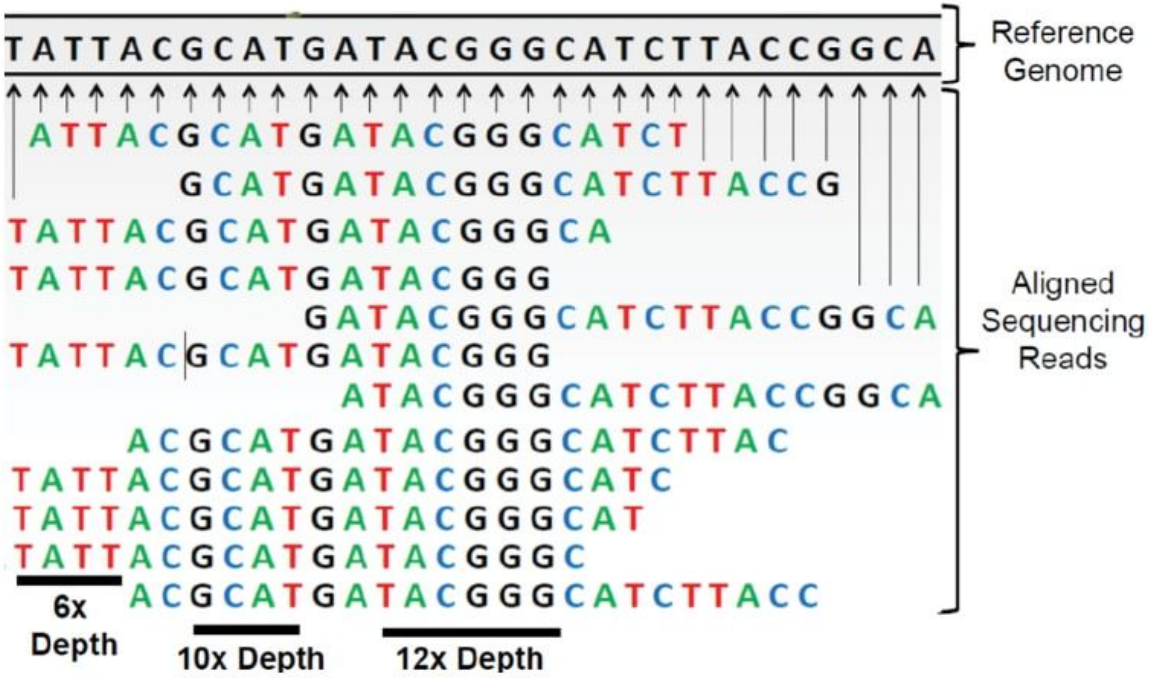
Sample Name	% GC	Ins. size	≥ 1X	≥ 5X	≥ 10X	≥ 30X	≥ 50X	Median cov	Mean cov	Error rate	% Aligned	M Aligned	M Total reads	Reads	Reads mapped
58.qualimap	38 %	105	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	554 X							
46.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	549 X							
81.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	524 X							
60.qualimap	38 %	105	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	502 X							
87.qualimap	38 %	102	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	483 X							
54.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	482 X							
66.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	477 X							
50.qualimap	38 %	102	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	475 X							
67.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	458 X							
51.qualimap	38 %	103	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	456 X							
47.qualimap	38 %	103	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	452 X							
76.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	447 X							
63.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	434 X							
59.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	430 X							
53.qualimap	38 %	104	100.0 %	100.0 %	100.0 %	100.0 %	100.0 %	426 X							

QualiMap

QualiMap is a platform-independent application to facilitate the quality control of alignment sequencing data and its derivatives like feature counts. *DOI: 10.1093/bioinformatics/btv566; 10.1093/bioinformatics/bts503.*

FastQC
Qualimap
MultiQC

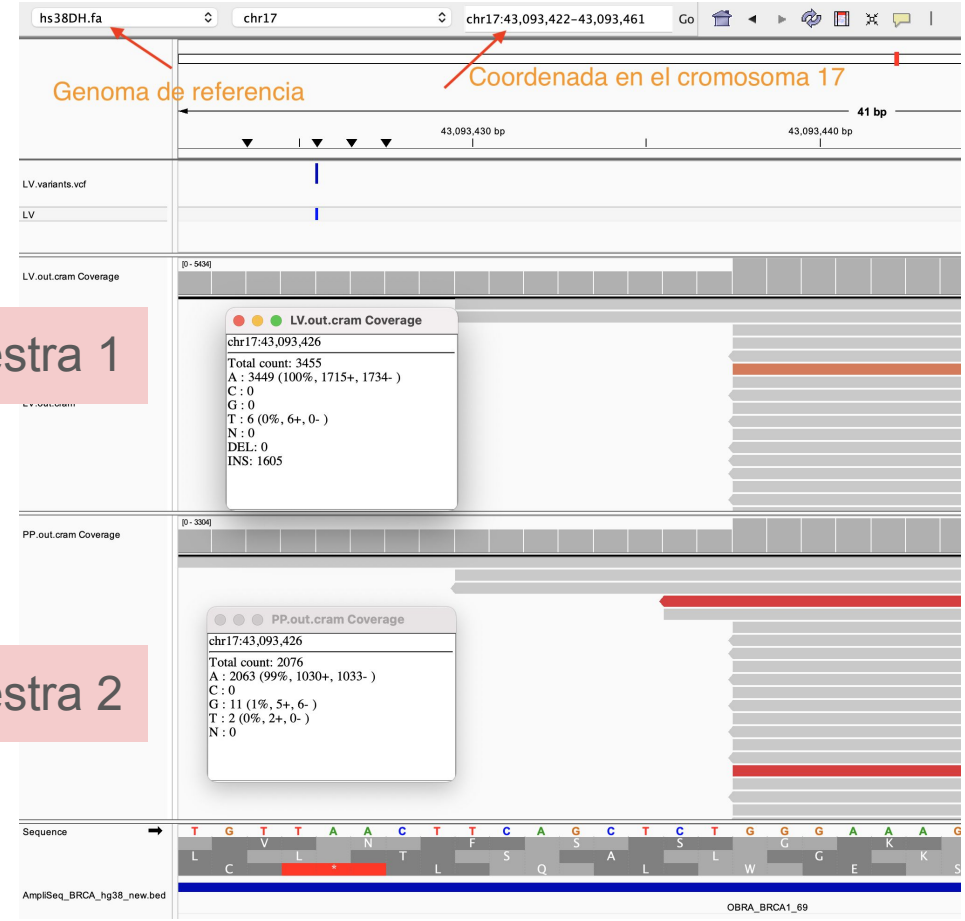
Archivos BAM: Lecturas alineadas al genoma



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.



Muestra 1

Muestra 2

Ejemplo: Variante frameshift insertion p.L655Ffs*10 *BRCA1*

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

Variant Call Format: VCF

A **VCF example**

Header

```
##fileformat=VCFv4.1
##fileDate=20110413
##source=VCFtools
##reference=file:///refs/human_NCBI36.fasta
##contig=<ID=1,length=249250621,md5=1b22b98cdeb4a9304cb5d48026a85128,species="Homo Sapiens">
##contig=<ID=X,length=155270560,md5=7e0e2e580297b7764e31dbc80c2540dd,species="Homo Sapiens">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	40	PASS	.	GT:DP	1/1:13	2/2:29
1	2	.	C	T,CT	.	PASS	H2;AA=T	GT	0|1	2/2
1	5	rs12	A	G	67	PASS	.	GT:DP	1|0:16	2/2:20
X	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=299	GT:GQ:DP	1:12:.	0/0:20:36

B **SNP**

Alignment

```
1234
ACGT
ATGT
 ^
```

VCF representation

POS	REF	ALT
2	C	T

C **Insertion**

```
12345
AC-GT
ACTGT
  ^
```

POS	REF	ALT
2	C	CT

D **Deletion**

```
1234
ACGT
A--T
 ^^
```

POS	REF	ALT
1	ACG	A

E **Replacement**

```
1234
ACGT
A-TT
 ^^
```

POS	REF	ALT
1	ACG	AT

100 results per page

back to HOME

Page: 1

Chr	Start	End	Ref	Alt	Func	Gene	GeneDetail	ExonicFunc	AAChange
chr13	32332343	32332343	A	C	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon10:c.A865C:p.N289H
chr13	32332592	32332592	A	C	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon10:c.A1114C:p.N372H
chr13	32332843	32332843	A	G	exonic	BRCA2		synonymous SNV	BRCA2:NM_000059.4:exon10:c.A1365G:p.S455S
chr13	32336584	32336584	T	C	exonic	BRCA2		synonymous SNV	BRCA2:NM_000059.4:exon11:c.T2229C:p.H743H
chr13	32336744	32336744	A	C	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon11:c.A2389C:p.K797Q
chr13	32337326	32337326	A	G	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon11:c.A2971G:p.N991D
chr13	32337751	32337751	A	G	exonic	BRCA2		synonymous SNV	BRCA2:NM_000059.4:exon11:c.A3396G:p.K1132K
chr13	32338162	32338162	T	C	exonic	BRCA2		synonymous SNV	BRCA2:NM_000059.4:exon11:c.T3807C:p.V1269V
chr13	32338918	32338918	A	G	exonic	BRCA2		synonymous SNV	BRCA2:NM_000059.4:exon11:c.A4563G:p.L1521L
chr13	32340099	32340099	C	T	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon11:c.C5744T:p.T1915M
chr13	32340678	32340678	G	A	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon11:c.G6323A:p.R2108H
chr13	32340868	32340868	G	C	exonic	BRCA2		synonymous SNV	BRCA2:NM_000059.4:exon11:c.G6513C:p.V2171V
chr13	32355095	32355095	A	G	exonic	BRCA2		synonymous SNV	BRCA2:NM_000059.4:exon14:c.A7242G:p.S2414S
chr13	32355250	32355250	T	C	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon14:c.T7397C:p.V2466A
chr13	32356461	32356461	T	C	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon15:c.T7469C:p.I2490T
chr13	32379413	32379413	G	A	exonic	BRCA2		nonsynonymous SNV	BRCA2:NM_000059.4:exon22:c.G8851A:p.A2951T
chr17	43071077	43071077	T	C	exonic	BRCA1		nonsynonymous SNV	BRCA1:NM_007297.4:exon14:c.A4696G:p.S1566G,E
chr17	43082453	43082453	A	G	exonic	BRCA1		synonymous SNV	BRCA1:NM_007297.4:exon11:c.T4167C:p.S1389S,E
chr17	43091983	43091983	T	C	exonic	BRCA1		nonsynonymous SNV	BRCA1:NM_007297.4:exon9:c.A3407G:p.K1136R,BR
chr17	43092412	43092412	C	T	exonic	BRCA1		nonsynonymous SNV	BRCA1:NM_007297.4:exon9:c.G2978A:p.S993N,BRC
chr17	43092418	43092418	T	C	exonic	BRCA1		nonsynonymous SNV	BRCA1:NM_007297.4:exon9:c.A2972G:p.E991G,BRC
chr17	43092919	43092919	G	A	exonic	BRCA1		nonsynonymous SNV	BRCA1:NM_007297.4:exon9:c.C2471T:p.P824L,BRC
chr17	43093220	43093220	A	G	exonic	BRCA1		synonymous SNV	BRCA1:NM_007297.4:exon9:c.T2170C:p.L724L,BRC
chr17	43093425	43093425	-	A	exonic	BRCA1		frameshift insertion	BRCA1:NM_007297.4:exon9:c.1964dupT:p.L655Ffs
chr17	43093449	43093449	G	A	exonic	BRCA1		synonymous SNV	BRCA1:NM_007297.4:exon9:c.C1941T:p.S647S,BRC
chr17	43093454	43093454	C	T	exonic	BRCA1		nonsynonymous SNV	BRCA1:NM_007297.4:exon9:c.G1936A:p.D646N,BRC

Page: 1

Sort by:

Buscar variantes “Pathogenic” en CLINVAR

Filter by:

CLNALLELEID: CLNDN: = CLNSIG: = athogenic

SIFT_pred: SIFT4G_pred: Polyphen2_HVAR_score:

LRT_pred: MutationTaster_pred: FATHMM_score:

PROVEAN_pred: MetaSVM_score: MetaLR_score:

MetaLR_pred: Reliability_index:

Chr:

Start:

End:

Gene:

CLNALLELEID:

VCF anotados con ANNOVAR

Filtrado de Variantes para la Identificación de:

- **Variantes Patogénicas**
- **VUS (Variantes de Significado Incierto)**
- **Variantes Novel**
- **Variantes Benignas**

Repositorio del pipeline BRCA en GitHub