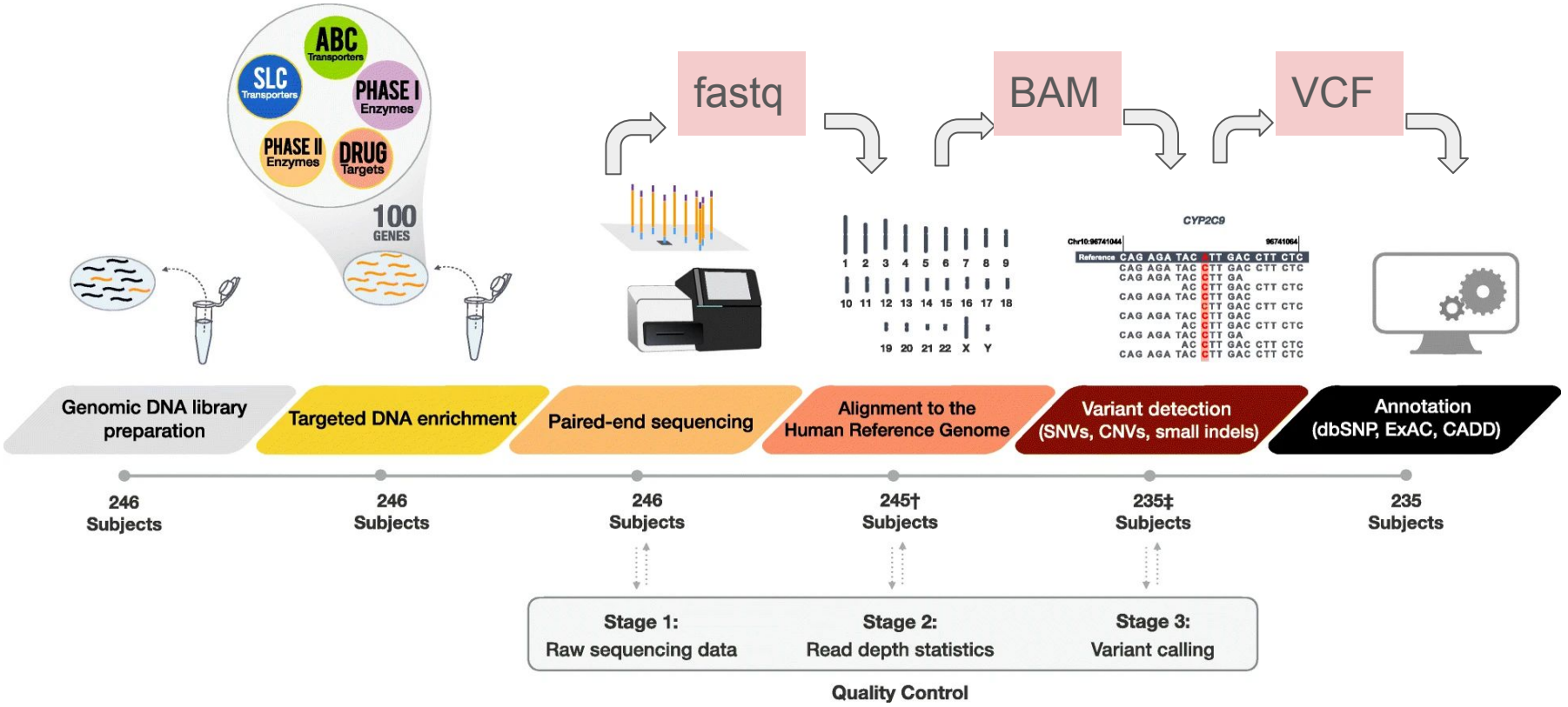
Clase 2: Llamado de variantes y visualización con IGV

Evelin González F.
evelyn.gonzalez@uoh.cl

Organización de las clases: Parte 1

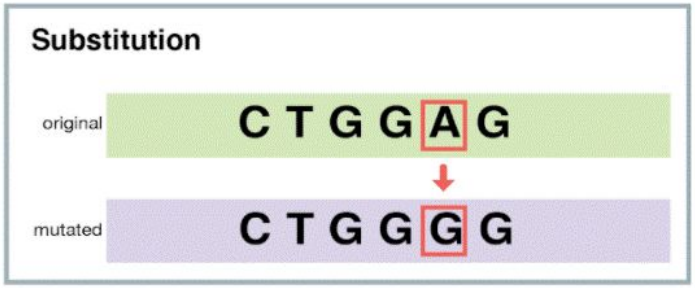
Clase 1: Preprocesamiento de los datos y reportes de calidad.
Clase 2: Llamado de variantes y visualización con IGV
Clase 3: Anotación de variantes (uso de bases de datos y clasificación de variantes patogénicas).
Clase 4: Netflow y automatización de pipeline de análisis, uso de pipeline y reproducibilidad.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

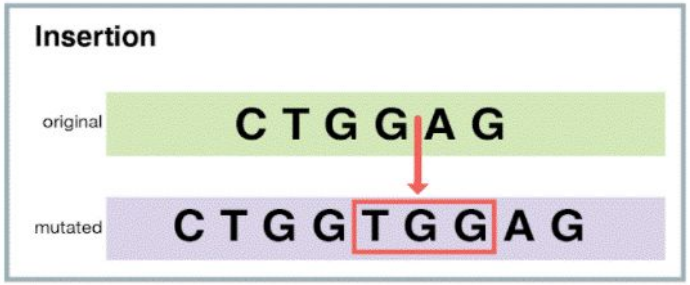
Flujo de trabajo “Targeting Sequencing”

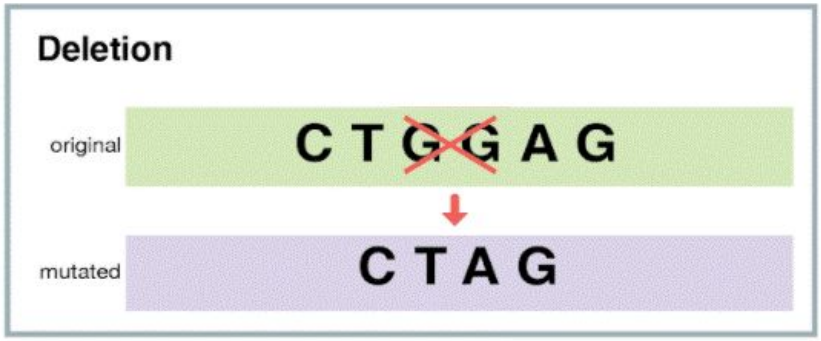


https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-019-0527-2

Concepto de variante genética (SNPs, INDELs)

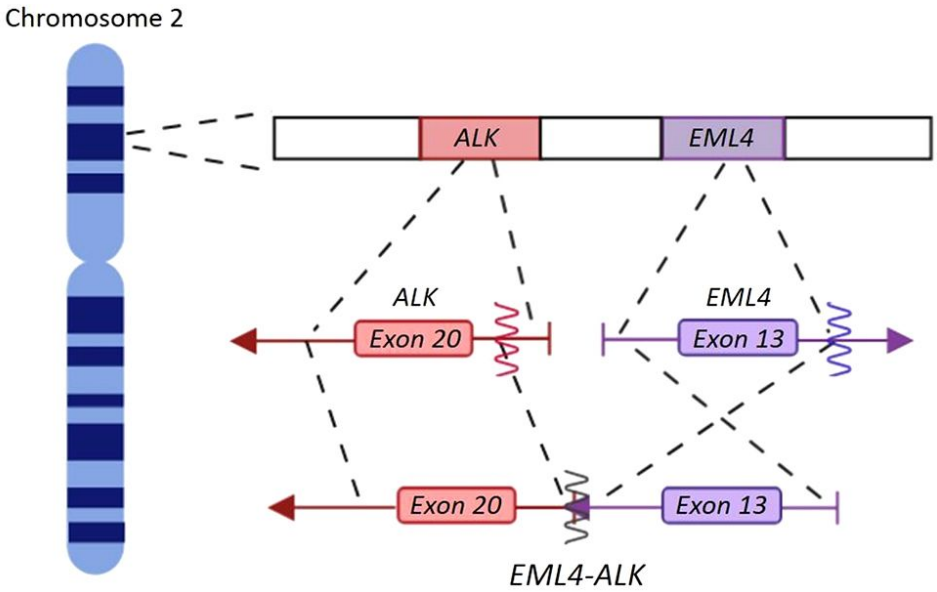






- **Sustitución (SNP)**: Cambio de una base por otra en una posición específica (ej. A → T).
- **Inserción**: Adición de una o más bases en la secuencia de ADN.
- **Deleción**: Pérdida de una o más bases en la secuencia de ADN.

Fusiones en Cáncer



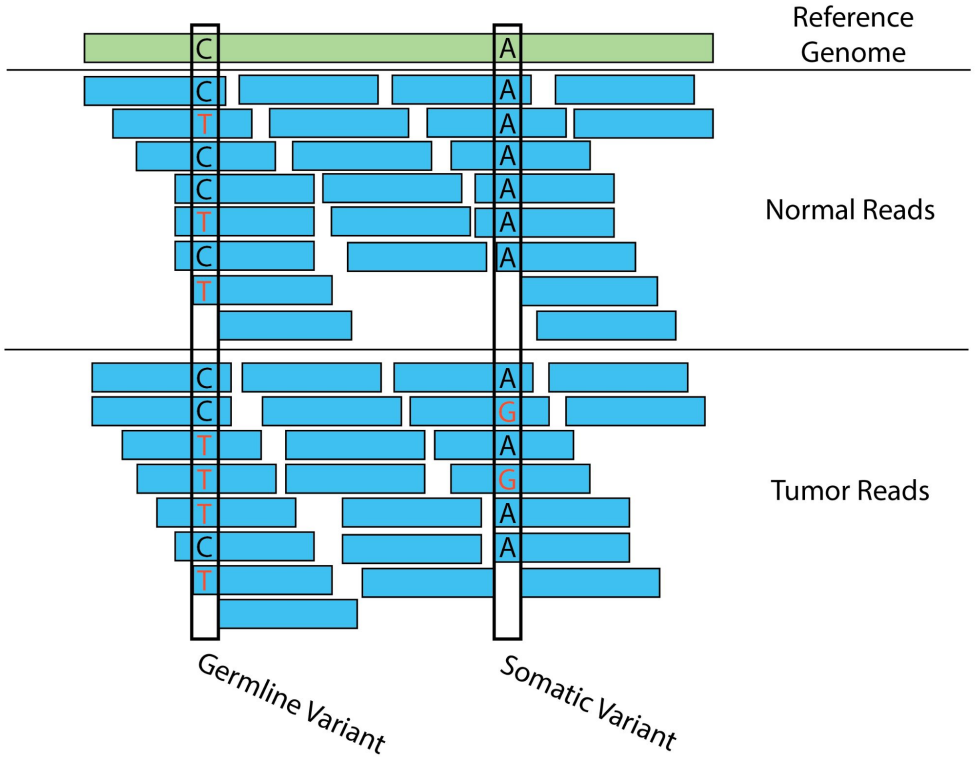
Las fusiones genéticas ocurren cuando dos genes diferentes se combinan para formar un nuevo gen híbrido.
Estas fusiones pueden surgir de:

- **Translocaciones cromosómicas**: intercambio de segmentos entre cromosomas.
- **Inversiones**: cuando un segmento del cromosoma se invierte en su posición.
- **Deleciones**: pérdida de fragmentos de cromosomas que facilitan la unión de genes adyacentes.

Generan proteínas anómalas que pueden activar rutas de señalización descontroladas.
Son frecuentes en varios tipos de cáncer y pueden influir en el pronóstico y tratamiento.
Su detección es posible mediante técnicas de secuenciación de ADN y ARN.

Algunas fusiones son objetivos de **terapias dirigidas**, ayudando a personalizar el tratamiento.

Diferencia en la identificación de variantes somáticas y germinales



Germinales: Incluyen variantes heredadas (SNVs, indels, CNVs, y variantes estructurales), que son típicamente estables y presentes en heterocigosis o homocigosis.

Somáticas: Se detectan variantes específicas del tumor (mutaciones puntuales, fusiones, CNVs, etc.) que no están en las células normales, y suelen variar en frecuencia al estar presentes en subpoblaciones tumorales.

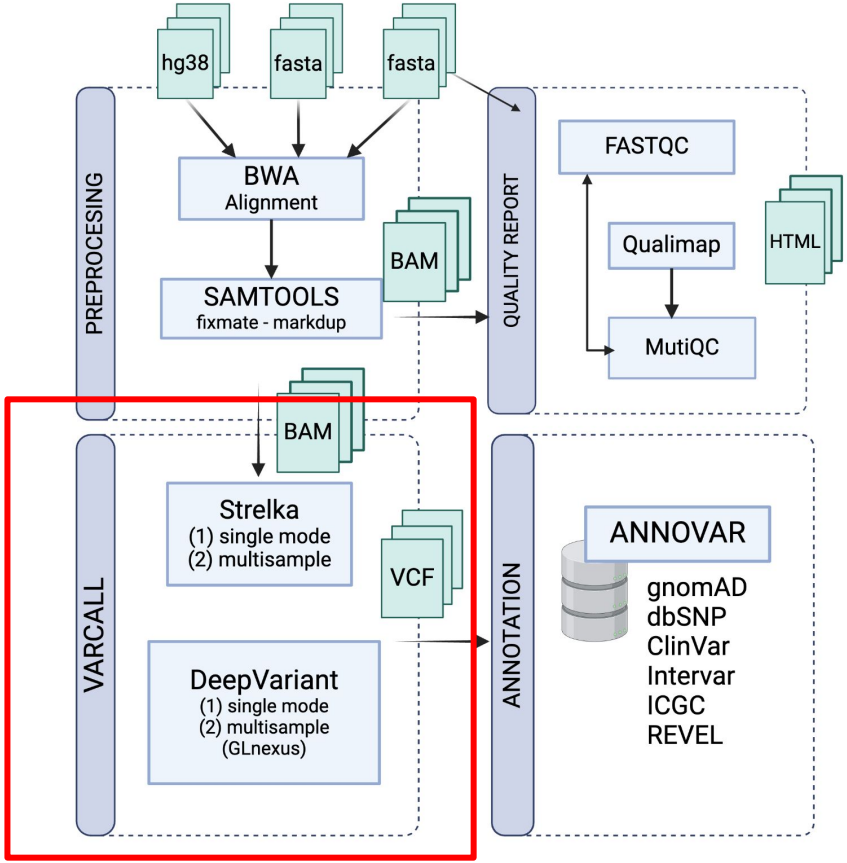
Tool	Approach, method	Application	References
Small variants			
GATK Haplotypecaller	Local reassembly of haplotypes	Germline, MNPs	(Poplin, Ruano-Rubio, et al., 2018)
BCFtools	Positional, pileups	Germline	(Danecek et al., 2021)
FreeBayes	Haplotype-based, Bayesian model	Germline, MNPs	(Garrison & Marth, 2012)
GATK Mutect2	Local reassembly	Somatic	(Cibulskis et al., 2013)
Strelka2	Tiered haplotype model	Germline, somatic	(Kim et al., 2018)
Structural variants			
Delly2	RP, SR, RD	Germline SVs	(Rausch et al., 2012)
Pindel	SR, RP	Germline SVs	(Ye et al., 2018)
Manta	SR, RP, AS	Germline, somatic	(Chen et al., 2016)
GRIDSS2	AS, SV Breakpoint	Somatic	(Cameron et al., 2021)
Varscan2	RD, Circular Binary Segmentation	Exome, somatic, CNVs	(Koboldt et al., 2012)

Abbreviations: AS, assembly; CNV, copy-number variants; GATK, Genome Analysis ToolKit; MEI, mobile element

Llamadores de variantes

Utilizar una herramienta adecuada para el tipo de dato es fundamental en bioinformática, especialmente en el análisis de datos genómicos, debido a varias razones clave.

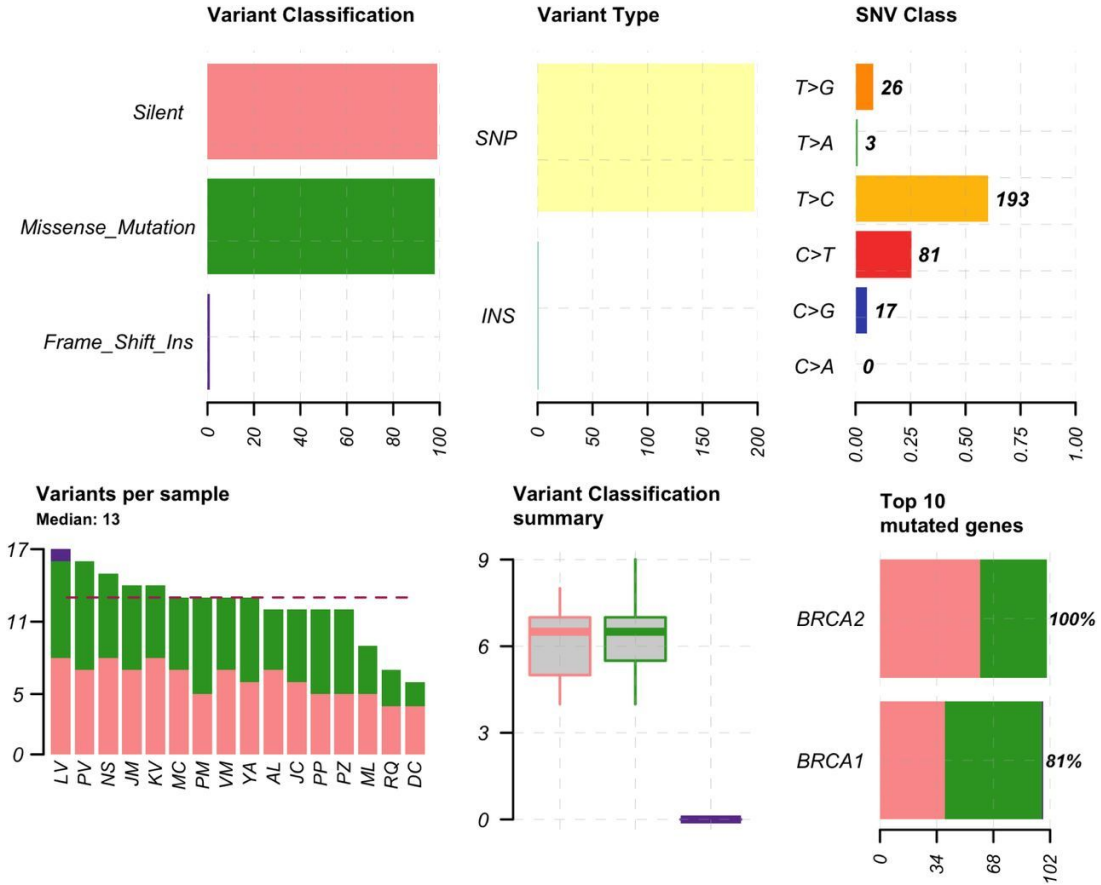
Flujo de trabajo bionformático para la detección de variantes en *BRCA1* y *BRCA2*



Por qué utilizar single y multi-sample mode?

Qué tipos de errores podemos detectar. ?

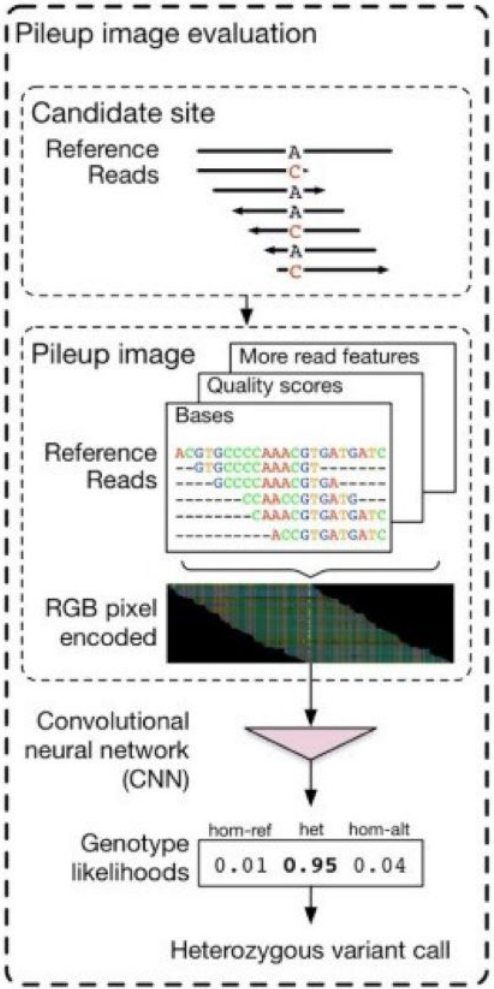
The workflow, including the configurations and tools, is publicly available on the GitHub repository: https://github.com/digenoma-lab/BRCA.



Resumen de mutaciones germinales encontradas en pacientes con cáncer de mama.

Estos paneles presentan la clasificación, el tipo y la distribución de variantes en 16 pacientes con cáncer de mama.

- Se detallan los cambios de bases SNV y se muestra una mediana de 13 variantes por muestra.
- Los genes BRCA2 y BRCA1 presentan tasas de mutación del 100 % y 81 %, respectivamente.



Llamado de variantes: DeepVariant

- Modelo de aprendizaje profundo desarrollado por Google para el llamado de variantes en ADN.
- Diseñado para detectar variantes de alta precisión en secuencias germinales y somáticas.

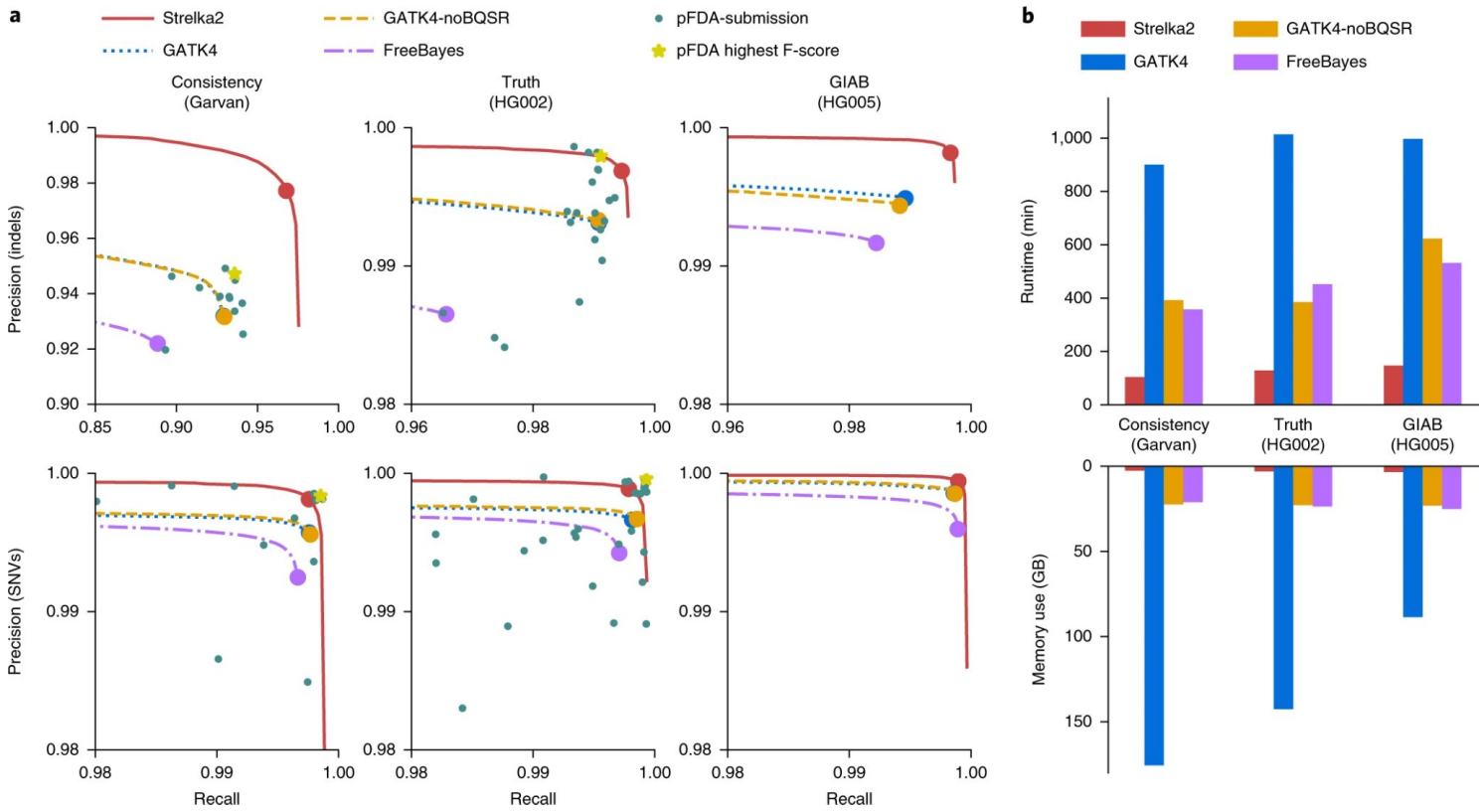
Funcionamiento de DeepVariant

- Convierte datos de secuenciación en imágenes y utiliza una red neuronal convolucional para clasificar cada variante.
- Proporciona una precisión avanzada gracias a la capacidad del modelo para reconocer patrones complejos en los datos.

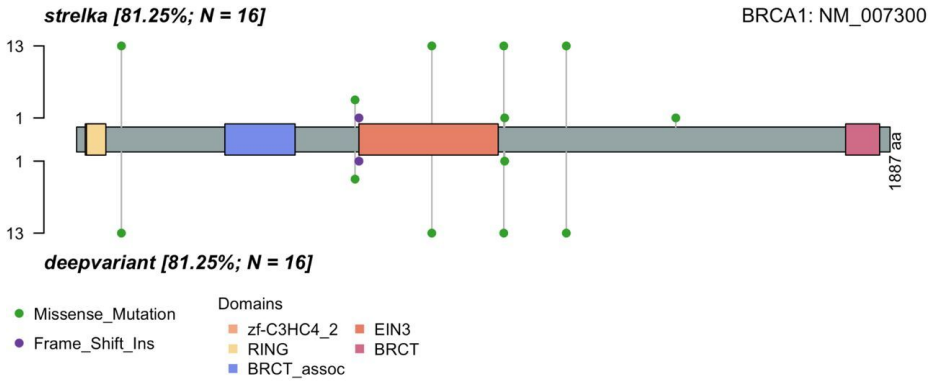
Ventajas para Análisis Germinal

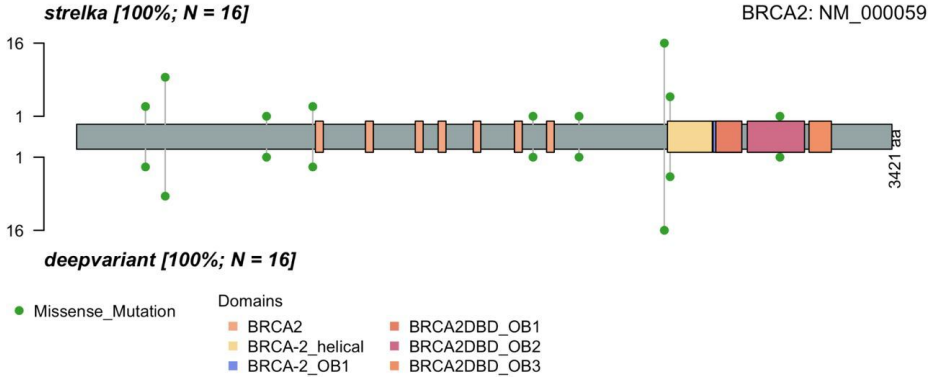
- Alta sensibilidad y especificidad en la detección de SNPs e indels.
- Soporte para múltiples tipos de datos de secuenciación (WGS, WES).
- Compatible con plataformas como Illumina y PacBio.

Llamado de variantes: Strelka2

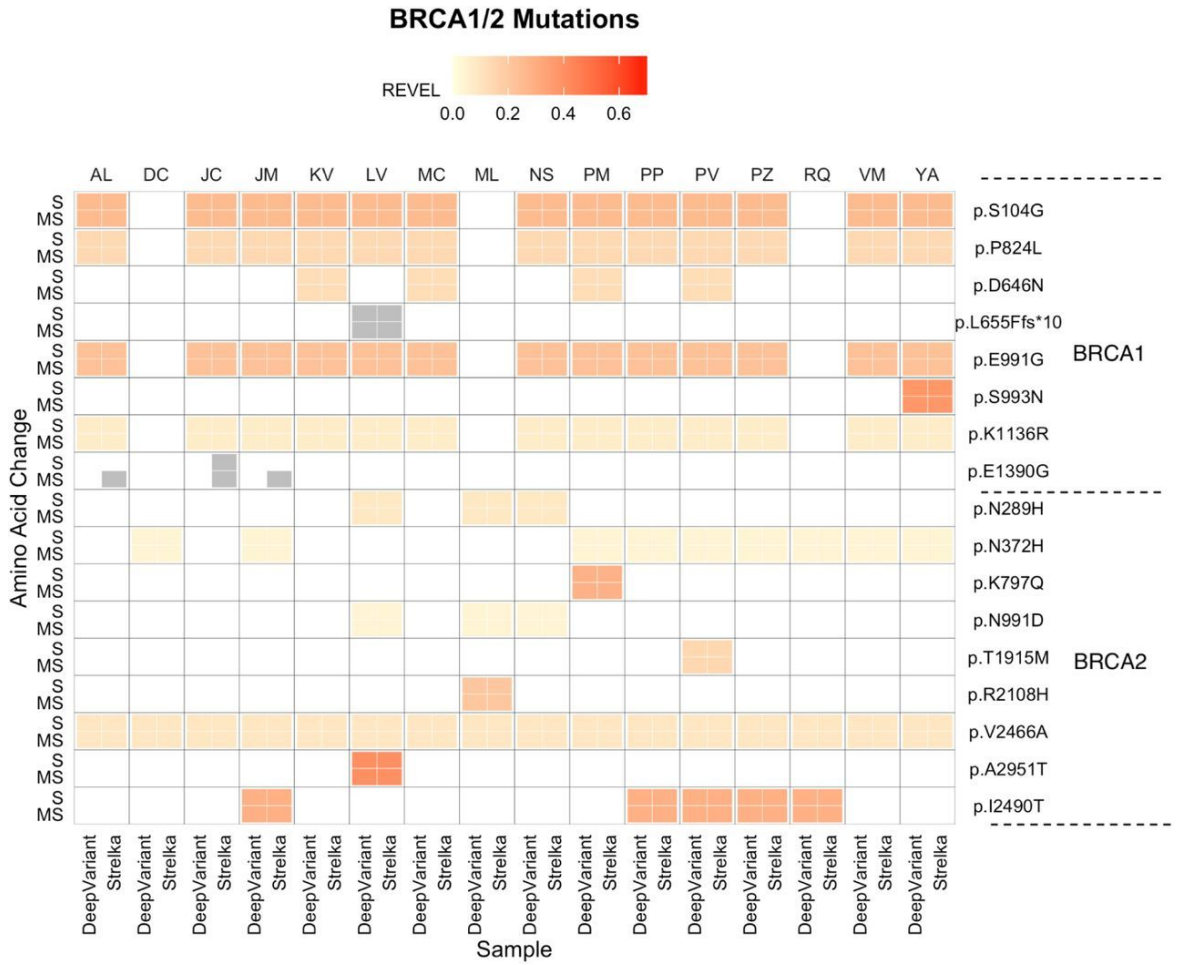


Lollipop Plots de variantes en *BRCA1* y *BRCA2* detectadas por los varcallers Strelka y DeepVariant





- Representación gráfica de los genes BRCA1 y BRCA2, comparando las mutaciones obtenidas por Strelka y DeepVariant.
- El eje Y muestra el número de pacientes con mutaciones de **Missense** (verde) y **frame shift Ins** (morado).
- La sección superior representa los resultados de Strelka y la sección inferior, los resultados de DeepVariant para cada gen. Los dominios están indicados en las etiquetas. Los resultados corresponden a las variantes obtenidas por ambos callers en modo individual.



HeatMap de variantes en BRCA1/2 reportadas en 16 pacientes con cáncer de mama utilizando DeepVariant y Strelka, tanto en modo individual (S) como en modo de múltiples muestras(MS).

- El eje Y muestra los cambios de aminoácidos (a la derecha) en modos individual y agrupado (a la izquierda), y el eje X muestra a los 16 pacientes (arriba) a través de los dos callers de variantes utilizados (abajo).
- El color indica la potencial patogenicidad según el puntaje REVEL. Las variantes sin puntaje REVEL están en gris.
- Modo de muestra individual (S), Modo de múltiples muestras (MS).

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**

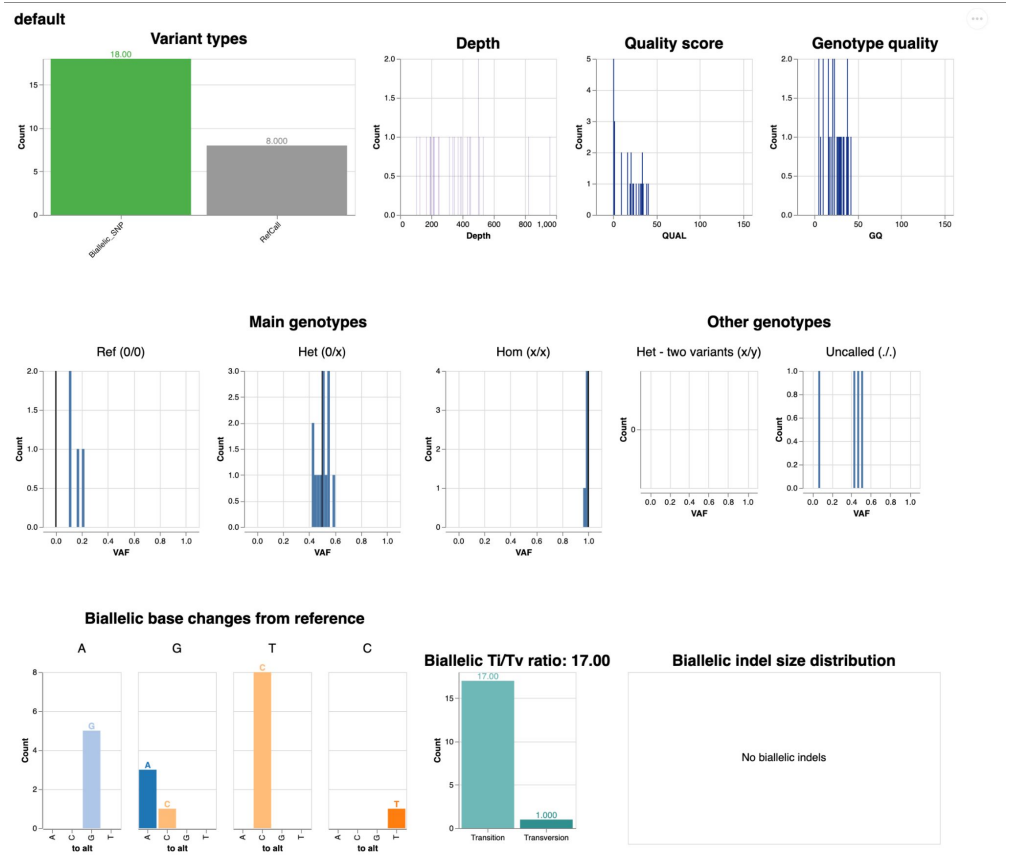
```
1234
ACGT
A-TT
 ^^
```

POS	REF	ALT
1	ACG	AT

VCF

```
##fileformat=VCFv4.2
##contig=<ID=2,length=51304566>
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
#CHROM  POS ID  REF ALT QUAL FILTER  INFO FORMAT     SAMPLE1     SAMPLE2     SAMPLE3     SAMPLE4     SAMPLE5     SAMPLE6     SAMPLE7
2  81170  .  C  T   .  .   AC=9;AN=7424   GT:DP:GQ  0/0:4:12    0/0:3:9     0/1:1:3     0/1:9:24    1/0:4:12    0/0:5:15    0/0:4:12
2  81171  .  G  A   .  .   AC=6;AN=7446   GT:DP:GQ  0/1:4:12    0/0:3:9     0/0:1:3     0/0:9:24    0/1:4:12    0/1:5:15    0/0:4:12
2  81182  .  A  G   .  .   AC=5;AN=7506   GT:DP:GQ  0/0:5:15    0/0:4:12    0/0:5:15    0/0:9:24    0/0:4:12    0/0:4:12    0/0:4:12
2  81204  .  T  G   .  .   AC=2;AN=7542   GT:DP:GQ  1/0:5:15    0/0:9:27    0/0:10:30   0/0:15:39   0/0:9:27    1/0:13:39   0/1:14:42
```

- **CHROM**: Cromosoma de la variante.
- **POS**: Posición de la variante.
- **REF**: Base de referencia.
- **ALT**: Base alternativa.
- **QUAL**: Calidad de la llamada.
- **INFO**: Información adicional (por ejemplo, profundidad de cobertura y frecuencia alélica).



Principales Métricas en el Llamado de Variantes

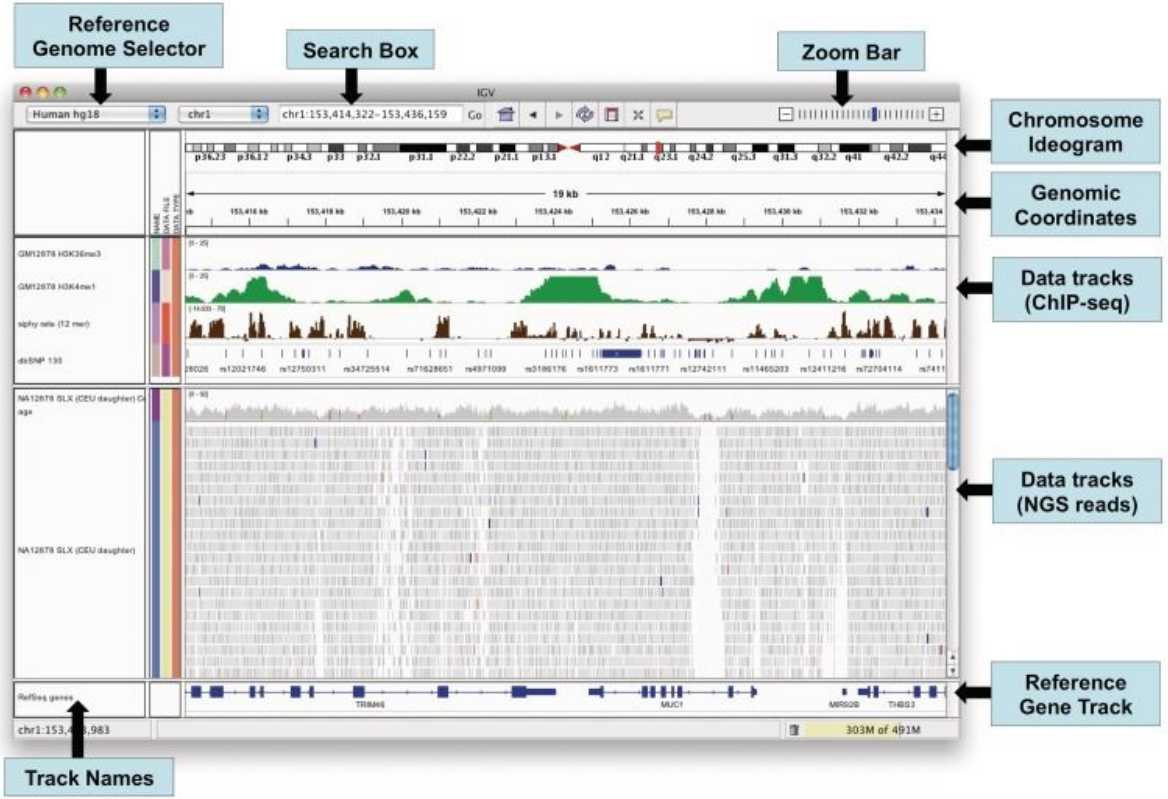
- **Main Genotypes (Genotipos Principales)**: Representación de los alelos observados en una posición (ej. 0/1 para heterocigoto, 1/1 para homocigoto variante).
- **Depth (Profundidad de Cobertura)**: Número de lecturas que cubren la posición variante, indicando la confiabilidad en la detección.
- **Quality Score (Puntuación de Calidad)**: Valor que mide la confianza en la presencia de una variante en una posición específica; cuanto mayor es el valor, mayor es la seguridad de que la variante es real.
- **Genotype Quality (Calidad del Genotipo)**: Estima la precisión de los genotipos asignados (ej. si un genotipo 0/1 realmente corresponde a heterocigoto). Un puntaje alto sugiere una menor probabilidad de error en el genotipo

Integrative Genomics Viewer (IGV)

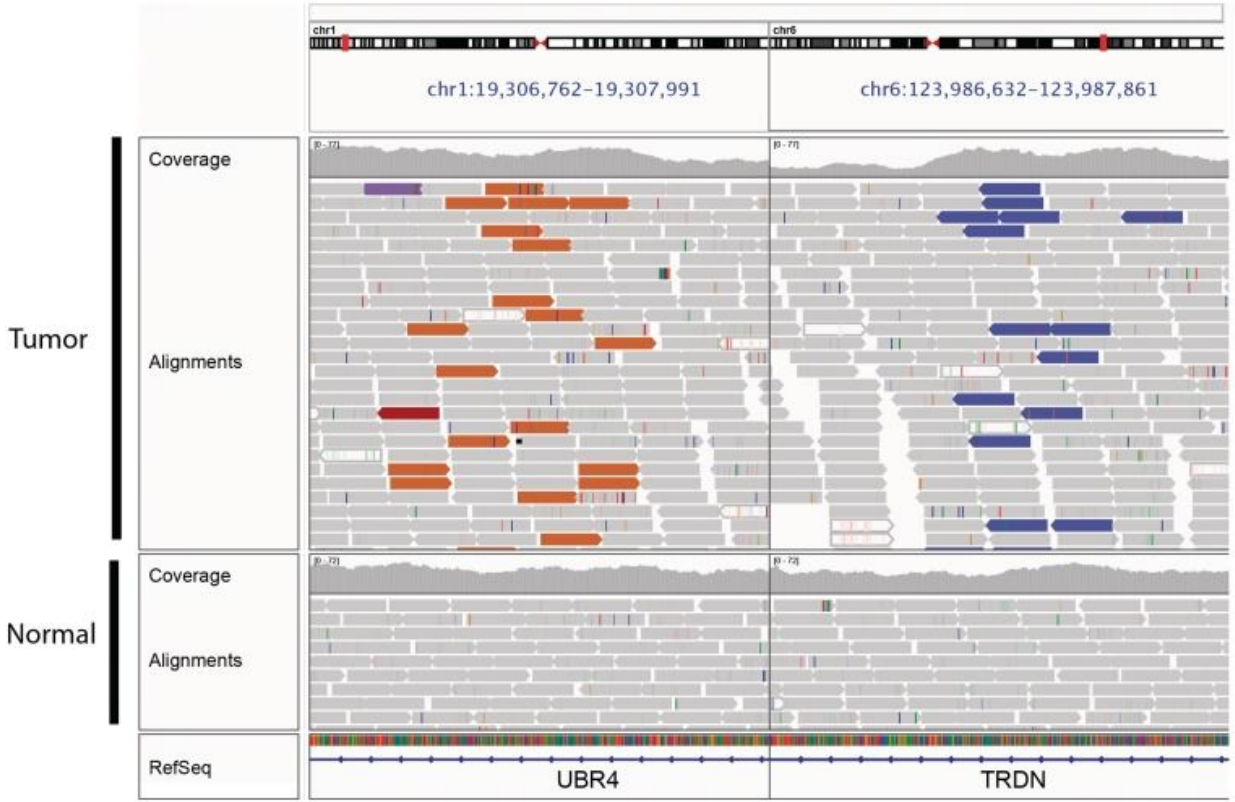
Características de IGV

1. **Visualización a múltiples niveles**: IGV permite navegar desde la vista del genoma completo hasta el nivel de nucleótidos individuales, facilitando la exploración de variantes en profundidad.
2. **Compatibilidad con varios tipos de archivos**: IGV soporta formatos como BAM, CRAM, VCF, BED, GFF, y más, lo cual facilita el uso de datos de diferentes etapas del pipeline de análisis de NGS.
3. **Exploración de datos interactiva**: Permite acercar y alejar, visualizar cobertura de secuencias, y ver alineaciones específicas de lectura en regiones genómicas de interés.

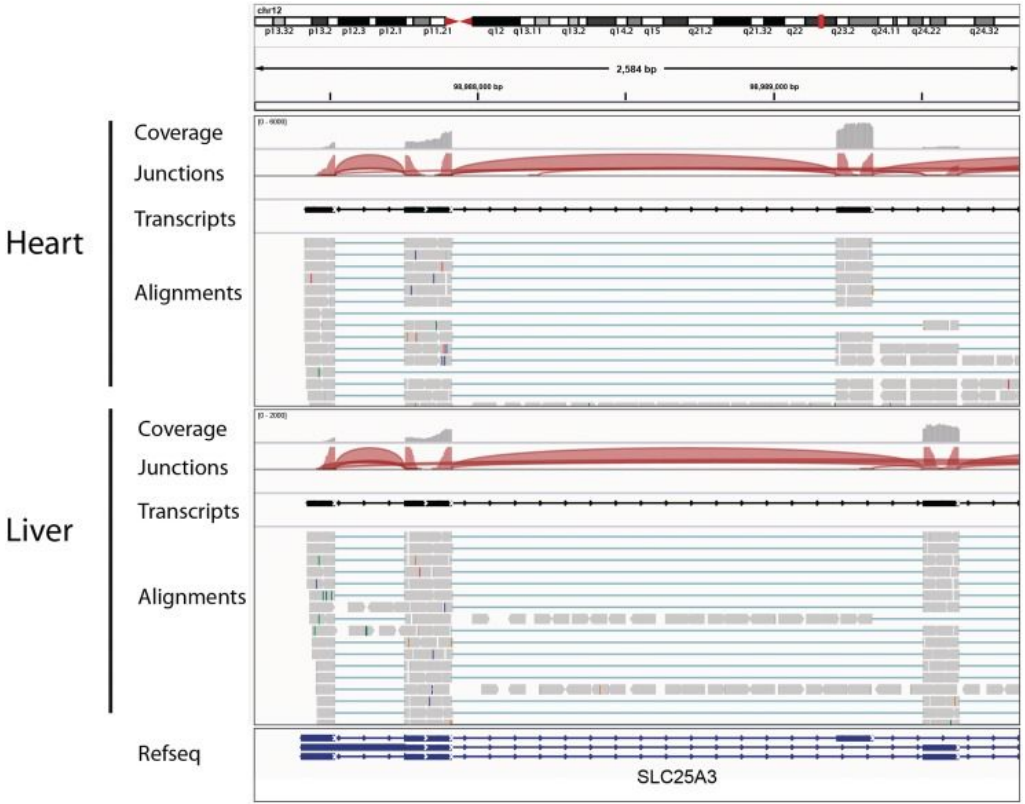
Integrative Genomics Viewer (IGV)



IGV: Comparación de Tumor vs normal de un paciente

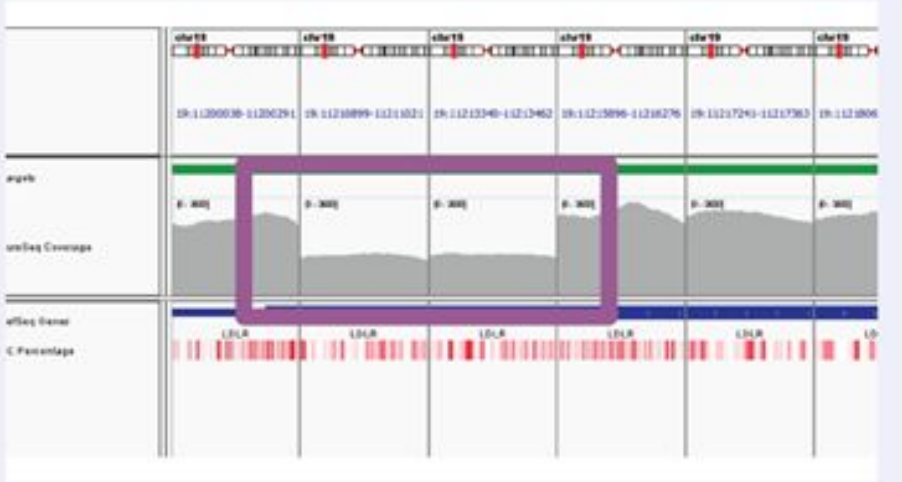


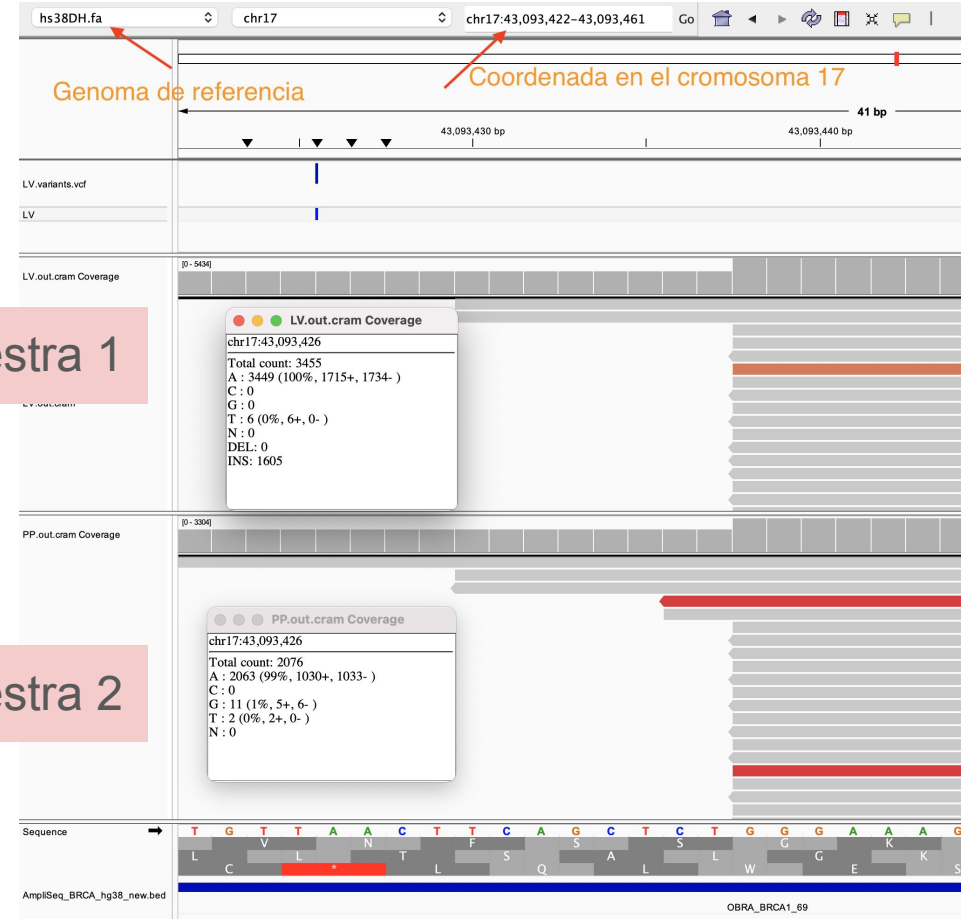
IGV: Visualización de datos de RNA-seq de muestras de tejido



Visualización de datos de RNA-seq de muestras de tejido cardíaco y hepático.

Cada panel incluye pistas para la cobertura total, cobertura de uniones, transcritos predichos y alineaciones de lecturas.
Las lecturas que abarcan las uniones están conectadas con líneas azules delgadas.
Hay evidencia clara de empalme alternativo entre los dos tejidos.





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.

Práctico de Llamado de Variantes y Visualización con IGV

1. Realizar el llamado de variantes en datos genómicos utilizando Strelka y DeepVariant en Galaxy.
2. Comprender qué es un llamado de variantes y cómo interpretar archivos de variantes (VCF).
3. Comparar las variantes detectadas entre Strelka y **DeepVariant** en una misma muestra.
4. Visualizar y analizar variantes en el genoma de referencia usando **IGV**.