Instituto Nacional de Medicina Genómica

Aprende a utilizar los flujos de análisis bioinformáticos internos del Inmegen

Identificación automatizada de variantes de datos de RNA-seq

Profesores: Dra. Alejandra Cervera
Dra. Laura Gómez
Dr. Daniel Pérez

nextflow

Repaso clase anterior

1. ¿Cuales son los principales pasos para correr un flujo de trabajo de NextFlow?
2. ¿Dudas del pipeline Q&DEA RNAseq?

Repositorio de la clase

https://github.com/INMEGEN/Clase_pipelines/tree/main

Temario del curso

Instituto Nacional de Medicina Genómica

- Repaso de Bash
- Introducción a Nextflow y Docker
- Cuantificación y análisis de expresión diferencial
- **Identificación automatizada de variantes de datos de RNA-seq**
- Identificación automatizada de variantes germinales
- Identificación automatizada de variantes somáticas

Instituto Nacional de Medicina Genómica

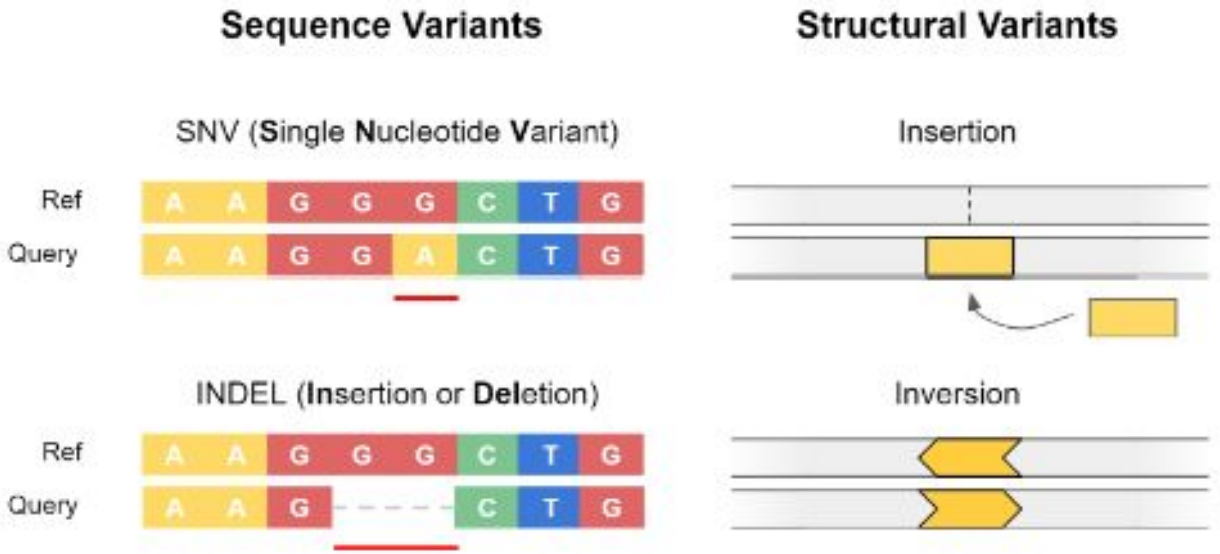
Identificación automatizada de variantes de datos de RNA-seq

¿Qué es una variante genética?

El Instituto Nacional del Cáncer (NIH) de Estados Unidos define una variante como:

Una alteración en la secuencia más común de nucleótidos del ADN. El término variante se usa para describir una alteración que puede ser benigna, patógena o de repercusión incierta. Este término se usa cada vez más en lugar del término mutación. También se llama variación de secuencia, variación genética, variante de secuencia y variante genética.

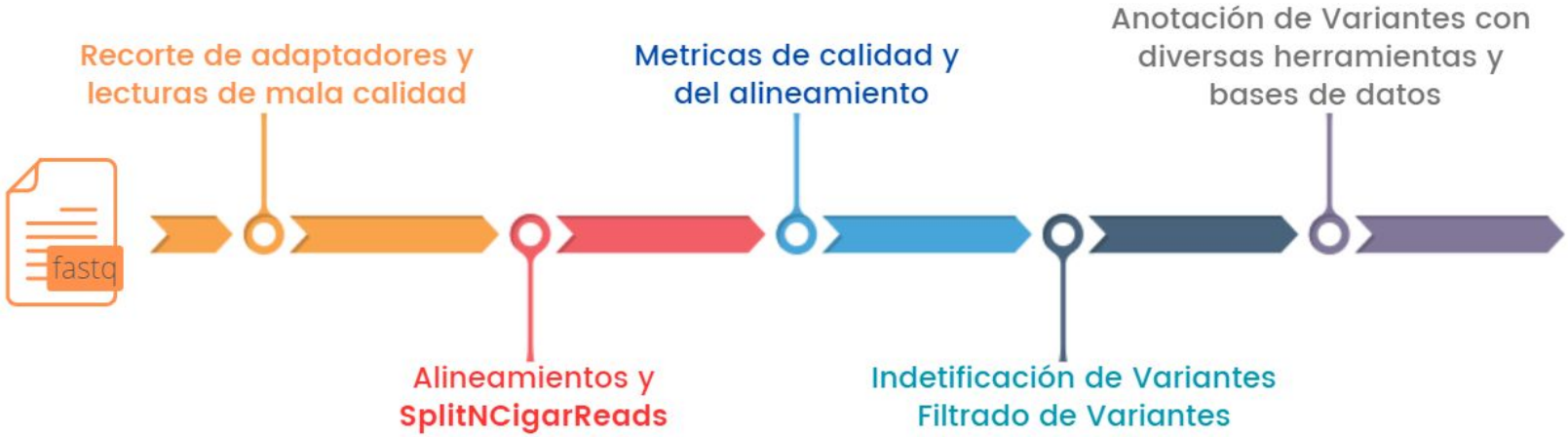
Variantes


Sequence Variants
Structural Variants
SNV (Single Nucleotide Variant)
Ref
Query
A A G G G C T G
A A G G A C T G
Insertion
INDEL (Insertion or Deletion)
Ref
Query
A A G G G C T G
A A G C T G
Inversion

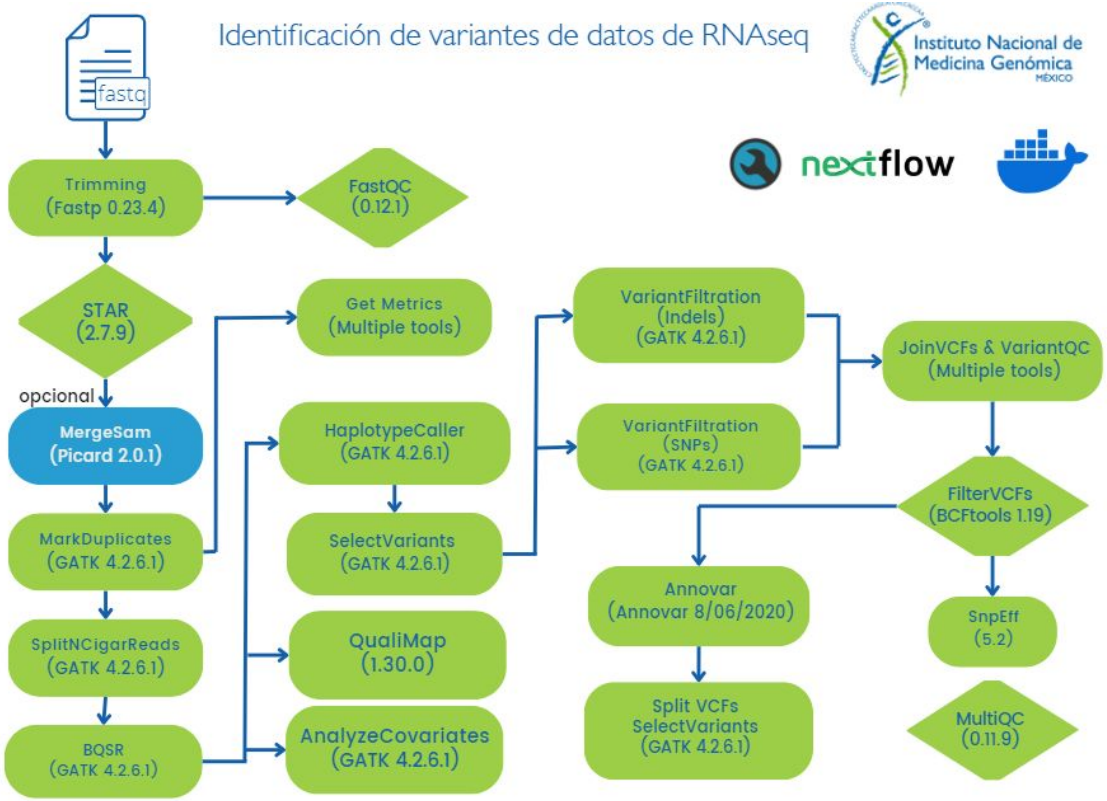
Identificación automatizada de variantes de datos de RNA-seq (VC-RNAseq)

Instituto Nacional de Medicina Genómica

Flujo de trabajo experimental de GATK
https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels



Herramientas del pipeline VC-RNAseq



Identificación automatizada de variantes de datos de RNA-seq (VC-RNAseq)

Instituto Nacional de Medicina Genómica

Exploren el repositorio

https://github.com/INMEGEN/Pipelines_INMEGEN/tree/Principal

Particularmente el directorio **VC-RNAseq**

¿Qué elementos del código de NextFlow reconocen?

¿Qué cambia con respecto a Q&DEA?

Archivo sample_info.tsv

Instituto Nacional de Medicina Genómica

Este archivo contiene:

- Nombre de la muestra `(Sample_name)`
- Identificador de la muestra `(SampleID)`
- Barcode de la flowcell y número de lane. `(RG_PU)`
- tenología de secuenciació `(RG_PL)`
- Barcode de la librería de secuenciación `(RG_LB)`
- Ruta absoluta al archivo R1 `(R1)`
- Ruta absoluta al archivo R2 `(R2)`

Ejemplo:

```
Sample_name      SampleID   RG_PU RG_PL RG_LB R1    R2
IDS1  ID_L1 FLOWCELL.1 ILLUMINA   BARCODE    Path/to/fastq_S1_R1.fastq  Path/to/fastq_S1_R2.fastq
IDS2  ID_L2 FLOWCELL.2 ILLUMINA   BARCODE    Path/to/fastq_S2_R1.fastq  Path/to/fastq_S2_R2.fastq
```

Archivo nextflow.config

Contiene los parámetros necesarios para correr el flujo de trabajo.

Los parámetros que es imprescindible editar son:

- Directorio de salida
- Nombre del proyecto
- Si las muestras se distribuyeron en diferentes carriles (lanes)
- Ruta absoluta y nombre del índice de STAR
- Nombre del archivo GTF
- Ruta absoluta de los archivos necesarios para BQSR
- Ruta absoluta de las bases de datos de annovar

Ejercicio

Seguir las instrucciones del repositorio para correr el flujo de trabajo de NextFlow que se encuentra en el directorio VC-RNAseq

¿dudas?