Análisis de resultados de secuenciación

Karen Oróstica Tapia, PhD
Universidad del Desarrollo

Contenido

1.Introducción

- Tipo de variantes
- Diluvio de datos

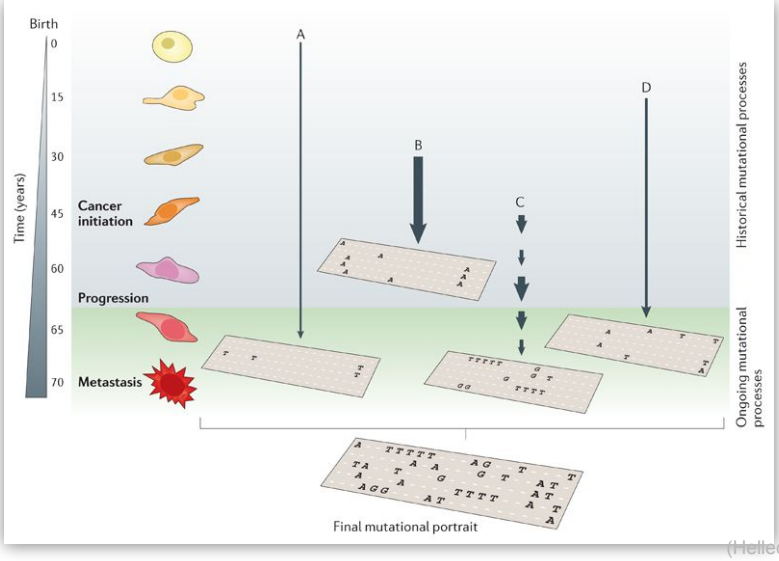
2.Pipeline BI

- Preprocesamiento
- Llamado de Variantes
- Anotación funcional

3.Interpretación

- Contenido VCF file
- Bases de datos y herramientas

Proceso Mutacional



(Helleday, Eshtad, and Nik-Zainal 2014)

Proceso Mutacional

Variantes

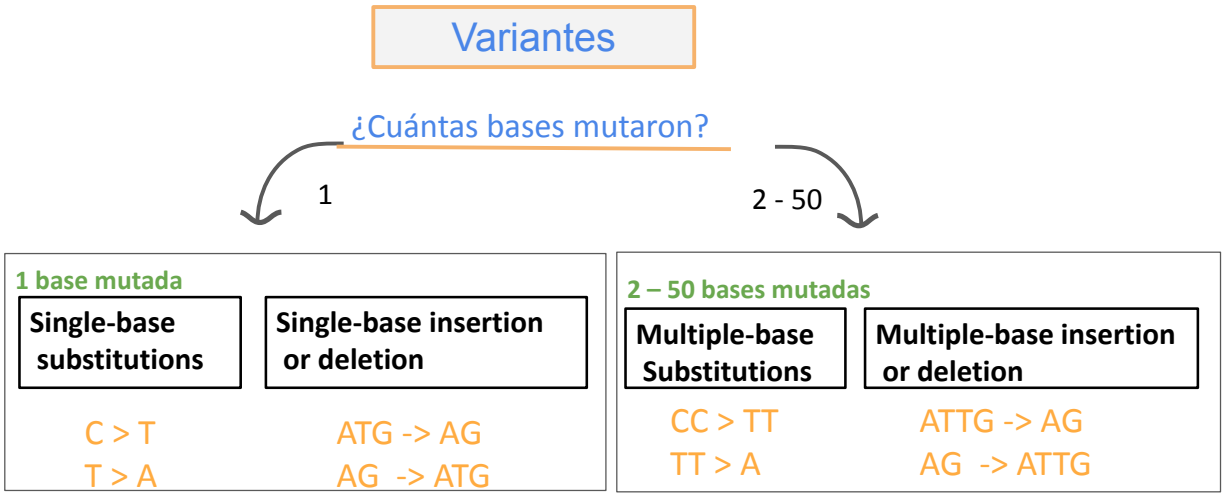
¿Cuántas bases mutaron?

1

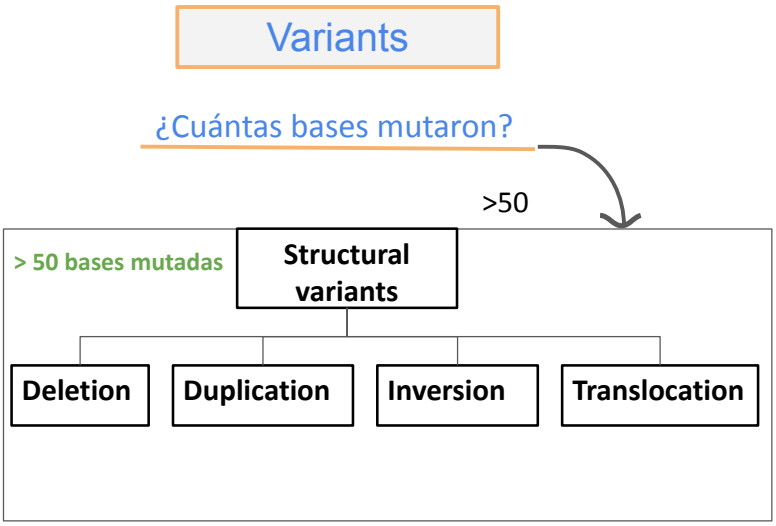
1 base mutada

Single-base substitutions	Single-base insertion or deletion
C > T	ATG -> AG
T > A	AG -> ATG

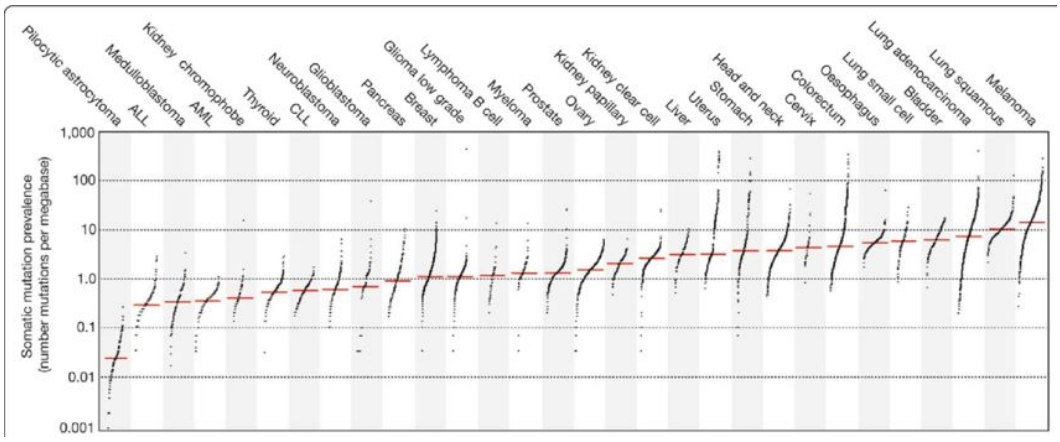
Proceso mutacional



Proceso mutacional



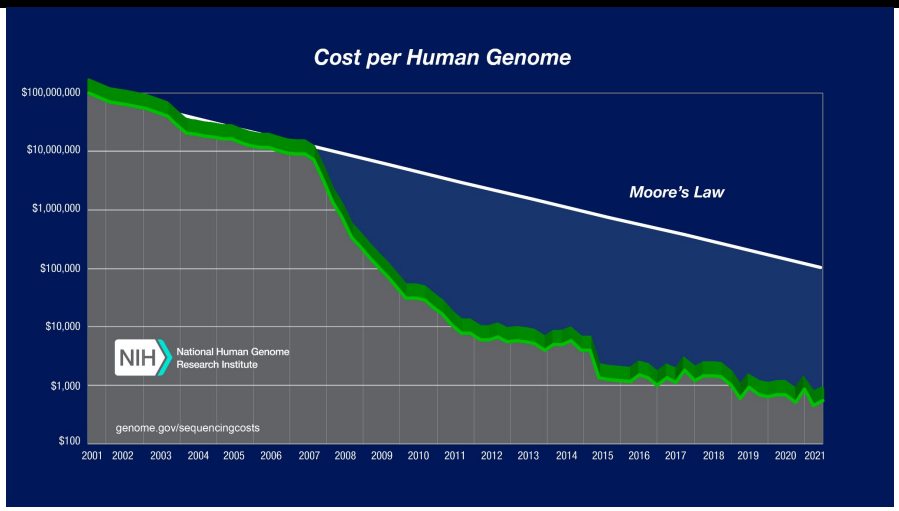
Carga Mutational por cáncer



Con el inicio de las tecnologías de secuenciación de próxima generación hemos podido identificar perfiles genómicos

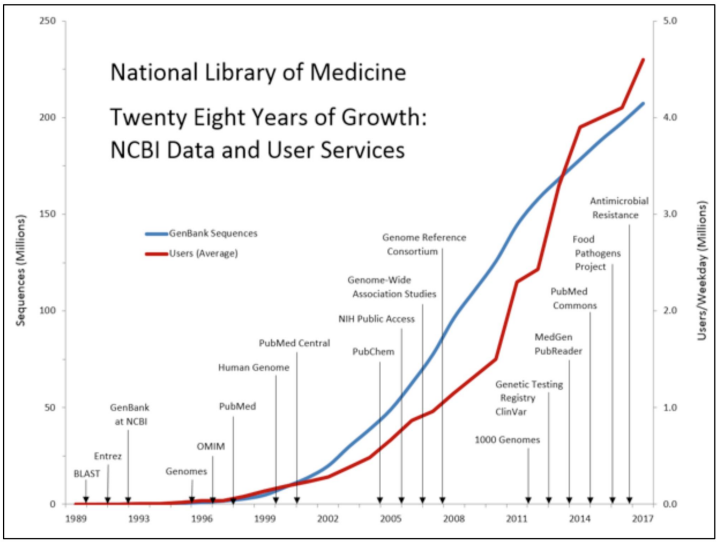
(Alexandrov et al. 2013)

Diluvio de información genómica

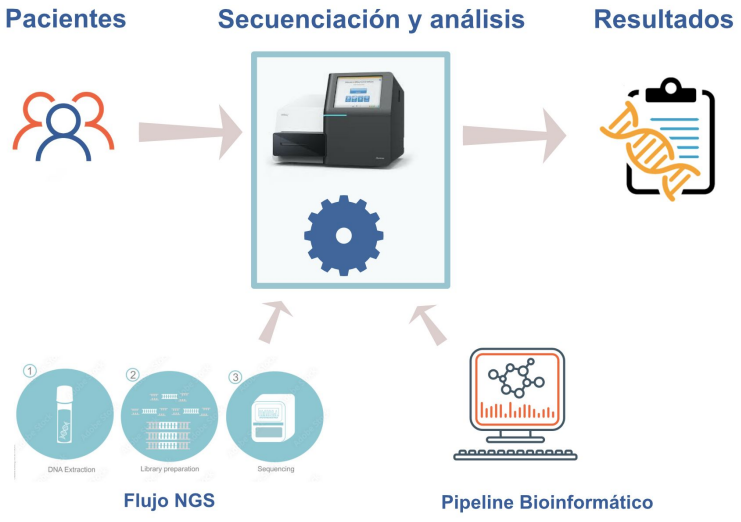


Cost of sequencing the human genome (2001-2020)
National Human Genome Research Institute

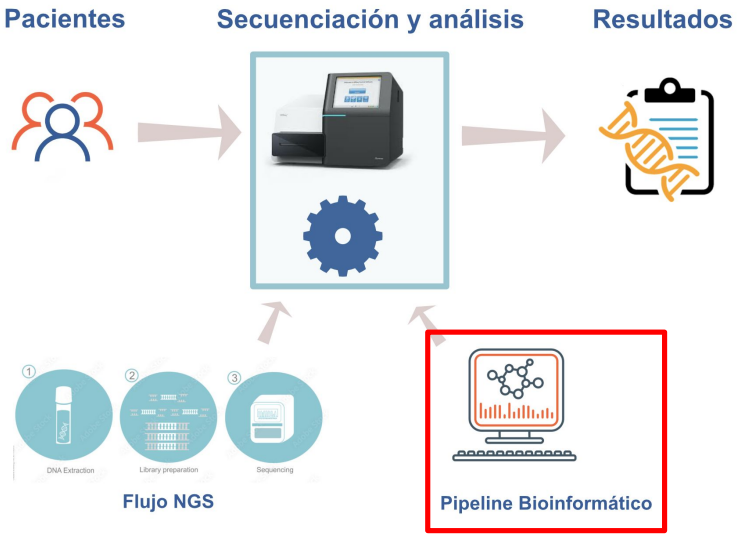
Diluvio de información genómica



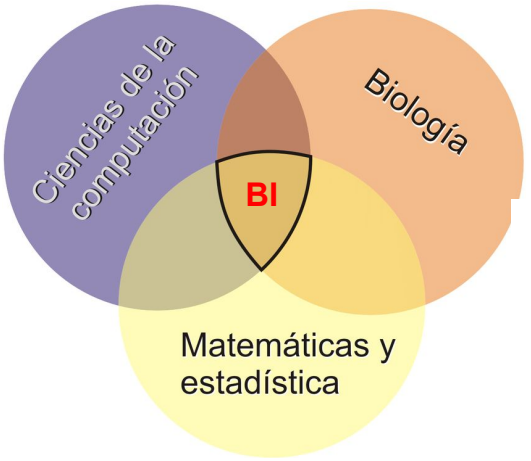
Proceso general de secuenciación



Proceso general de secuenciación

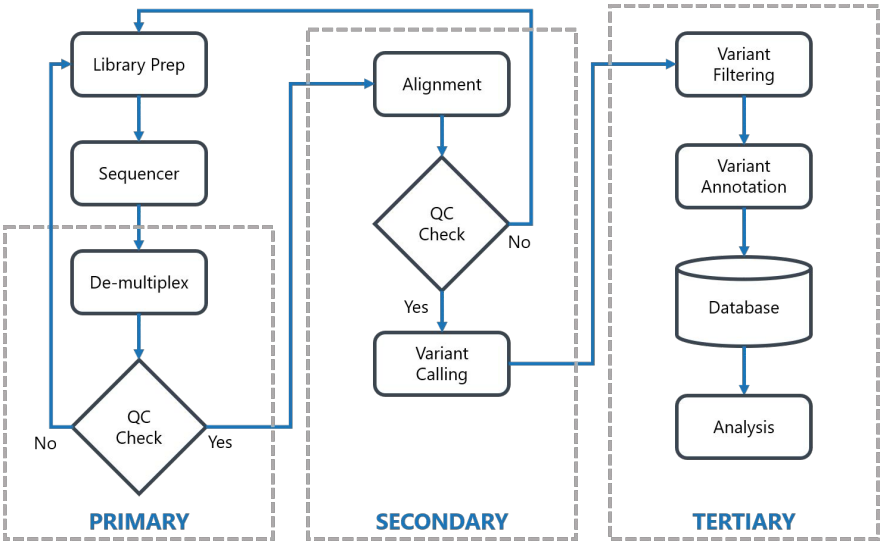


¿Qué es un pipeline bioinformático?

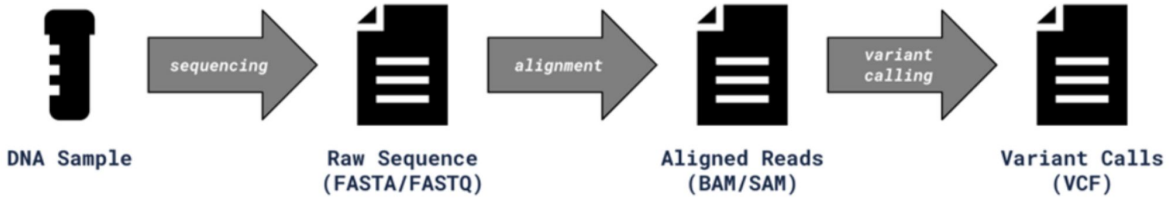


Un pipeline bioinformático es **una serie de algoritmos o softwares que procesan datos** biológicos brutos para luego generar interpretaciones a partir de estos datos.

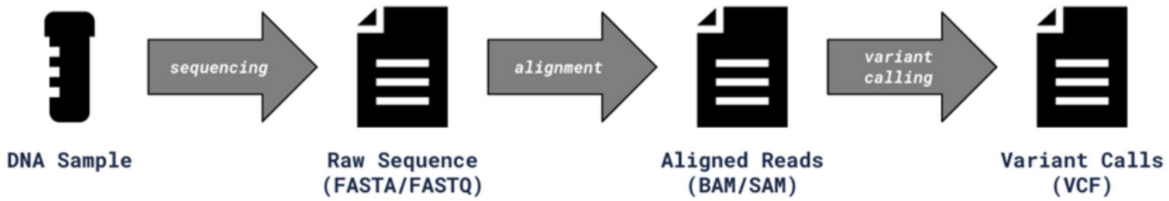
¿Qué es un pipeline bioinformático?



Pipeline bioinformático en genómica del cáncer

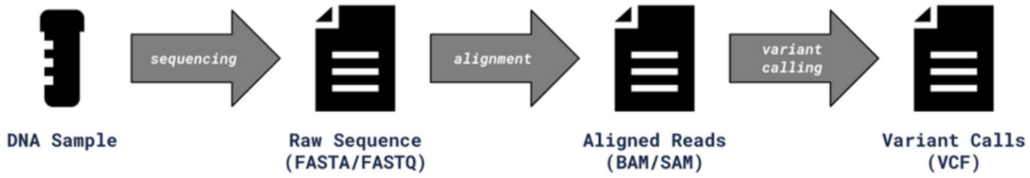


Pipeline bioinformático en genómica del cáncer



RAW data → **Datos procesados**

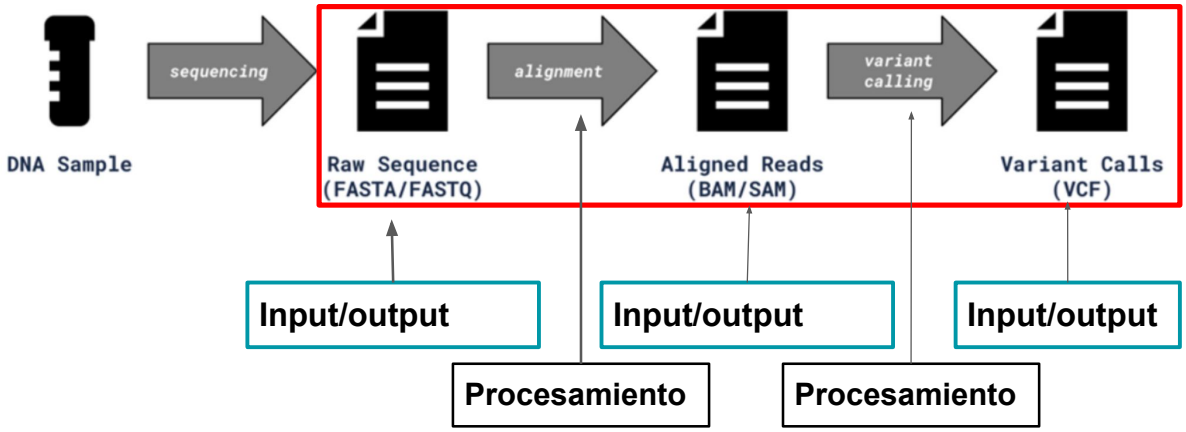
Volumen de información por tipo de archivo



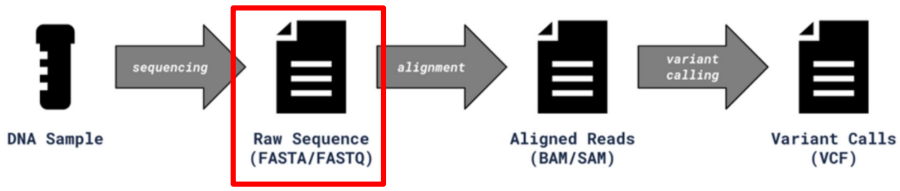
	Depth	FASTQ.gz	BAM	VCF.gz
Genome	30x	100-200Gb	100-200 Gb	2Gb
Exome (50 Mb)	120x	5-10 Gb	10-30Gb	3-6Mb
Deep Targeted (150 kb)	1500x	300-500Mb	150Mb	3-6Mb

Assume 100 bp read length

Pipeline bioinformático en genómica del cáncer



Estructura de un Fastq file




```
@IL31_4368:1:1:996:3507/2
TCCCTTACCCCCAAGCTCCATACCCTCCTAATGCCCACACCTCTTACCTTAGGA
+
FFCEFFFEEFFFFFFFEFFEFFFEFCFC<EEFEFFFCEFF<;EEFF=FEE?FCE
@IL31_4368:1:1:996:21421/2
CAAAAACTTTCACTTTACCTGCCGGGTTTCCCAGTTTACATTCCACTGTTTGAC
+
>DBDDB,B9BAA4AAB7BB?7BBB=91;+*@;5<87+*=/*@@?9=73=.7)7*
@IL31_4368:1:1:997:10572/2
GATCTTCTGTGACTGGAAGAAAATGTGTTACATATTACATTTCTGTCCCCATTG
+
E?=EECE<EEEE98EEEEAEEBD??BE@AEAB><EEABCEEDEC<<EBDA=DEE
@IL31_4368:1:1:997:15684/2
CAGCCTCAGATTCAGCATTCTCAAATTCAGCTGCGGCTGAAACAGCAGCAGGAC
+
EEEEDEEE9EAEEDEEEEEEEEEEECEEAAEEDEE<CD=D=*BCAC?;CB,<D@,
@IL31_4368:1:1:997:15249/2
AATGTTCTGAAACCTCTGAGAAAGCAAATATTTATTTTAATGAAAAATCCTTAT
+
EDEEC;EEE;EEE?EECE;7AEEEEEE07EECEA;D6D>+EE4E7EEE4;E=EA
@IL31_4368:1:1:997:6273/2
ACATTTACCAAGACCAAAGGAAACTTACCTTGCAAGAATTAGACAGTTCATTTG
+
EEAAFFFEEFEFCFAFFAFCCFFEFEF>EFFFFB?ABA@ECEE=<F@DE@DDF;
@IL31_4368:1:1:997:1657/2
CCCACCTCTCTCAATGTTTTCCATATGGCAGGGACTCAGCACAGGTGGATTAAT
(...)
```

- Instrument serial #
- Lane
- Swath
- X coord
- Y coord
- Read direction

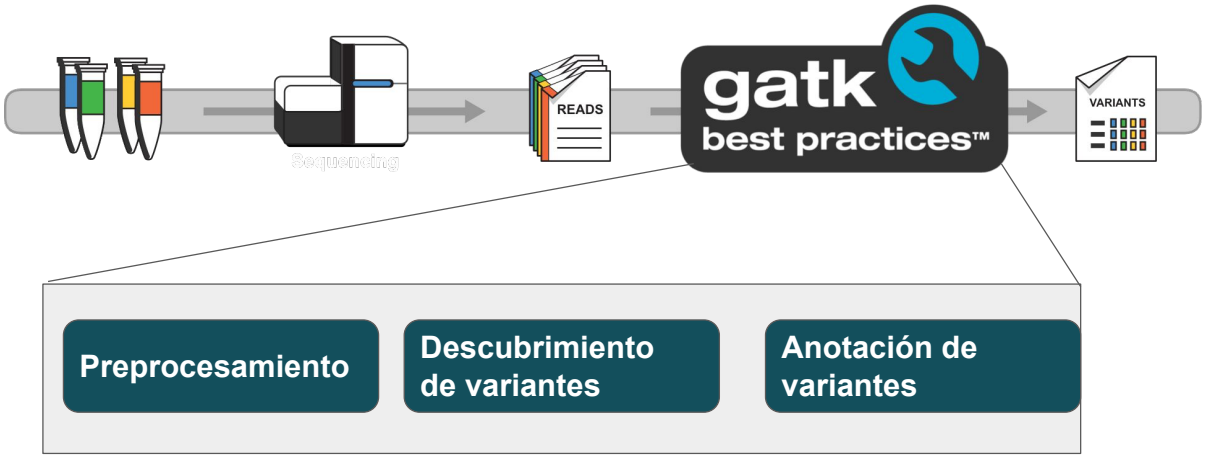
```
+SEQ_ID
!''*((((***+))%%%++)(%%%%).1**
```

Codificación ASCII del valor de calidad Q, el cual representa la probabilidad *P* de que el base calling sea incorrecto

$$Q = -10\ \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

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 10000	99.99%
50	1 in 100000	99.999%

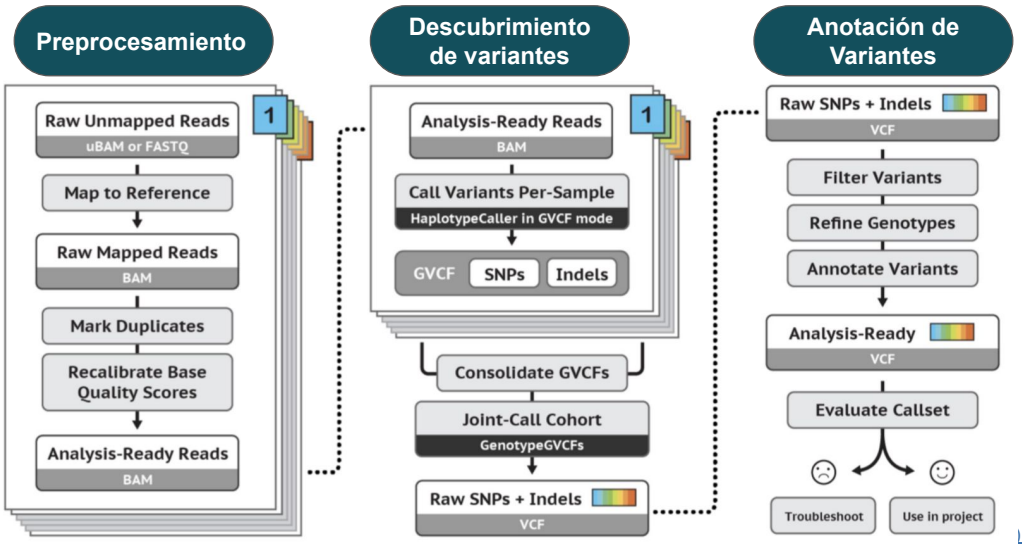
Pipeline bioinformático en genómica del cáncer



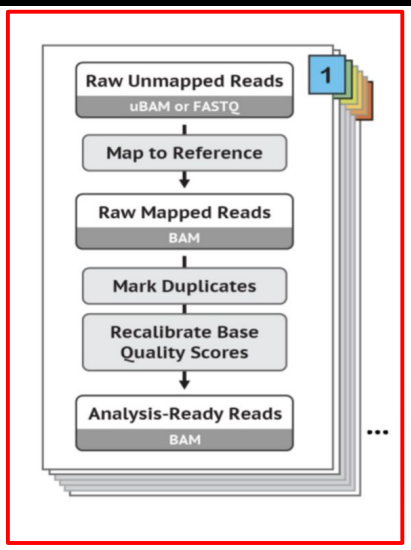
GATK:paquete de software para analizar datos de secuenciación de alto rendimiento

Genome Analysis Tookit (GATK)

Pipeline bioinformático en genómica del cáncer



Preprocesamiento de datos: análisis de calidad

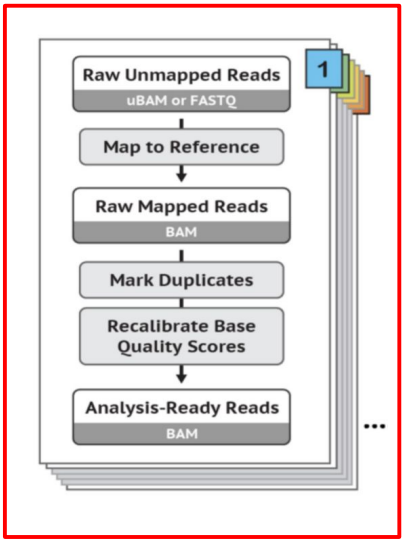


- Las tecnologías de NGS generan gran cantidad de lecturas de secuencias.
- Generación de diferentes tipos y cantidades de errores (como la identificación de nucleótidos incorrectos).
- Es necesario comprender, identificar y excluir los tipos de errores que pueden afectar la interpretación del análisis posterior.
- Por lo tanto, el control de calidad de la secuencia es un primer paso esencial en su análisis. La detección temprana de errores ahorra tiempo más adelante.

Necesitamos filtrar y preparar las lecturas y los datos de alineamiento para el llamado de variantes

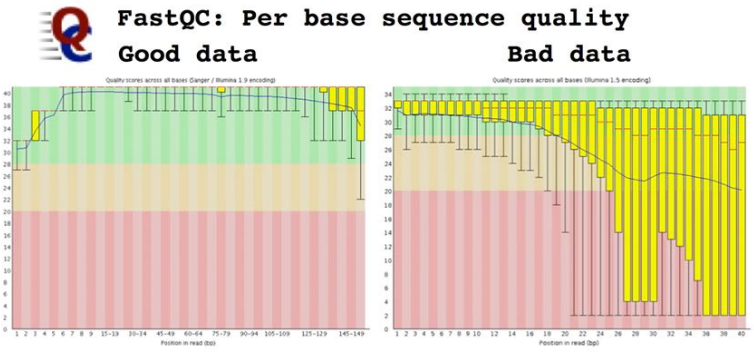
Genome Analysis Tookit (GATK)

Preprocesamiento de datos



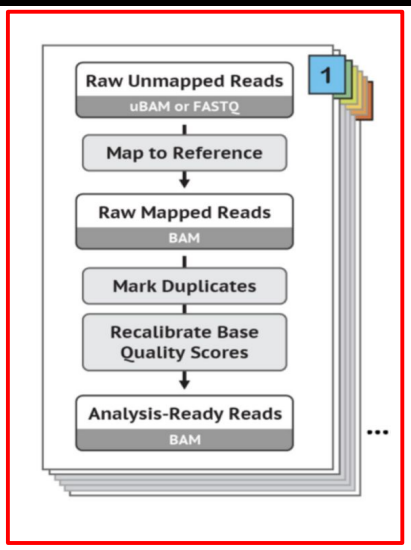
Calidad de las secuencias por base

- Sequence Analysis Viewer (SAV) en Basespace
- FastQC
- MultiQC



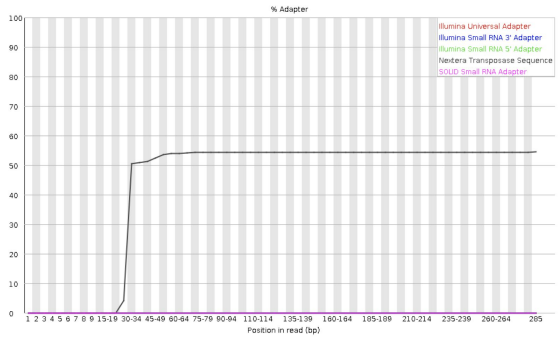
Genome Analysis Tookit (GATK)

Preprocesamiento de datos

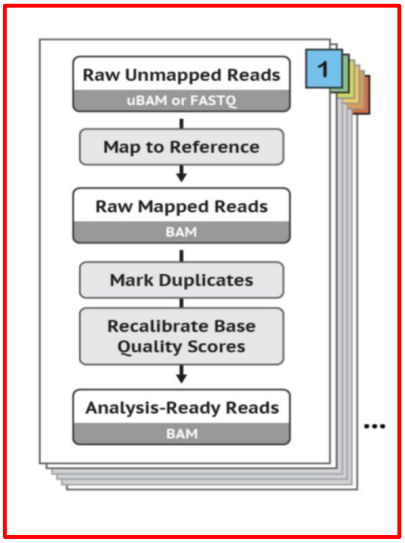


Contenido de adaptadores
El gráfico muestra el porcentaje acumulado de lecturas con las diferentes secuencias de adaptadores en cada posición.

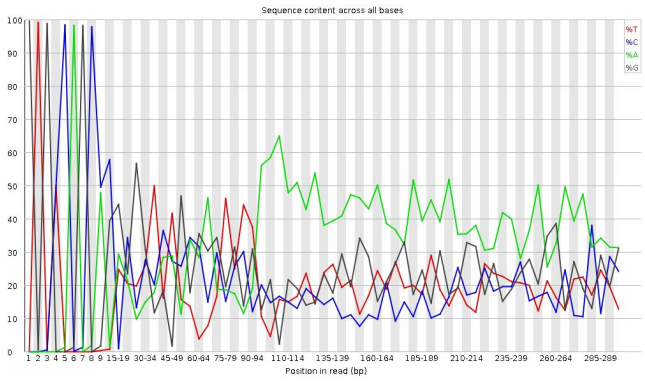
FastQC puede detectar algunos adaptadores de forma predeterminada (por ejemplo, Illumina, Nextera).



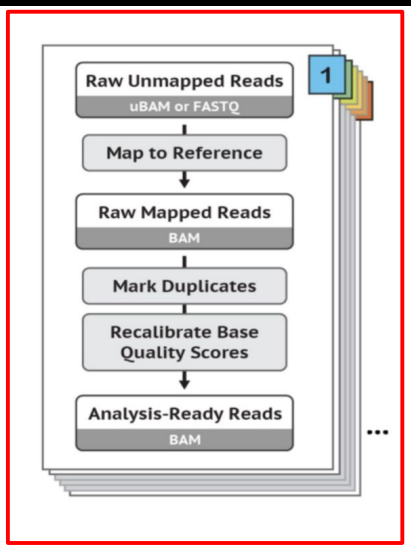
Preprocesamiento de datos



Por contenido de secuencia base
traza el porcentaje de cada uno de los cuatro nucleótidos (T, C, A, G) en cada posición en todas las lecturas en el archivo de secuencia de entrada.



Preprocesamiento de datos



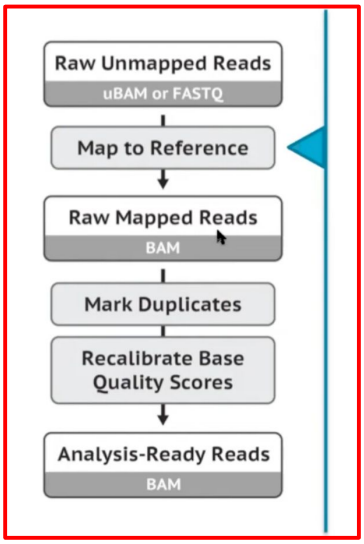
Trimming y filtro de secuencias
La baja calidad de nucleótidos podría causar un sesgo en los análisis posteriores.
Las secuencias deben **pre-procesarse** para reducir el sesgo en el análisis posterior.

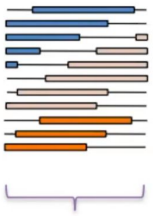
El trimming incluye:

1. Recorte/corte/enmascaramiento de secuencias
   a. Regiones con bajo nivel de calidad
   b. Principio/fin de secuencia
   c. Eliminación de adaptadores

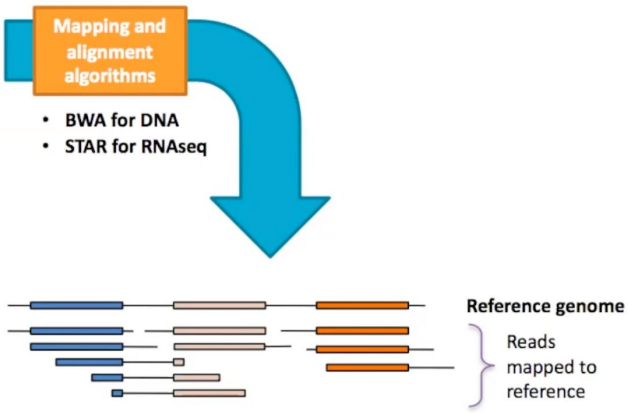
2. Filtrado de secuencias:
   a. Con puntuación de calidad media baja
   b. demasiado corto
   c. Con demasiadas bases ambiguas (N)

Preprocesamiento de datos

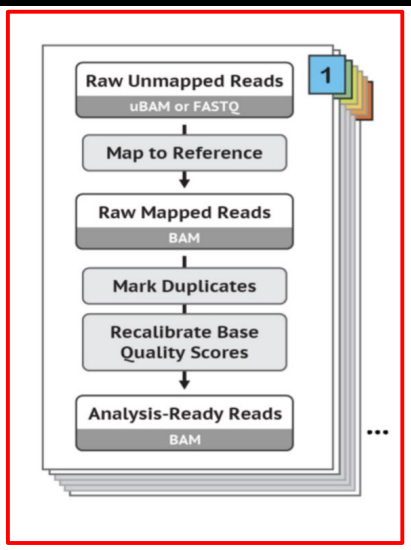




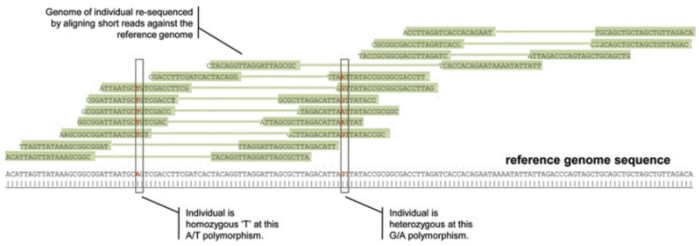
Enormous pile of short reads from sequencer



Preprocesamiento de datos

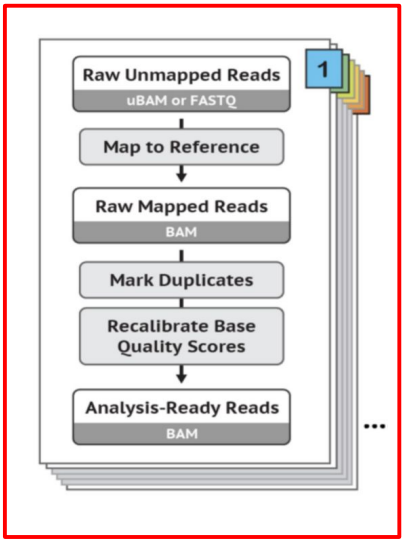


Alineamiento o mapeo de secuencias



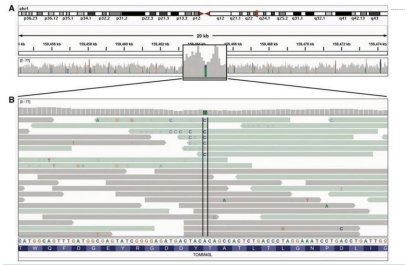
Actualmente, hay más de 60 mapeadores diferentes.
BWA es una herramienta rápida y eficiente en memoria, particularmente buena para alinear lecturas de secuenciación de alrededor de 50 hasta 1000 de bases para genomas relativamente largos.

Preprocesamiento de datos



Alineamiento o mapeo de secuencias

Integrative Genomic Viewer

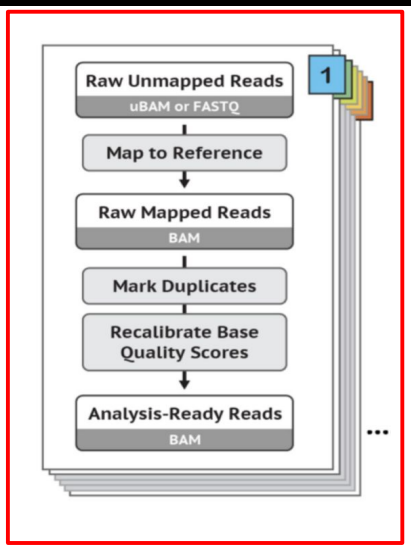


¿Qué mirar en una visualización de alineamiento?

- Cobertura en regiones de interés
- Posibles sesgos de variantes entre R1 y R2 por hebra
- Variantes e INDELs en los extremos

https://software.broadinstitute.org/software/igv/

Preprocesamiento de datos

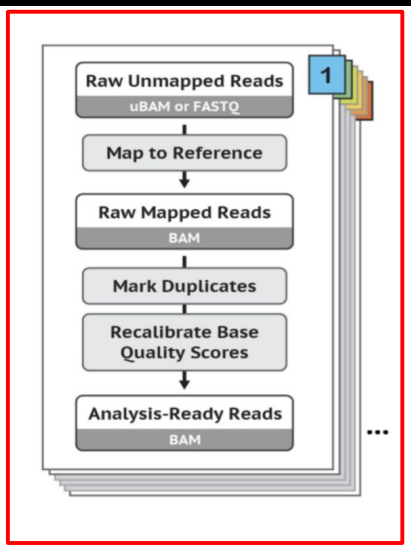


Métricas básicas de calidad del alineamiento

- % de lecturas mapeados
- % de lecturas únicamente mapeados
- % de lecturas efectivamente mapeados (luego de eliminar duplicados)
- Profundidad promedio (x)
- Cobertura (% de la región blanco que fue cubierta con lecturas)
- Calidad de mapeo (QMAP para cada lectura)-> tasa de error
- % de lecturas en el blanco (ontarget)

https://software.broadinstitute.org/software/igv/

Preprocesamiento de datos



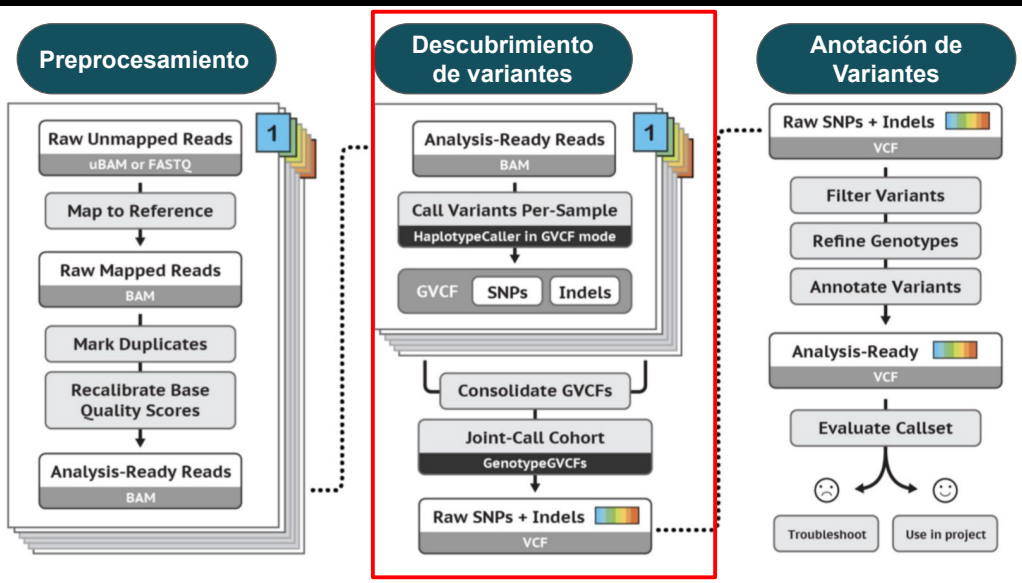
Conclusiones del preprocesamiento

Después del control de calidad, el mapeo es un paso importante de la mayoría de los análisis de datos de secuenciación (RNA-Seq, ChIP-Seq, etc.) ya que nos ayuda a determinar el puzzle y usar esta información para análisis posteriores.

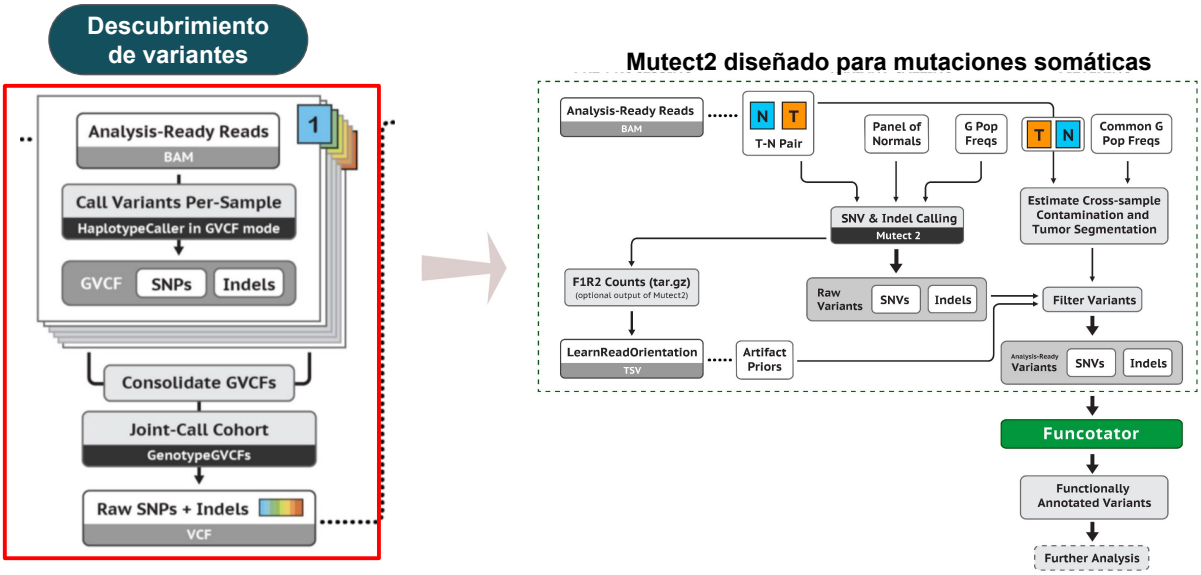
Puntos clave

1. **¡Conoce tus datos!**
2. **El mapeo no es trivial**
3. **Hay muchos algoritmos de mapeo, depende de sus datos cuál elegir.**

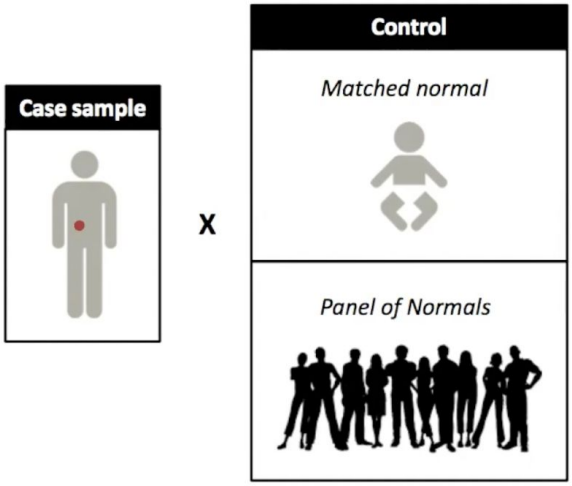
Pipeline bioinformático



Pipeline bioinformático en genómica del cáncer

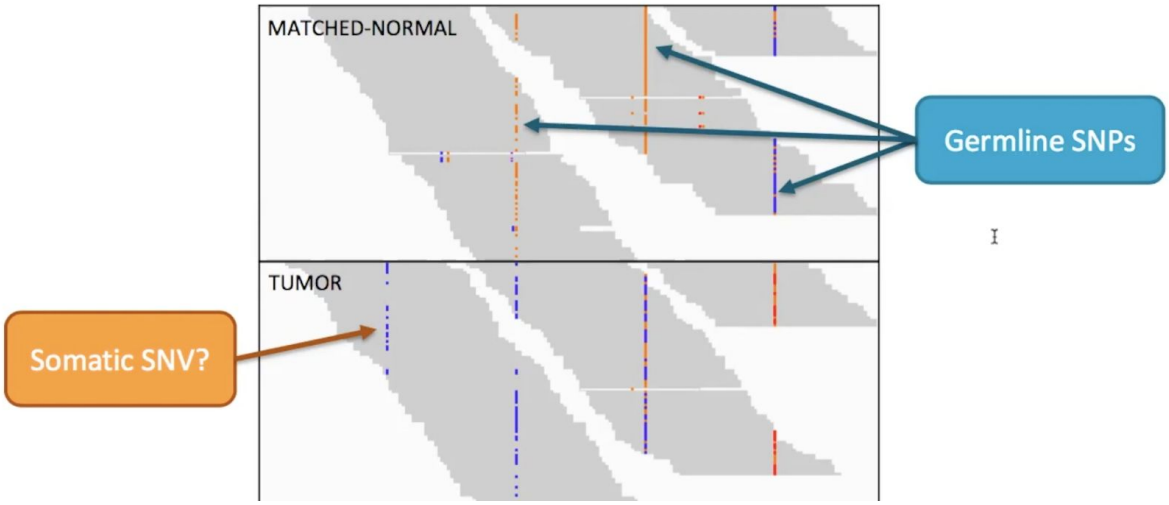


Pipeline bioinformático



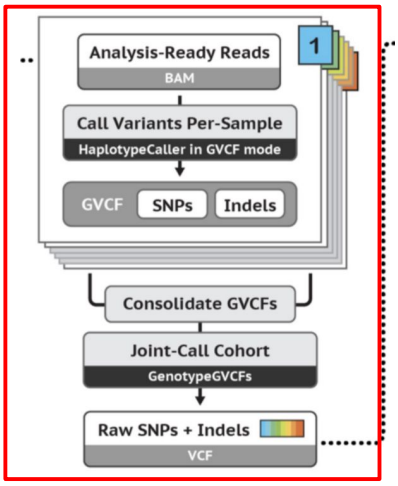
- *Matched normal*
  to subtract germline background
- *Panel of Normals*
  to subtract systematic noise

Pipeline bioinformático

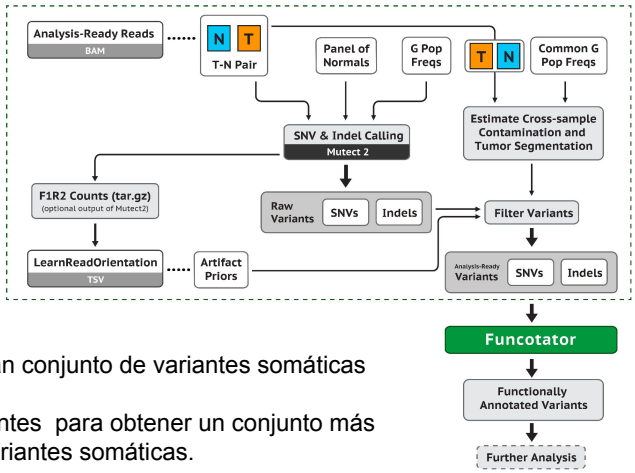


Pipeline bioinformático en genómica del cáncer

Descubrimiento de variantes



Identifique variantes somáticas (SNV e Indels) en una o más muestras de tumores de un solo individuo, con o sin una muestra normal coincidente.



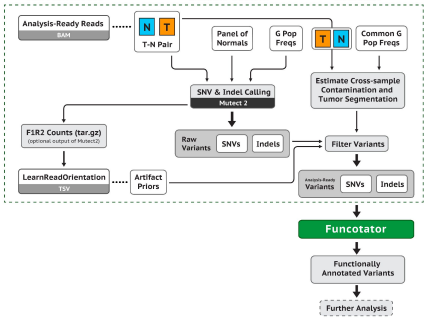
Objetivos

Generar un gran conjunto de variantes somáticas candidatas.

Filtrar las variantes para obtener un conjunto más confiable de variantes somáticas.

Pipeline bioinformático en genómica del cáncer

Descubrimiento de variantes



Mutect2 llama a SNV e indels simultáneamente a través del ensamblaje local de novo de haplotipos en una región activa.

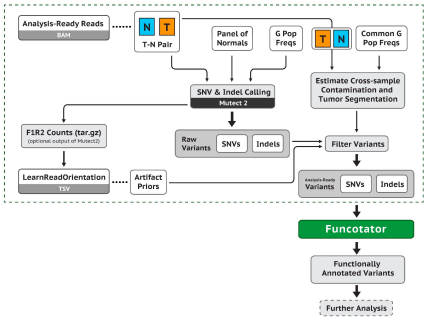
Mutect2 encuentra una región que muestra signos de variación somática, descarta la información de mapeo existente y vuelve a ensamblar completamente las lecturas en esa región para generar variantes de haplotipos candidatos.

Mutect2 luego alinea cada lectura con cada haplotipo a través del algoritmo Pair-HMM para obtener una matriz de probabilidades.

Mutect2 aplica un modelo bayesiano de probabilidades somáticas para obtener las probabilidades logarítmicas de que los alelos sean variantes somáticas versus errores de secuenciación.

Pipeline bioinformático en genómica del cáncer

Descubrimiento de variantes

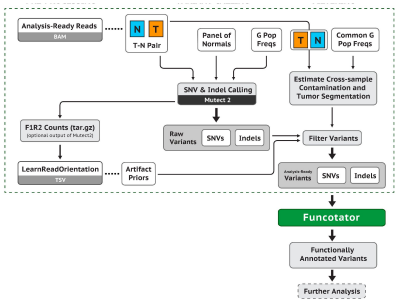


GetPileupSummaries, CalculateContamination
Este paso emite una estimación de la fracción de lecturas debido a la contaminación entre muestras para cada muestra de tumor.

A diferencia de otras herramientas de contaminación, CalculateContamination está diseñado para funcionar bien sin una muestra normal coincidente.

Pipeline bioinformático en genómica del cáncer

Descubrimiento de variantes



Variant Calling Format (VCF) es un archivo de texto delimitado por tabuladores que se utiliza para describir variantes de un solo nucleótido (SNV), así como indels y otras variaciones de secuencia.

VCF es un formato preferido porque es inequívoco, escalable y flexible, lo que permite agregar información adicional al campo de información. Se pueden almacenar muchos millones de variantes en un solo archivo VCF.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA19909
11	5248232	rs334	T	A	100	PASS	AA=T|||;AC=1;AF=0.0273562;AFR_AF=0.0998;AMR_AF=0.0072;AN=2;DP=22876;EAS_AF=0;EUR_AF=0;EX_TARGET;NS=2504;SAS_AF=0;VT=SNP	GT	0|1

Pipeline bioinformático en genómica del cáncer

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA19909
11	5248232	rs334	T	A	100	PASS	AA=T|||;AC=1;AF=0.0273562;AFR_AF=0.0998;AMR_AF=0.0072;AN=2;DP=22876;EAS_AF=0;EUR_AF=0;EX_TARGET;NS=2504;SAS_AF=0;VT=SNP	GT	0|1

Column Content Description

#CHROM Chromosome

POS Co-ordinate - The start coordinate of the variant.

ID Identifier

REF Reference allele - The reference allele is whatever is found in the reference genome. It is not necessarily the major allele.

ALT Alternative allele - The alternative allele is the allele found in the sample you are studying.

QUAL Score - Quality score out of 100.

FILTER Pass/fail - If it passed quality filters.

INFO Further information - Allows you to provide further information on the variants. Keys in the INFO field can be defined in header lines above the table.

FORMAT Information about the following columns - The GT in the FORMAT column tells us to expect genotypes in the following columns.

NA19909 Individual identifier (optional) - The previous column told us to expect to see genotypes here. The genotype is in the form 0|1, where 0 indicates the reference allele and 1 indicates the alternative allele, i.e it is heterozygous. The vertical pipe | indicates that the genotype is phased, and is used to indicate which chromosome the alleles are on. If this is a slash / rather than a vertical pipe, it means we don't know which chromosome they are on.

Estructura de VCF

```
##fileformat=VCFv4.1
##reference=1000GenomesPilot-NCBI36
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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">
#CHROM POS    ID         REF ALT  QUAL  FILTER INFO                FORMAT     NA00001   NA00002   NA00003
20    14370   rs6054257 G   A    29    PASS   DP=14;AF=0.5  GT:GQ:DP   0/0:48:1  1/0:48:8  1/1:43:5
20    1230237 .         T   .    47    PASS   DP=13         GT:GQ:DP   0/0:54:7  0/0:48:4  0/0:61:2
20    1234567 .         GT  G    50    PASS   DP=9          GT:GQ:DP   0/1:35:4  0/2:17:2  1/1:40:3
```

Header

Records

Identificadores de variantes

Los VCF incluyen identificadores de variantes.

Los identificadores son combinaciones únicas de letras y números que se asignan a entidades conocidas, como genes, variantes y proteínas.

Los identificadores de variantes son particularmente útiles cuando se busca información sobre una variante porque son **inequívocos, únicos y estables, a diferencia de los nombres descriptivos,** que pueden ser utilizados de manera diferente por diferentes personas, ser idénticos entre especies y cambiar con el tiempo.

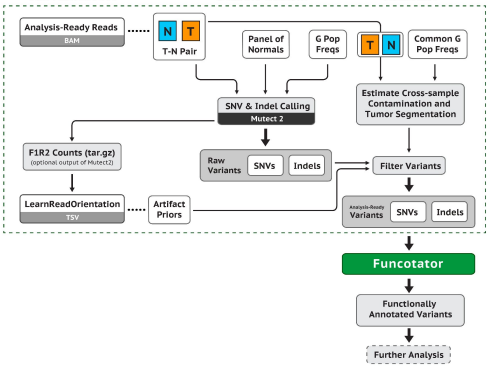
Las variantes pueden tener identificadores de múltiples bases de datos.

Identificadores de variantes

Identifier type	Example	Description
ssID	ss335	ID de SNP asignado por dbSNP
HGVS*	ENST00000366667.4:c.803T>C	Expresa la ubicación de la variante en términos de una transcripción o proteína.
COSMIC ID	COSM1290	ID asignado por COSMIC para variantes somáticas.
ClinVar	RCV000016573	ID asignado a variantes anotadas de dbSNP o dbVar/DGVa, vinculándolas a la salud humana.
UniProt	VAR_010085	ID asignado por UniProt curados por humanos.

Pipeline bioinformático en genómica del cáncer

Descubrimiento de variantes

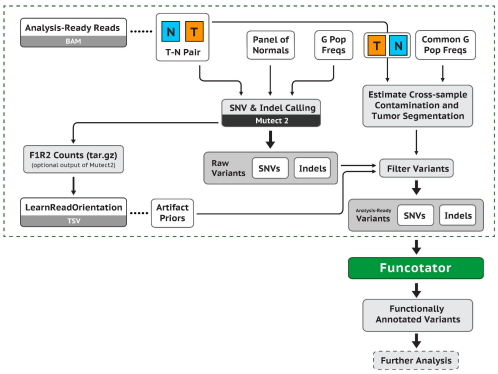


FilterMutectCalls da cuenta de los errores correlacionados, es decir, la posibilidad de que todas las lecturas variantes en un sitio se deban a alguna fuente común de error. Logra esto a través de varios filtros para detectar artefactos de alineación y modelos probabilísticos para artefactos.

Aprende un modelo bayesiano para la tasa de mutación general de SNV e indel y el espectro de fracción alélica del tumor para refinar las probabilidades logarítmicas emitidas por Mutect2. A continuación, establece automáticamente un umbral de filtrado para optimizar la puntuación F, la media armónica de sensibilidad y precisión.

Pipeline bioinformático en genómica del cáncer

Descubrimiento de variantes



Funcotator

Agregar información a nivel genético a cada variante. Funcotator es una herramienta de anotación funcional en el conjunto de herramientas principal de GATK y fue diseñado para manejar casos de uso somáticos y de línea germinal.

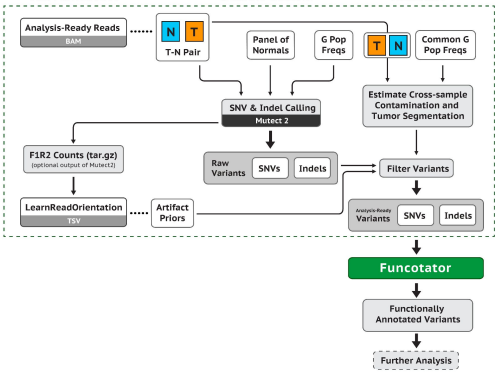
Funcotator lee un archivo VCF, etiqueta cada variante con una de las veintitrés clasificaciones de variantes distintas, produce información genética (por ejemplo, gen afectado, secuencia de aminoácidos variante predicha, etc.) y asociaciones con información en fuentes de datos.

Las fuentes de datos admitidas incluyen GENCODE (información genética y predicción de cambio de proteínas), dbSNP, gnomAD y COSMIC (entre otras).

Funcotator produce un archivo VCF (con anotaciones en el campo INFO) o un archivo de formato de anotación de mutación (MAF).

Pipeline bioinformático en genómica del cáncer

Descubrimiento de variantes



gnomAD es una gran base de datos de variantes conocidas. Este subconjunto contiene un subconjunto muy reducido de campos INFO, que contiene principalmente datos de frecuencia de alelos. gnomAD se divide en dos partes: una basada en los datos del exoma y otra basada en los datos del genoma completo. Estas dos fuentes de datos no son equivalentes y para una cobertura completa con gnomAD, recomendamos anotar con ambas.

gnomAD
genome aggregation database

Análisis de variantes

Las variantes se pueden analizar de diferentes maneras. Por ejemplo, es posible que desee determinar a qué genes afectan las variantes y qué efectos tienen sobre ellos. Para ello se pueden utilizar herramientas como **Ensembl VEP y SnpEff.**

Después de filtrar sus datos e identificar genes o variantes interesantes, puede explorarlos más a fondo utilizando bases de datos como:

Ensembl UniProt GWAS Catalog The NHGRI-EBI Catalog of published genome-wide association studies European Variation Archive

Estándares AMP/ASCO/CAP para clasificación de variantes somáticas

Las guías conjuntas de la Association for Molecular Pathology (AMP), la American Society of Clinical Oncology (ASCO) y el College of American Pathologists (CAP), publicadas en 2017, establecen un marco estandarizado fundamental para la interpretación clínica de variantes somáticas en oncología.

Marco de clasificación jerárquico en 4 niveles

Este sistema de clasificación representa un avance crucial en la medicina de precisión oncológica, estableciendo criterios uniformes para evaluar el significado clínico de las variantes somáticas detectadas en tumores. La clasificación se estructura en cuatro niveles jerárquicos (Tiers) que reflejan el grado de evidencia disponible y su aplicabilidad clínica directa.

El marco considera múltiples dimensiones de evidencia: terapéutica (respuesta a tratamientos específicos), diagnóstica (utilidad para clasificación tumoral), pronóstica (predicción de evolución clínica) y la calidad metodológica de los datos que respaldan cada interpretación.

Los cuatro niveles de clasificación clínica

Tier I – Variantes Accionables

Nivel A: Terapia aprobada por agencias regulatorias para ese tumor específico

Nivel B: Evidencia clínica sólida, ensayos fase I/II con resultados consistentes

Tier II – Potencial Clínico

Nivel C: Evidencia clínica establecida en otros tipos tumorales

Nivel D: Evidencia preclínica o datos clínicos limitados

Tier III – Significado Desconocido

Variantes de Significado Incierto (VUS): requieren más investigación para determinar su impacto clínico

Tier IV – Benignas

Variantes benignas o probablemente benignas, sin impacto clínico relevante conocido

Este esquema se detalla ampliamente en el artículo seminal del Journal of Molecular Diagnostics (2017, doi: 10.1016/j.jmoldx.2016.10.002) y en revisiones posteriores como Ouyang et al., 2022 (PMC9081216).

Criterios de evaluación para clasificación clínica

La clasificación AMP/ASCO/CAP se fundamenta en un análisis exhaustivo de múltiples criterios, diseñado específicamente para orientar la acción clínica en oncología. A diferencia de las guías ACMG/AMP 2015 enfocadas en patogenicidad de variantes germinales, este marco prioriza la utilidad clínica inmediata.

Evidencia terapéutica

- Respuesta a fármacos dirigidos
- Resistencia terapéutica documentada
- Resultados de ensayos clínicos

Evidencia diagnóstica

- Utilidad para clasificación tumoral
- Biomarcadores establecidos
- Valor en diagnóstico diferencial

Evidencia pronóstica

- Predicción de evolución clínica
- Estratificación de riesgo
- Supervivencia global y libre de enfermedad

01

Datos funcionales

Estudios in vitro e in vivo que demuestren el efecto biológico de la variante

02

Frecuencia poblacional

Análisis de cohortes tumorales y patrones mutacionales recurrentes

03

Bases de conocimiento

Reportes previos en ClinVar, COSMIC, OncoKB, CiViC y literatura científica

Clasificación funcional de variantes a nivel proteico

Las anotaciones de SnpEff/VEP clasifican las variantes según el efecto molecular sobre la proteína:

1 Missense (cambio de aminoácido)

- Una base cambia → cambia un aminoácido
- Ejemplo: p.Gly12Asp (KRAS G12D)
- Impacto: MODERATE
- Efecto: alteración de función, activación/inhibición, hotspot oncogénico

2 Nonsense / Stop-gained

- Introduce un codón de parada prematuro
- Ejemplo: p.Arg213*
- Impacto: HIGH
- Efecto: pérdida de función, truncamiento → frecuente en genes supresores (TP53, BRCA1)

3 Frameshift

- Inserción/deleción que altera el marco de lectura
- Ejemplo: p.Lys100fs
- Impacto: HIGH
- Efecto: proteína anómala o truncada tempranamente

4 Silent / Synonymous

- Cambio de nucleótido sin alterar aminoácido
- Ejemplo: p.Leu50=
- Impacto: LOW
- Efecto: No altera el aminoácido, pero ocasionalmente puede afectar el splicing

Variantes que afectan el splicing

Las variantes que afectan el splicing son críticas para la función génica:

1

Splice donor (+1, +2)

Impacto: HIGH

Efecto: rompe unión exón–intrón

Ejemplo: c.425+1G>A → afecta sitio donador → potencial pérdida de función

2

Splice acceptor (-1, -2)

Impacto: HIGH

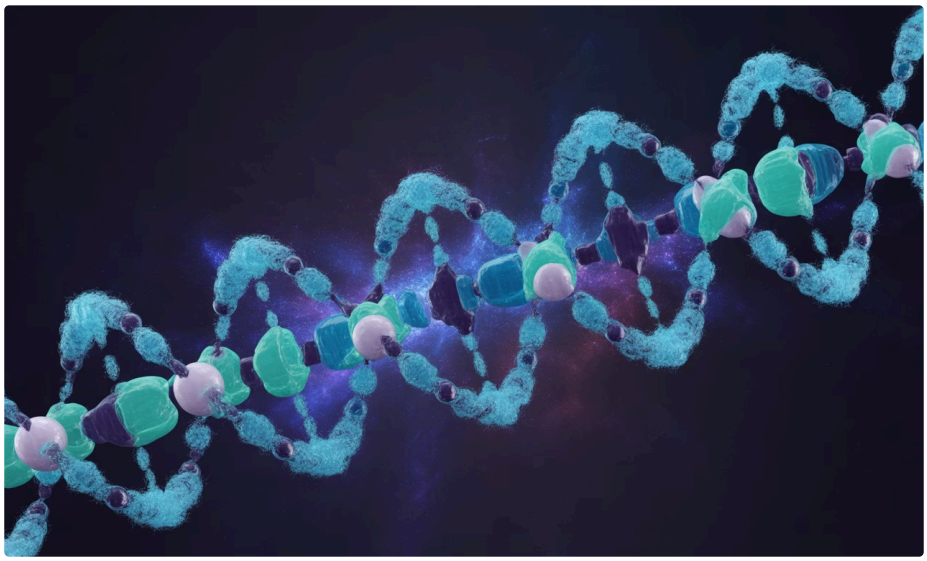
Efecto: afecta ensamblaje exacto de exones

3

Splice region (±3–8)

Impacto: MODERATE/LOW

Efecto: posible efecto, depende del contexto



Relación entre tipo de variante y su acción clínica

La clasificación AMP/ASCO/CAP integra el tipo de variante con su relevancia clínica:

Nonsense / Frameshift / Splice críticos

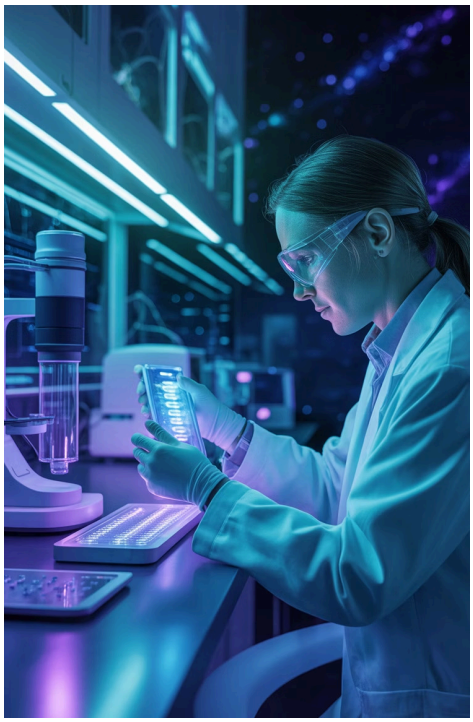
- Frecuentes en genes supresores
- Suelen clasificarse en Tier I/II si hay evidencia clínica
- Mecanismo: pérdida de función

Missense

- Frecuentes en oncogenes
- Hotspots (p.ej., BRAF V600E, KRAS G12C) → Tier I (A/B)
- Mecanismo: ganancia de función o cambio de actividad

Sinónimas

- Generalmente Tier IV, salvo evidencia de efecto en splicing



Ejemplos clásicos según el tipo de variante

Ejemplos reales de oncología clasificados por tipo de variante:

Missense – Activadoras en oncogenes

- BRAF p.V600E → melanoma, tiroides, colon
- KRAS p.G12D/G12C → resistencia anti-EGFR
- EGFR p.L858R → sensibilidad a TKI

Nonsense / Stop-gained – Inactivadoras

- TP53 p.R196* → pérdida completa de función
- BRCA1 p.Y1853* → riesgo hereditario + sensibilidad a PARP

Frameshift

- BRCA2 p.S1982fs → deleción → pérdida de función
- PTEN p.K267fs → tumor suppressor inactivation

Splice

- MET exon 14 skipping → activación oncogénica (importante para TKIs)

Cómo identificar tipos de variantes en el campo ANN del VCF

En el VCF anotado por SnpEff o VEP, el campo ANN= contiene la información del tipo de efecto:

Formato general:

```
ANN=<ALT>|<tipo_efecto>|<impacto>|<gen>|...|<proteína>
```

Ejemplos prácticos:

Missense:

```
ANN=T|missense_variant|MODERATE|BRAF|...|p.Val600Glu
```

Stop-gained:

```
ANN=A|stop_gained|HIGH|TP53|...|p.Arg213*
```

Frameshift:

```
ANN=G|frameshift_variant|HIGH|BRCA1|...|p.Lys50fs
```

Splice:

```
ANN=G|splice_acceptor_variant|HIGH|MET|...|
```

Resumen: tipos de variantes en interpretación del cáncer

Puntos clave del módulo:

La clasificación de variantes (missense, stop, frameshift, splice) define el mecanismo molecular esperado.

1 Patrones principales

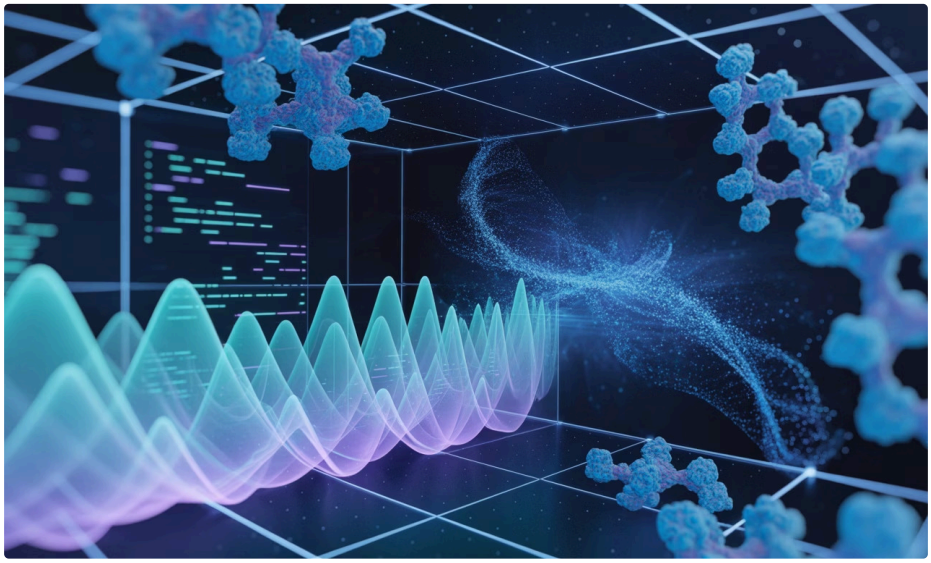
- **Oncogenes:** generalmente missense activadoras en hotspots
- **Genes supresores:** generalmente stop/frameshift/splice → pérdida de función

2 Integración clínica

- Estos tipos se integran dentro del esquema Tier I–IV de AMP/ASCO/CAP
- Las anotaciones HGVS permiten buscar variantes en herramientas como CancerVar, CACSV, OncoKB

3 Aplicación práctica

- El tipo de variante orienta sobre el mecanismo biológico
- La evidencia clínica determina la clasificación final
- La nomenclatura estandarizada es clave para la búsqueda precisa



Herramientas de clasificación automatizada

Software principal

CancerVar

Implementación integral que combina criterios AMP/ASCO/CAP con ACMG y algoritmos computacionales avanzados. Genera reportes clínicos completos incluyendo evidencia terapéutica, diagnóstica y pronóstica. Interfaz web disponible y pipeline descargable.

CACSV

Clinical Annotation of Cancer Somatic Variants aplica directamente la clasificación AMP/ASCO/CAP. Disponible como API programática y pipeline integrable en workflows bioinformáticos institucionales.

Bases de conocimiento esenciales

OncoKB

Clasifica niveles de evidencia terapéutica con estructura análoga a Tiers I–IV. Base de referencia del Memorial Sloan Kettering Cancer Center.

CiViC

Curación comunitaria de evidencia clínica. Modelo abierto y colaborativo con revisión por expertos.

JAX-CKB

Especializado en terapias dirigidas y biomarcadores. Actualizaciones continuas del Jackson Laboratory.

Recomendación práctica: La integración de múltiples bases de datos mejora la sensibilidad en la detección de variantes accionables y reduce falsos negativos en la interpretación clínica.

Nomenclatura estandarizada: clave para búsqueda precisa

La búsqueda e interpretación correcta de variantes en bases de datos especializadas depende críticamente del uso de nomenclatura estandarizada. Las herramientas bioinformáticas requieren formatos específicos para identificar inequívocamente las variantes y recuperar toda la evidencia clínica disponible.

Nivel de DNA (c.)

Notación de la variante a nivel de cDNA según HGVS

Ejemplo: c.1799T>A

Nivel de proteína (p.)

Cambio aminoacídico resultante

Ejemplo: p.Val600Glu (BRAF V600E)

Coordenadas genómicas

Posición exacta en genoma de referencia (GRCh37/GRCh38)

Ejemplo: chr7:140453136 A>T

Identificadores rsID

Número de referencia en dbSNP

Ejemplo: rs113488022

Nota técnica importante: CancerVar, CACSV, OncoKB y ClinVar optimizan sus resultados cuando se utilizan coordenadas genómicas precisas combinadas con nomenclatura HGVS completa. Una misma variante puede describirse de múltiples formas; la estandarización previene pérdida de información clínica relevante.

Análisis en la nube con TERRA

¿Por qué trabajar en la nube?

1. Para analizar y compartir muchos datos
   a. Facilita colaboraciones internacionales
   b. Buena alternativa cuando no se cuenta con servidor local
2. Permite probar diferentes herramientas ya instaladas (Docker)
3. Hace que los estudios sean reproducibles

¿Qué es TERRA?

Es una plataforma escalable genómica diseñada para realizar todos sus procesos en la nube.
Contiene:

1. Recursos: Librerías de datos y métodos
2. Disponibilidad: Datos, métodos y espacio de trabajos se pueden compartir.
3. Análisis en tiempo real: Los resultados de un análisis pueden ser manipulados y organizados desde TERRA usando Jupyter notebook.

¿Qué es TERRA?

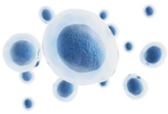
42,000 TERRA USERS

2.38 MILLION STUDY PARTICIPANTS

39.54 MILLION SINGLE CELLS

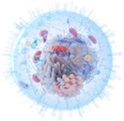
Terra supports researchers in many biomedical disciplines

Cancer Genomics



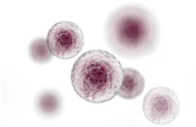
The Van Allen Lab is using Terra to advance clinical oncology through cancer genomics.

Single-Cell Transcriptomics



The Shalek Lab is using Terra to improve the scalability, accessibility, and reproducibility of single-cell analysis.

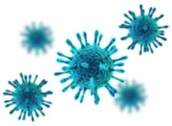
Medical and Population Genetics



The Natarajan Lab is using Terra to study genetic factors of heart diseases.

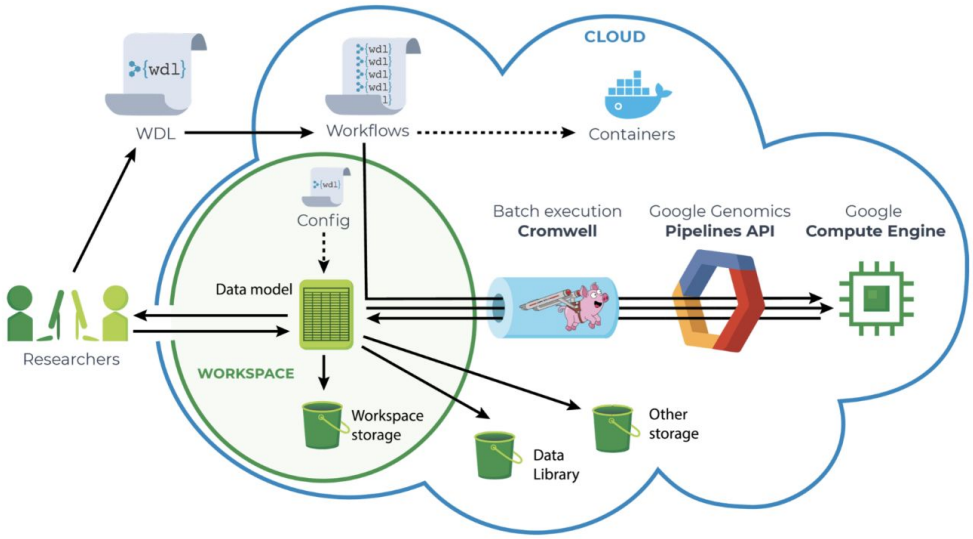
Read More

Infectious Diseases

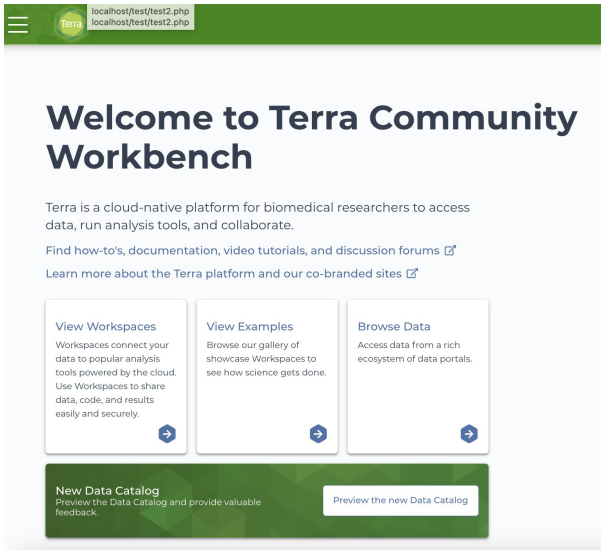


Terra is being used to advance genomic epidemiology and surveillance of pathogens

Análisis en la nube con TERRA



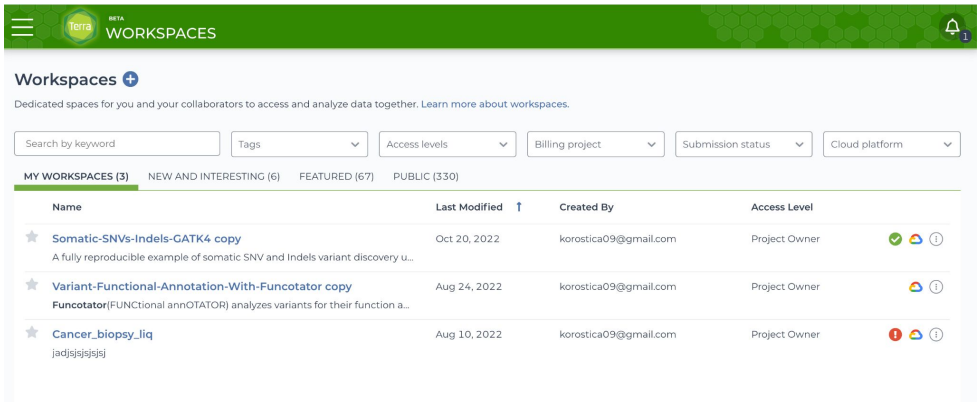
¿Qué es TERRA?



Creamos y hacemos login con nuestra cuenta

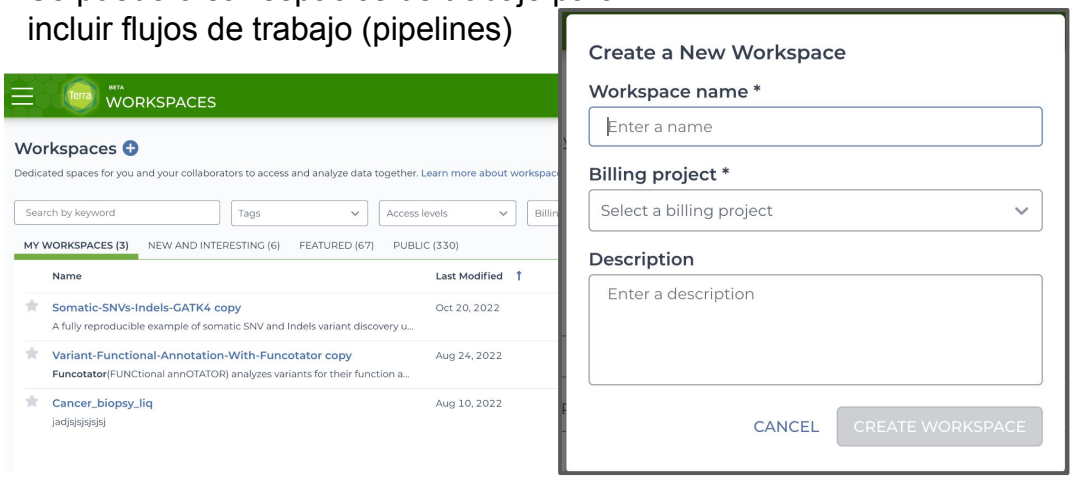
¿Qué es TERRA?

Se puede crear espacios de trabajo para incluir flujos de trabajo (pipelines)



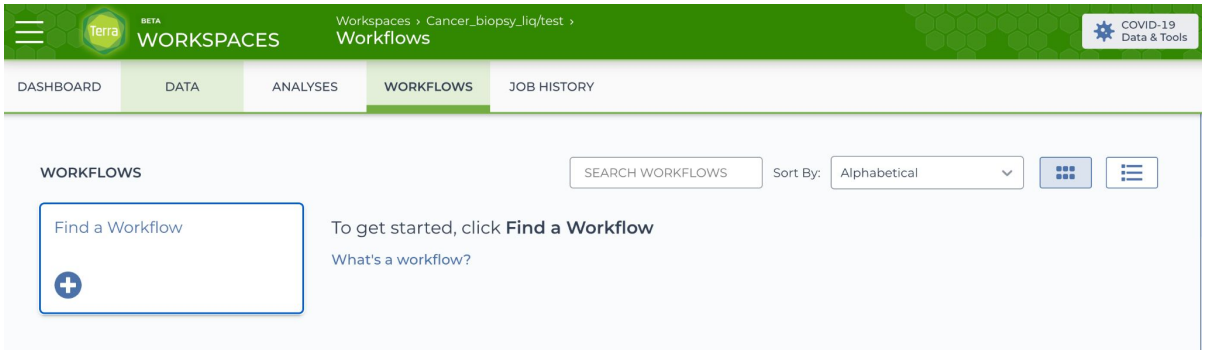
¿Qué es TERRA?

Se puede crear espacios de trabajo para incluir flujos de trabajo (pipelines)



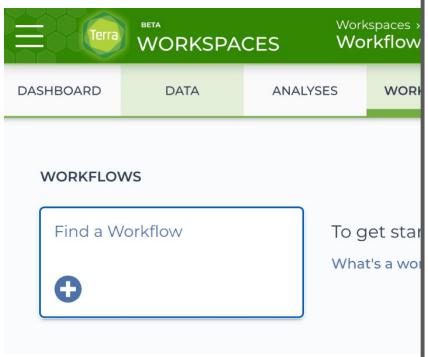
¿Qué es TERRA?

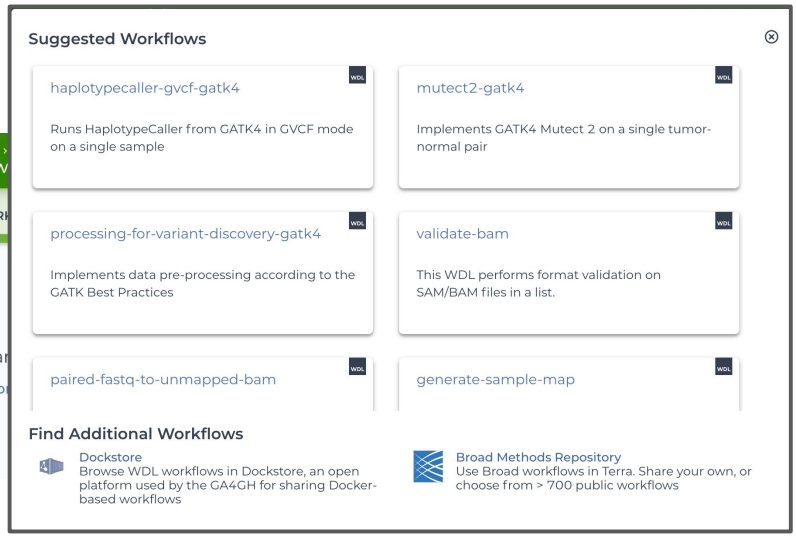
Búsqueda de workflows



¿Qué es TERRA?

Búsqueda de workflows

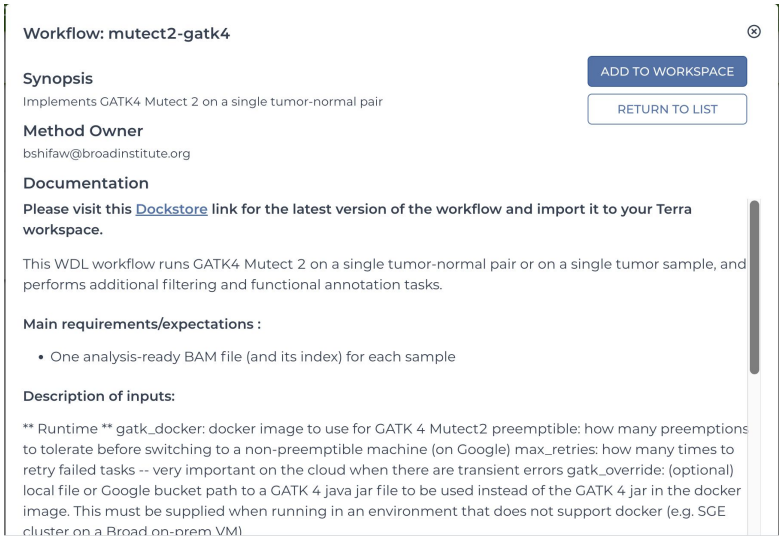




¿Qué es TERRA?

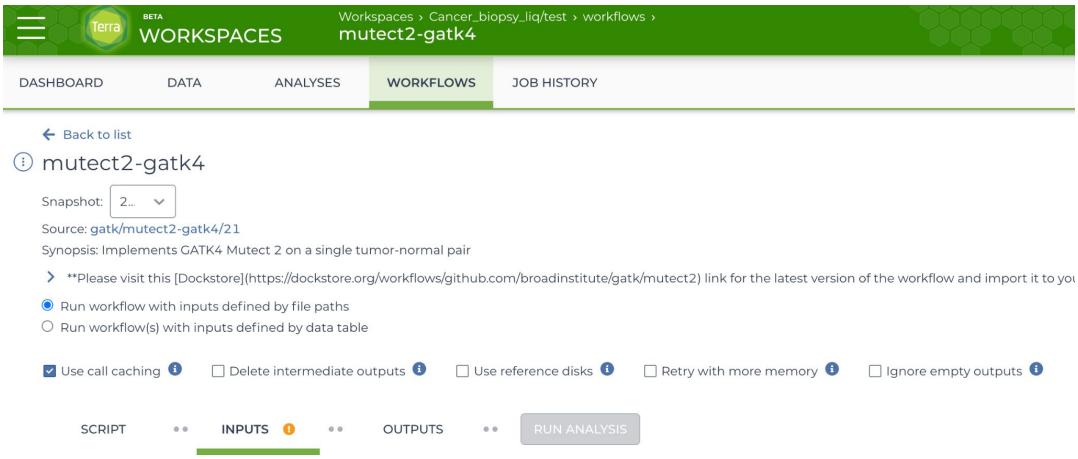
Descripción de workflows

El cual se agregará a mi espacio de trabajo



¿Qué es TERRA?

Descripción de workflows



¿Qué es TERRA?

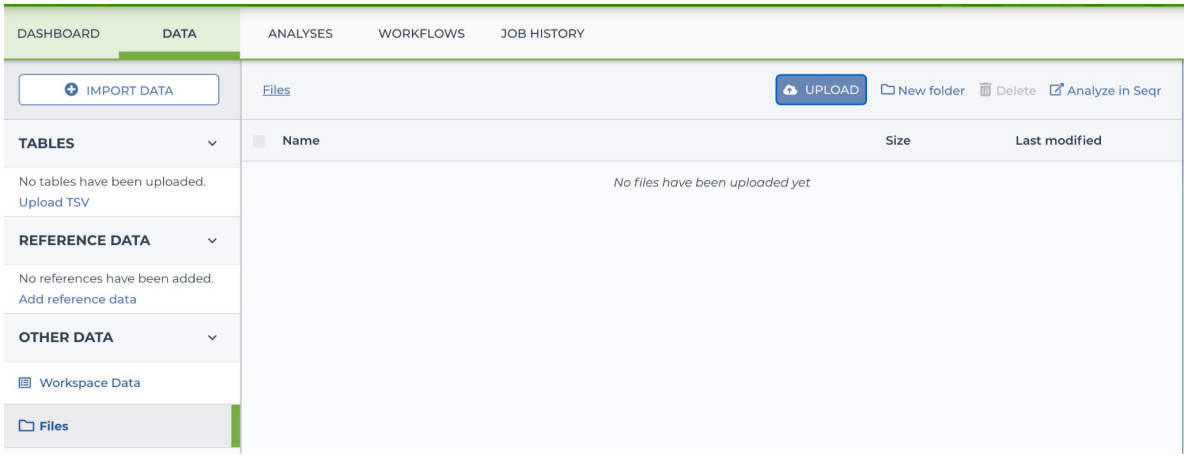
Edición de los input (datos parámetros, bases de datos)

Hide optional inputs

Task name	Variable	Type	Attribute
Mutect2	gatk_docker	String	Required
Mutect2	ref_dict	File	Required
Mutect2	ref_fai	File	Required
Mutect2	ref_fasta	File	Required
Mutect2	scatter_count	Int	Required
Mutect2	tumor_reads	File	Required
Mutect2	tumor_reads_index	File	Required
CalculateContamination	*intervals*	*String*	Optional

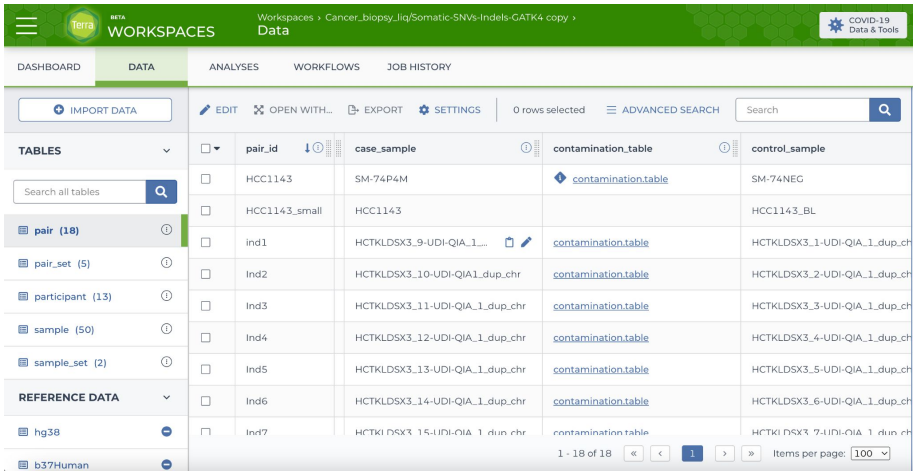
¿Qué es TERRA?

Upload de datos propios



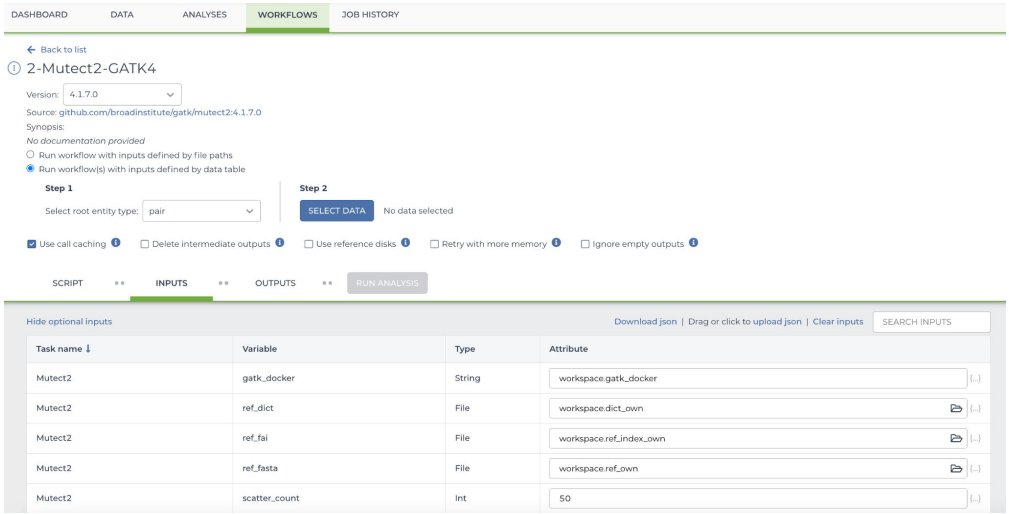
¿Qué es TERRA?

Upload de datos propios



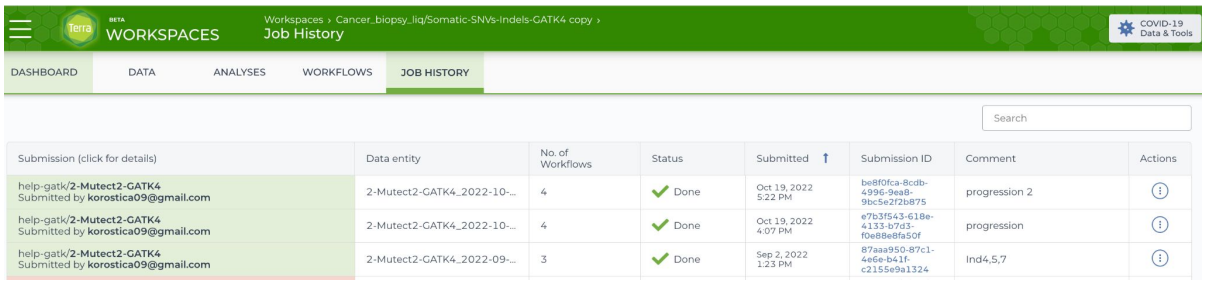
¿Qué es TERRA?

Upload de datos propios



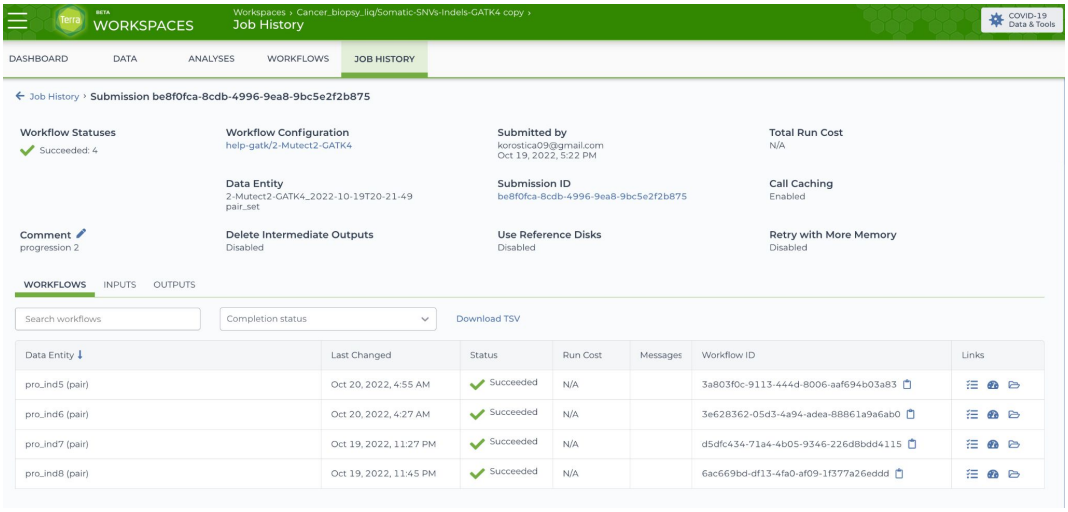
¿Qué es TERRA?

Lanzamiento de trabajo

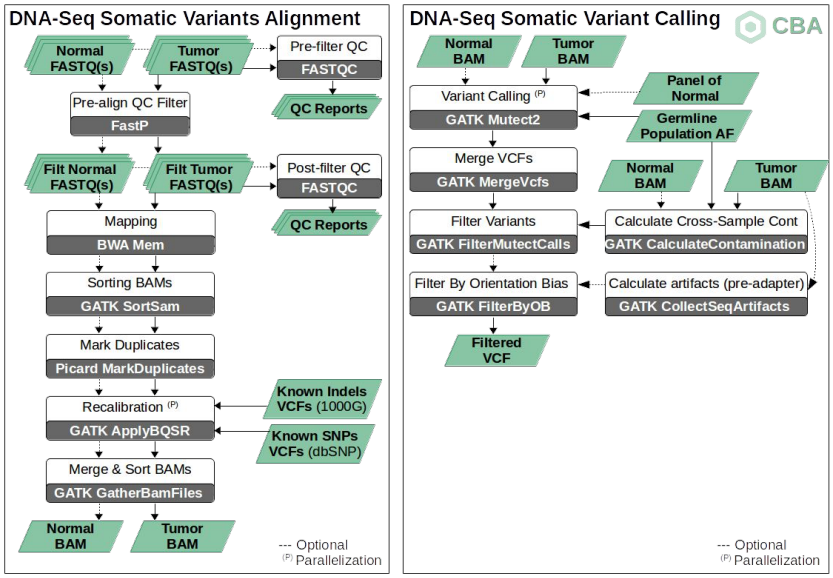


¿Qué es TERRA?

Lanzamiento de trabajo



Gracias por su asistencia

Karen Oróstica, korostica09@gmail.com / korostica@udd.cl
Math-Bio Lab
Unidades de investigación/prácticas/tesis disponibles

Pipeline bioinformático en genómica del cáncer