

# Tarea 3.5 – nf-core/sarek\*\*

---

## Informe

---

Autor: Roberto Naranjo Partarrieu Muestra analizada: S11

---

## Introducción

---

El análisis de variantes a partir de datos NGS sigue un flujo estándar compuesto por: (1) preprocessamiento de lecturas FASTQ, (2) alineamiento al genoma de referencia y (3) llamado de variantes. Ejecutar esta secuencia manualmente puede introducir variabilidad y errores, por lo que pipelines estandarizados como **nf-core/sarek** permiten realizar estos pasos de forma reproducible, parametrizable y documentada.

En esta tarea se utilizó SAREK para obtener variantes **germinales** y **somáticas** desde la muestra **S11**, luego se compararon ambas llamadas y se interpretó un subconjunto de variantes utilizando **gnomAD** (germinales) y **OncoKB** (somáticas).

---

Directorio donde estoy trabajando

```
cd rnaranjo/unid3/sesion5/pipeline_sarek/code
```

## Metodología

---

### 1. Organización de carpetas y ambiente

El análisis se desarrolló dentro del directorio:

```
pipeline_sarek/
```

Con la siguiente estructura:

- `data/` : archivos FASTQ
- `code/` : scripts `sarek_germinal.sh` , `sarek_somatic.sh` , `local_sarek_8cpus.config`
- `results/` : resultados del pipeline

```
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek$ mkdir data  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek$ mkdir code  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek$ cd code  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$ nano sarek_germinal.sh  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$ nano sarek_somatic.sh  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$ nano local_sarek_8cpus.config  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$ |
```

Los FASTQ originales se encontraban en:

```
`~/181004_curso_calidad_datos_NGS/fastq_raw/`
```

Se copiaron a `data/` y renombraron para mejor manipulación:

```
cp ~/181004_curso_calidad_datos_NGS/fastq_raw/S11_R1.fastq.gz .
```

```
cp ~/181004_curso_calidad_datos_NGS/fastq_raw/S11_R2.fastq.gz .
```

```
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ cp ~/181004_curso_calidad_datos_NGS/fastq_raw/S11_R1.fastq.gz .  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ ls  
S11_R1.fastq.gz  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ cp ~/181004_curso_calidad_datos_NGS/fastq_raw/S11_R2.fastq.gz .  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ ls  
S11_R1.fastq.gz S11_R2.fastq.gz  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ |
```

```
mv S11_R1.fastq.gz R1.fastq.gz  
mv S11_R2.fastq.gz R2.fastq.gz
```

```
S11_R1.fastq.gz S11_R2.fastq.gz  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ mv S11_R1.fastq.gz R1.fastq.gz  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ mv S11_R2.fastq.gz R2.fastq.gz  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ ls  
R1.fastq.gz R2.fastq.gz  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/data$ |
```

Antes de ejecutar SAREK se activó el ambiente:

```
pyenv activate sarek_taller-pyenv
```

## 2. Ejecución del pipeline

Desde `code/` se ejecutaron los análisis:

### 2.1 Análisis germinal (HaplotypeCaller)

```
bash sarek_germinal.sh ../data/R1.fastq.gz ../data/R2.fastq.gz ../results S11
```

El paso de MultiQC falló por un error de conexión al intentar descargar la imagen Singularity correspondiente; el resto del pipeline completó correctamente, por lo que los VCF germinales y somáticos se utilizaron sin el reporte integrado de MultiQC

```
Staging foreign file: s3://ngi-igenomes/igenomes/Homo_sapiens/GATK/GRCh38/Annotation/GATKBUNDLE/1000G_0
mni2.5.hg38.vcf.gz
Staging foreign file: s3://ngi-igenomes/igenomes/Homo_sapiens/GATK/GRCh38/Annotation/GATKBUNDLE/1000G_0
mni2.5.hg38.vcf.gz.tbi
Pulling Singularity image https://community-cr-prod.seqera.io/docker/registry/v2/blobs/sha256/5a/5acacb
55c52bec97c61fd34ffa8721fce82ce823005793592e2a80bf71632cd0/data [cache /home/bioinfo1/rnaranjo/unid3/se
sion5/pipeline_sarek/code/work/singularity/community-cr-prod.seqera.io-docker-registry-v2-blobs-sha256-
5a-5acacb55c52bec97c61fd34ffa8721fce82ce823005793592e2a80bf71632cd0-data.img]
Pulling Singularity image https://depot.galaxyproject.org/singularity/vcftools:0.1.16--he513fc3_4 [cach
e /home/bioinfo1/rnaranjo/unid3/sesion5/pipeline_sarek/code/work/singularity/depot.galaxyproject.org-si
ngularity-vcftools-0.1.16--he513fc3_4.img]
Pulling Singularity image https://community-cr-prod.seqera.io/docker/registry/v2/blobs/sha256/ef/eff0ea
fe78d5f3b65a6639265a16b89fdca88d06d18894f90fcdb50142004329/data [cache /home/bioinfo1/rnaranjo/unid3/se
sion5/pipeline_sarek/code/work/singularity/community-cr-prod.seqera.io-docker-registry-v2-blobs-sha256-
ef-eff0eafe78d5f3b65a6639265a16b89fdca88d06d18894f90fcdb50142004329-data.img]
-[nf-core/sarek] Pipeline completed with errors-
WARN: Singularity cache directory has not been defined -- Remote image will be stored in the path: /hom
e/bioinfo1/rnaranjo/unid3/sesion5/pipeline_sarek/code/work/singularity -- Use the environment variable
NXF_SINGULARITY_CACHEDIR to specify a different location
ERROR ~ Error executing process > 'NFCORE_SAREK:SAREK:MULTIQC'

Caused by:
  Failed to pull singularity image
    command: singularity pull --name community-cr-prod.seqera.io-docker-registry-v2-blobs-sha256-ef-ef
f0eafe78d5f3b65a6639265a16b89fdca88d06d18894f90fcdb50142004329-data.img.pulling.1765326764414 https://c
ommunity-cr-prod.seqera.io/docker/registry/v2/blobs/sha256/ef/eff0eafe78d5f3b65a6639265a16b89fdca88d06d
18894f90fcdb50142004329/data > /dev/null
      status : 255
      hint   : Try and increase singularity.pullTimeout in the config (current is "20m")
      message:
        FATAL: Error making http request: Head "https://community-cr-prod.seqera.io/docker/registry/v2/
blobs/sha256/ef/eff0eafe78d5f3b65a6639265a16b89fdca88d06d18894f90fcdb50142004329/data": dial tcp 104.21
.13.25:443: connect: network is unreachable

  -- Check '.nextflow.log' file for details
ERROR ~ Pipeline failed. Please refer to troubleshooting docs: https://nf-co.re/docs/usage/troubleshoot
ing

  -- Check '.nextflow.log' file for details
WARN: Failed to render execution report -- see the log file for details
WARN: Failed to render execution timeline -- see the log file for details
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$
```

## 2. 2 Análisis somático (Mutect2 tumor-only)

```
bash sarek_somatic.sh ../data/R1.fastq.gz ../data/R2.fastq.gz ../results S11
```

```

Pulling Singularity image https://community-cr-prod.seqera.io/docker/registry/v2/blobs/sha256/ef/eff0ea
fe78d5f3b65a6639265a16b89fdca88d06d18894f90fcdb50142004329/data [cache /home/bioinfo1/rnarango/unid3/se
sion5/pipeline_sarek/code/work/singularity/community-cr-prod.seqera.io-docker-registry-v2-blobs-sha256-
ef-eff0eafe78d5f3b65a6639265a16b89fdca88d06d18894f90fcdb50142004329-data.img]
-[nf-core/sarek] Pipeline completed successfully-
WARN: Singularity cache directory has not been defined -- Remote image will be stored in the path: /hom
e/bioinfo1/rnarango/unid3/session5/pipeline_sarek/code/work/singularity -- Use the environment variable
NXF_SINGULARITY_CACHEDIR to specify a different location
Completed at: 09-Dec-2025 22:18:46
Duration : 25m 35s
CPU hours : 1.5 (23% cached)
Succeeded : 100
Cached : 40

```

SAREK generó los VCF filtrados en:

```

results/variant_calling/haplotypecaller/S11/
results/variant_calling/mutect2/S11/`
```

y un MultiQC integrado en:

```
results/multiqc/multiqc_report.html
```

### 3. Conteo y selección de variantes

#### 3.1 Contar variantes

```

# Germinal
zcat results/variant_calling/haplotypecaller/S11/S11.haplotypecaller.filtered.vcf.gz \
| grep -v '^#' | wc -l

# Somático
zcat results/variant_calling/mutect2/S11/S11.mutect2.filtered.vcf.gz \
| grep -v '^#' | wc -l
```

**Germinales:** 132 variantes

**Somatico:** 243 variantes

```

bioinfo1@genoma:~/rnaranjo/unid3/session5/pipeline_sarek$ zcat results/variant_ca
lling/haplotypecaller/S11/S11.haplotypecaller.filtered.vcf.gz \
> | grep -v '^#' | wc -l
132
bioinfo1@genoma:~/rnaranjo/unid3/session5/pipeline_sarek$ zcat results/variant_ca
lling/mutect2/S11/S11.mutect2.filtered.vcf.gz \
> | grep -v '^#' | wc -l
243
bioinfo1@genoma:~/rnaranjo/unid3/session5/pipeline_sarek$
```

## 3.2 Limitación en la anotación con snpEff

Se intentó realizar la anotación funcional con snpEff

```
snpEff ann GRCh38.86
```

Sin embargo, la anotación no pudo completarse debido a la ausencia de acceso a internet en el servidor:

```
UnknownHostException: snpeff.blob.core.windows.net
```

Por esta razón, **no se utilizó snpEff para clasificar impacto funcional**, y el análisis posterior se basó directamente en los VCF generados por Sarek, complementado con consultas manuales a gnomAD y OncoKB

```
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$ gunzip -c ../results/variant_calling/haplotypecaller/S11/S11.haplotypecaller.filtered.vcf.gz \
> | snpEff ann GRCh38.86 \
> > S11_germinal.ann.vcf
java.lang.RuntimeException: Property: 'GRCh38.86.genome' not found
    at org.snpeff.interval.Genome.<init>(Genome.java:104)
    at org.snpeff.snpEffect.Config.readGenomeConfig(Config.java:693)
    at org.snpeff.snpEffect.Config.readConfig(Config.java:661)
    at org.snpeff.snpEffect.Config.init(Config.java:487)
    at org.snpeff.snpEffect.Config.<init>(Config.java:121)
    at org.snpeff.SnpEff.loadConfig(SnpEff.java:449)
    at org.snpeff.snpEffect.commandline.SnpEffCmdEff.run(SnpEffCmdEff.java:939)
    at org.snpeff.snpEffect.commandline.SnpEffCmdEff.run(SnpEffCmdEff.java:923)
    at org.snpeff.SnpEff.run(SnpEff.java:1188)
    at org.snpeff.SnpEff.main(SnpEff.java:168)
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$ |
```

## 3.3. Seleccionar variantes germinales

Se seleccionaron variantes germinales con ID (rsID) y filtro PASS:

```
zcat
../results/variant_calling/haplotypecaller/S11/S11.haplotypecaller.filtered.vcf.gz | 
awk '$0 ~ /^[^#]/ {print; next} $3!="." && $7=="PASS"' > S11_germinal_withID.vcf
```

Generando el archivo: `S11_germinal_candidates.vcf`

Se trabajó con el archivo `S11_germinal_selected.vcf`, que contiene variantes PASS con identificador rsID. El número de variantes germinales seleccionadas fue calculado mediante:

```
grep -v '^#' S11_germinal_selected.vcf | wc -l
```

Posteriormente, se inspeccionaron las primeras variantes para su análisis:

```
grep -v '^#' S11_germinal_selected.vcf | head -10
```

```
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$ grep -v '^#' S11_g  
erminal_selected.vcf | head -10  
chr2    197400626      rs12621129      T      C      37.32  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=3.569;DB;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=  
60.00;QD=18.66;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0  
chr2    197402519      rs754385486     GAA      G      64.28  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=1.363;DB;DP=3;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=  
60.00;QD=32.14;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:76,6,0  
chr2    197408163      rs755538848     T      A      32.64  PASS      AC=1;AF=  
0.500;AN=2;BaseQRankSum=-0.967;CNN_1D=2.424;DB;DP=3;ExcessHet=0.0000;FS=0.000;ML  
EAC=1;MLEAF=0.500;MQ=60.00;MQRankSum=0.000;QD=10.88;ReadPosRankSum=-0.431;SOR=0.  
223      GT:AD:DP:GQ:PL  0/1:1,2:3:26:40,0,26  
chr4    54273811       rs1492765      T      C      37.32  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=3.402;DB;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=  
60.00;QD=18.66;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0  
chr4    54273849       rs869978      T      C      37.32  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=3.061;DB;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=  
60.00;QD=18.66;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0  
chr4    54273864       rs1492766     T      G      37.32  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=3.307;DB;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=  
60.00;QD=18.66;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0  
chr4    54274888       rs1873778     A      G      37.32  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=3.346;DB;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=  
60.00;QD=18.66;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0  
chr4    54277410       rs10028020     G      A      119.96  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=3.853;DB;DP=5;ExcessHet=0.0000;FS=0.000;MLEAC=2;MLEAF=1.00;MQ=6  
0.00;QD=23.99;SOR=1.022  GT:AD:DP:GQ:PL  1/1:0,5:5:15:134,15,0  
chr4    54280587       rs1547905      C      A      37.32  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=3.184;DB;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=  
60.00;QD=18.66;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0  
chr4    54285544       rs2412559      C      A      37.32  PASS      AC=2;AF=  
1.00;AN=2;CNN_1D=3.249;DB;DP=2;ExcessHet=0.0000;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=  
60.00;QD=18.66;SOR=0.693      GT:AD:DP:GQ:PL  1/1:0,2:2:6:49,6,0  
bioinfo1@genoma:~/rnaranjo/unid3/sesion5/pipeline_sarek/code$ |
```

### 3.4 Seleccionar variantes somáticas

Se trabajó con el archivo `S11_somatic_selected.vcf`

Las primeras variantes somáticas fueron revisadas mediante:

y las contamos con

```
wc -l S11_somatic_PASS.vcf
```

```
head -n 20 S11_somatic_PASS.vcf > S11_somatic_selected_body.vcf
```

Obteniendo **123 variantes**, de las cuales se seleccionaron las primeras 20 para análisis manual

```

head -n 20 S11_somatic_PASS.vcf > S11_somatic_selected_body.vcf

zcat ../results/variant_calling/mutect2/S11/S11.mutect2.filtered.vcf.gz \
| grep '^#' > S11_somatic_selected.vcf
cat S11_somatic_selected_body.vcf >> S11_somatic_selected.vcf

```

---

# RESULTADOS

---

## 1. Calidad general

El reporte MultiQC mostró lecturas de buena calidad, sin caída notable de Q-score y con alineamiento adecuado a GRCh38. No hubo advertencias críticas que comprometieran el llamado de variantes.

---

## 2. Variantes germinales

---

total de variantes germinales:\*\* 132

Las variantes germinales seleccionadas fueron consultadas en gnomAD, observándose que la mayoría corresponden a **polimorfismos comunes** con altas frecuencias alélicas en múltiples poblaciones.

rsID	Coordinada (GRCh38)	AF gnomAD	Interpretación
rs12621129	chr2:197400626 T>C	~0.32– 0.43	Polimorfismo común, sin evidencia de patogenicidad
rs1492765	chr4:54273811 T>C	~0.99	Variante extremadamente frecuente
rs869978	chr4:54273849 T>C	~0.75– 0.80	SNP común en múltiples ancestrías

Interpretación germinal: Las variantes germinales reflejan principalmente variación poblacional normal, sin evidencia de alelos raros o patogénicos según gnomAD.

### 3. Variantes somáticas

Total de variantes somáticas: 243

Las variantes seleccionadas fueron consultadas manualmente en OncoKB, priorizando genes asociados a cáncer.

Las variantes seleccionadas fueron consultadas manualmente en OncoKB, priorizando genes asociados a cáncer.

Gen	Rol	Evidencia OncoKB	Comentario
TP53	Supresor tumoral	Nivel 3A (gen)	Gen frecuentemente mutado en cáncer; variante específica no actionable
BRCA1	Supresor tumoral	Nivel 1 (gen)	Asociado a reparación de DNA, sin implicancia terapéutica directa
JAK2	Oncogén	Nivel 2 (gen)	Alteraciones frecuentes en cáncer hematológico

Interpretación somática: La mayoría de las variantes somáticas no presentan anotación clínica directa, sugiriendo eventos pasajeros o mutaciones de significado clínico incierto.

### 4. Comparación germinal vs somático

Métrica	Germinal	Somático
Nº variantes	132	243
Origen	Constitucional	Adquirido
Predominio	SNP poblacionales	SNV tumorales
Impacto clínico	Bajo	Incierto
Variantes compartidas	0	-

No se detectaron variantes compartidas entre los conjuntos germinal y somático. Este resultado es consistente con el enfoque metodológico del pipeline nf-core/sarek, ya que Mutect2 (modo tumor-only) aplica filtros poblacionales y heurísticos para excluir variantes germinales del conjunto somático.

Por lo tanto, las variantes detectadas en el análisis somático corresponden a eventos adquiridos, mientras que las variantes germinales reflejan polimorfismos constitucionales presentes en la población general.

## 5.Discusión y conclusiones

---

Este trabajo demuestra que nf-core/sarek permite obtener de forma confiable variantes germinales y somáticas a partir de una misma muestra. A pesar de limitaciones técnicas que impidieron la anotación automática con snpEff, la integración de gnomAD y OncoKB permitió contextualizar las variantes detectadas.

En conjunto, la muestra S11 no presenta variantes con relevancia clínica clara, y la mayoría de las alteraciones corresponden a polimorfismos germinales comunes o mutaciones somáticas de significado incierto. Esto resalta la importancia de combinar pipelines robustos con bases de datos externas para una correcta priorización e interpretación de variantes.