1. Problema

Actualmente, las herramientas comúnmente utilizadas para detectar CNVs en los genomas humanos son costosas desde el punto de vista computacional (en memoria y complejidad temporal). Por ello, nosotros proponemos una solución de bajo consumo de recursos, logarítmica y probabilística, como una forma de análisis preventivo —no como un reemplazo completo— para la detección de CNVs antes del uso de herramientas exactas pero más costosas.

2. Conceptos importantes

- **Genoma:** Conjunto completo de ADN de un organismo. En humanos son aproximadamente **3 mil millones de pares de bases** (letras A, T, G, C).
- **Cobertura:** Número promedio de veces que cada base del genoma es leída por la secuenciación. **Alta cobertura** implica mayor confianza en las variantes detectadas. **Cobertura desigual** puede indicar **CNVs**.
- **BAM file:** Formato de archivo **binario** que contiene **lecturas (reads) alineadas** al genoma de referencia.
- **CNV (Copy Number Variation):** Tipo de **variante estructural** donde una región del genoma tiene un **número de copias diferente al normal** (2 en organismos diploides).
- **Bin: Ventana o segmento del genoma** de tamaño fijo (por ejemplo, 1 kb o 10 kb) usada para **agrupar lecturas o datos de cobertura**.
- **Quantile sketch: Estructura de datos probabilística** que **resume una distribución de valores**, permitiendo **consultas aproximadas de cuantiles** con **memoria constante**.

3. Propuesta de Proyecto

El objetivo de este trabajo es detectar **variaciones en el número de copias** (CNVs) en genomas humanos a partir de la **distribución de coberturas de secuenciación**, utilizando estructuras de datos compactas denominadas *quantile sketches*.

En secuenciación masiva, cada fragmento de ADN del genoma es leído múltiples veces por el secuenciador. Estas lecturas (*reads*) se almacenan en archivos `FASTQ` y posteriormente se alinean contra un genoma de referencia (`FASTA`) mediante un alineador como BWA o Bowtie. El resultado es un archivo `BAM`, que contiene la posición exacta de cada lectura en el genoma y permite calcular la **cobertura**.

La **cobertura** representa cuántas veces cada posición del genoma fue leída. Si el genoma se secuenció a una profundidad de 30×, significa que, en promedio, cada base fue leída 30 veces. Sin embargo, esta medida global no es suficiente para detectar variaciones estructurales, ya que las duplicaciones y deleciones afectan solo regiones específicas del genoma.

3.1. Cobertura local y detección de CNVs

Para analizar la variación local, el genoma se divide en ventanas de tamaño fijo llamadas **bins** (por ejemplo, de 1 kb). Para cada bin se calcula su cobertura promedio, es decir, el número total de bases de lecturas que caen dentro del bin dividido por su tamaño. Regiones con cobertura significativamente mayor o menor que la mediana global pueden indicar **duplicaciones** o **deleciones**, respectivamente.

$$C_i = \frac{\text{bases leídas dentro del bin}_i}{\text{tamaño del bin}} \tag{1}$$

Donde C_i es la cobertura del bin i. A partir de todos los C_i del cromosoma, se obtiene una distribución de coberturas $\{C_1, C_2, \ldots, C_n\}$.

3.2. Uso de quantile sketches

Guardar la cobertura de millones de bins puede ser costoso en memoria. Por ello, se utilizarán **quantile sketches** (KLL), que permiten construir una representación compacta de la distribución de coberturas sin almacenar todos los valores. Estos sketches permiten consultas como percentiles o cuantiles de forma aproximada, con errores acotados.

- El **percentil 50 (p50)** representa la cobertura típica del cromosoma.
- Los percentiles altos (p95, p99) indican bins con cobertura elevada → posibles duplicaciones.
- Los percentiles bajos (p5, p1) indican bins con cobertura reducida → posibles deleciones.

A partir de los cuantiles obtenidos se pueden definir indicadores simples para detectar anomalías en la distribución de coberturas. La mediana (p50) representa la cobertura típica del cromosoma, mientras que las colas (p5, p95) reflejan desviaciones locales.

Un desplazamiento pronunciado de la cola superior ($p95 \gg p50$) sugiere la existencia de regiones con cobertura anormalmente alta, indicativas de **duplicaciones**, mientras que un desplazamiento de la cola inferior ($p5 \ll p50$) indica **deleciones**.

En este trabajo se generará un **sketch por cromosoma**, de modo que cada distribución refleje el comportamiento de las coberturas dentro de ese cromosoma en particular.

El análisis no busca determinar la posición exacta de las duplicaciones o deleciones, sino detectar su **presencia o evidencia estadística** a partir de la forma de la distribución de coberturas.
Esto permite identificar cromosomas que presentan señales de CNVs sin requerir un análisis base por base ni almacenar todos los datos.
Estos indicadores permiten evaluar la existencia de duplicaciones o deleciones sin almacenar todas las coberturas individuales, sino únicamente un resumen probabilístico de su distribución.

3.3. Resumen del flujo de trabajo

1. **Entrada:** archivo `BAM` con lecturas alineadas al genoma.
2. **Procesamiento:**

- Leer lecturas por cromosoma.
- Dividir el cromosoma en bins de tamaño fijo (1 kb).
- Contar cuántas lecturas cubren cada bin.
- Calcular la cobertura promedio por bin.
- Insertar cada cobertura en un quantile sketch (KLL).

3. **Salida:**
   - Distribución de coberturas por cromosoma (percentiles).
   - Identificación de cromosomas con patrones anómalos (presencia de CNVs).

3.4. Resultados esperados

El sistema debe permitir identificar cromosomas cuya distribución de coberturas difiera de la distribución típica.

La interpretación se centrará en la **detección global por cromosoma**, es decir, determinar si un cromosoma muestra evidencia estadística de duplicaciones o deleciones, sin precisar su ubicación exacta.

Se espera observar desplazamientos en los cuantiles o colas de la distribución asociados a duplicaciones o deleciones, demostrando que las estructuras de resumen probabilístico (sketches) pueden utilizarse como una alternativa eficiente para la detección preliminar de CNVs.

$$\text{RankError}(r) = \left|\hat{r} - r\right| \leq \frac{c}{\sqrt{K}}$$

donde:

- r es el rank verdadero (posición relativa en el conjunto ordenado).
- $\hat{r}$ es el rank estimado por el sketch.
- K es el parámetro del KLL sketch.
- c es una constante del algoritmo (aprox. 1).