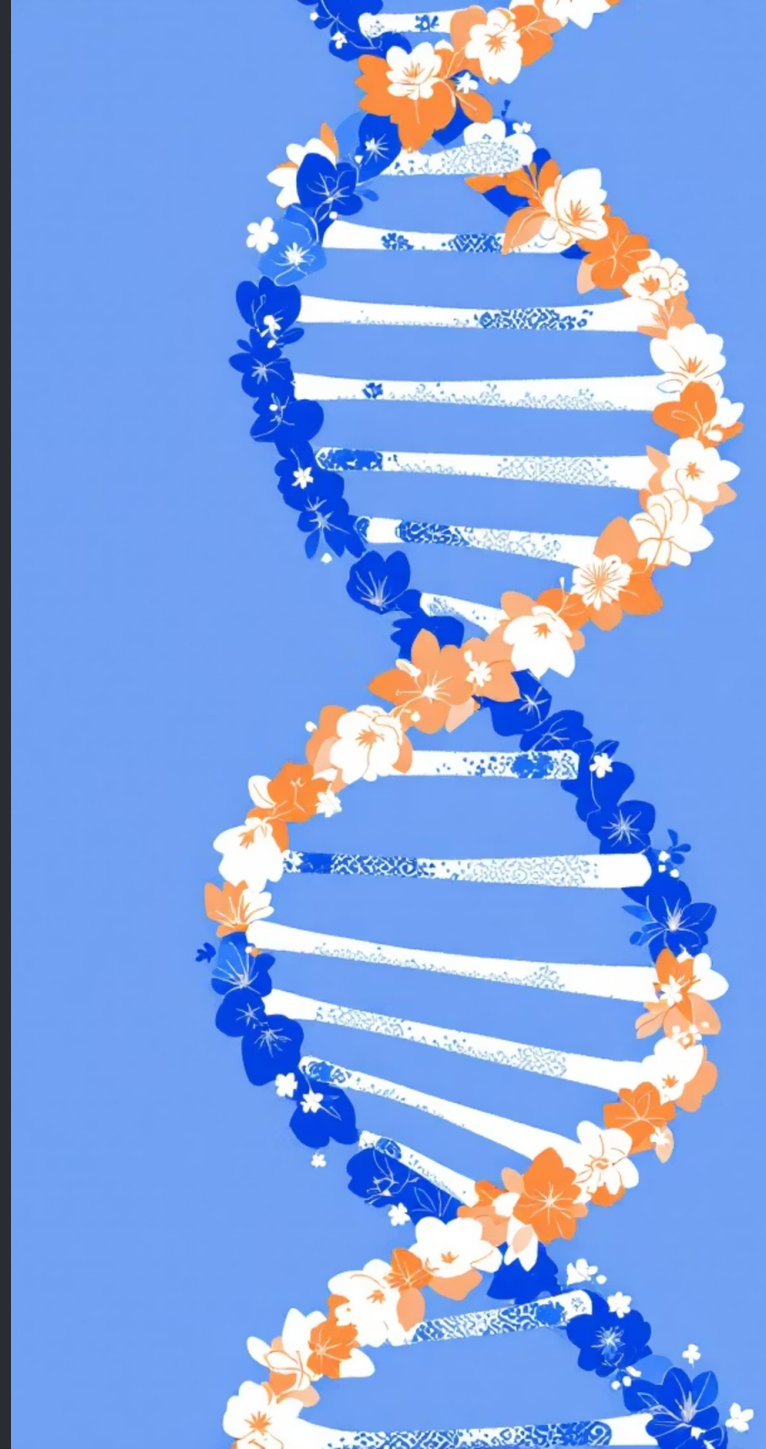
Detección Eficiente de CNVs con KLL Sketches

Análisis de Cobertura Genómica en Streaming

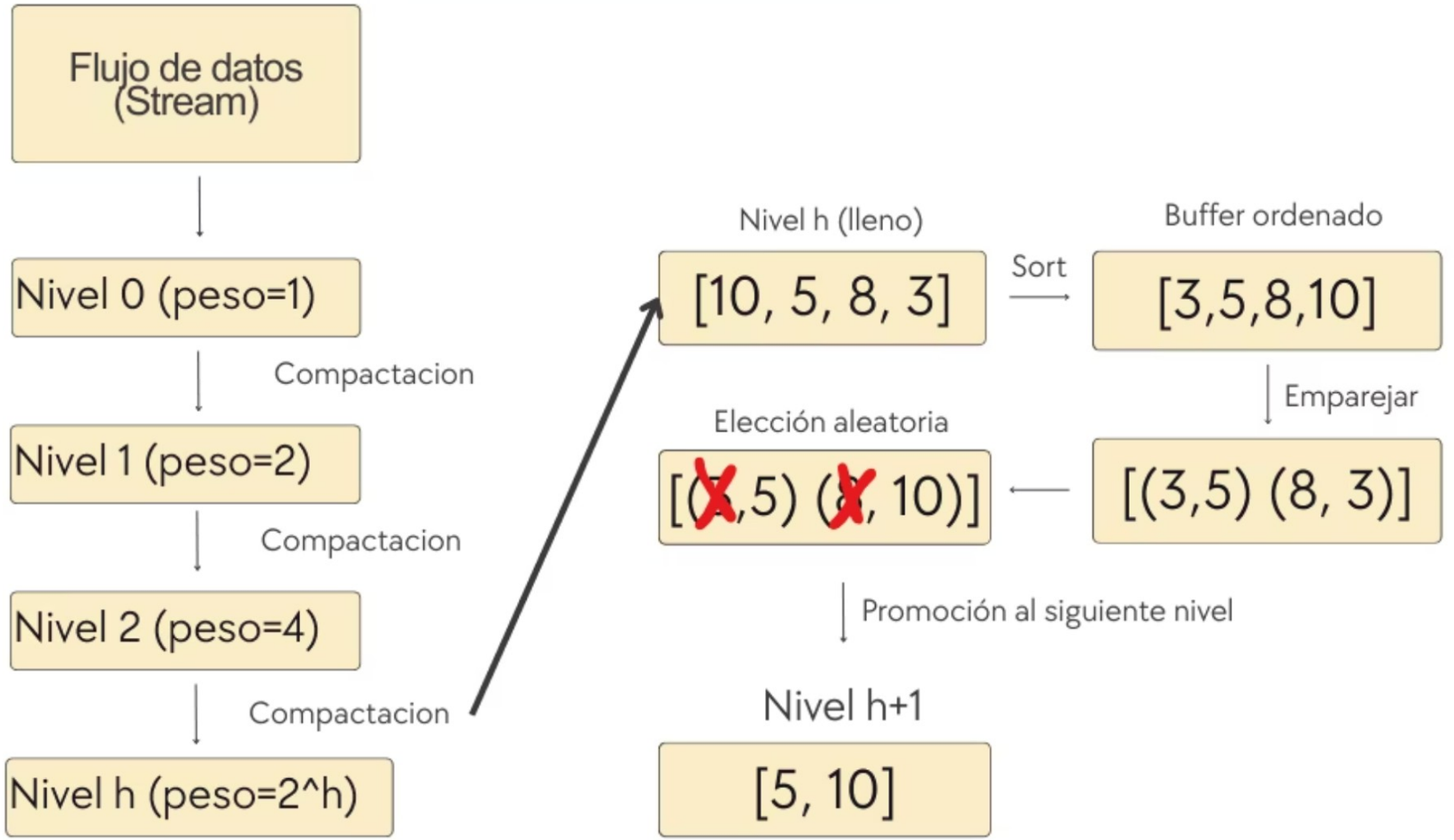


El Desafío del Análisis de Cobertura Genómica

El análisis de datos genómicos a gran escala presenta retos significativos. Los archivos BAM, que almacenan información de secuenciación, pueden alcanzar tamaños de 40 a 150 GB por genoma, haciendo inviable el procesamiento con métodos tradicionales para grandes cohortes.

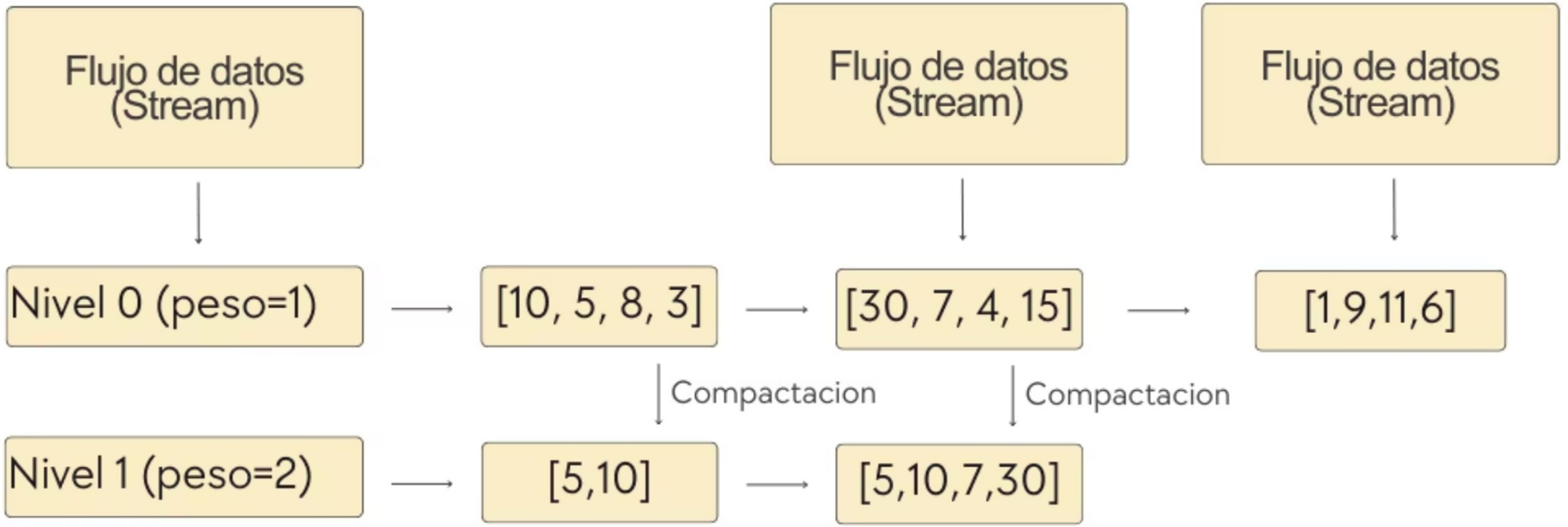
- Archivos BAM voluminosos.
- Necesidad de calcular estadísticas de cobertura (cuantiles).
- Métodos tradicionales: carga completa de datos en memoria. Costosos computacionalmente

Arquitectura KLL



Cada vez que se compactan datos se eliminan los datos del nivel en el que se hizo la compactación

Estado del KLL



Consulta de cuantiles

Una vez finalizada la lectura de datos, se combinan todos los elementos almacenados en los diferentes buffers en una lista única, manteniendo el peso asociado a cada valor

Nivel 0 (peso=1) ⟶ [1,9,11,6]

Nivel 1 (peso=2) ⟶ [5,10,7,30]

par.first es el valor y par.second es el peso asociado a ese valor

[(1,1),(9,1),(11,1),(6,1),(5,2),(10,2),(7,2),(30,2)]

Sort basado únicamente en el valor (par.first)

[(1,1),(5,2),(6,1),(7,2),(9,1),(10,2),(11,1),(30,2)]

Masa total de Datos: $N = \sum_{h=0}^{L-1} \sum_{i=1}^{M_h} W_h$

Table 1: Lista Combinada y Ordenada de Elementos del Sketch KLL

Orden Final	**Valor**	**Peso (W)**	**Fuente Original**
1.	1	1	Nivel 0
2.	5	2	Nivel 1
3.	6	1	Nivel 0
4.	7	2	Nivel 1
5.	9	1	Nivel 0
6.	10	2	Nivel 1
7.	11	1	Nivel 0
8.	30	2	Nivel 1

Table 2: Cálculo de Rango Acumulado y Determinación de la Mediana

Orden Final	**Valor**	**Peso (W)**	**Peso Acumulado (Rango)**	**¿Rango $\geq$ 6?**
1.	1	1	1	No
2.	5	2	3	No
3.	6	1	4	No
4.	**7**	**2**	**6**	**Sí**
5.	9	1	7	Sí
6.	10	2	9	Sí
7.	11	1	10	Sí
8.	30	2	12	Sí

Masa Total de Datos (N) = 12. Rango Objetivo (Mediana P50) = 6.
*Resultado: El cuantil se alcanza en el **Valor 7**.*

Metodología e Implementación: El Pipeline de Análisis

Desarrollamos un pipeline eficiente que integra KLL Sketches para procesar datos de secuenciación genómica y detectar variaciones en la cobertura.

Dividir Genoma

En bins de 1-10 kb.

Leer BAM

En streaming, contando reads por bin.

Insertar Cobertura

En el KLL Sketch.

Consultar Cuantiles

(P50, P95, P99).

Detectar Anómalos

Bins con posibles CNVs.

Tecnología Utilizada

Apache DataSketches C++

Librería robusta para sketches.

HTSlib

Lectura eficiente de archivos BAM.

```
kll_sketch coverage(200);
for (auto& bin : bins) {
 coverage.update(bin.coverage);
}
float median = coverage.get_quantile(0.5);
```

Resultados del Análisis de Cobertura

Nuestro pipeline procesó un archivo BAM de 40 GB, demostrando eficiencia y precisión en la cuantificación de la cobertura genómica.

Archivo BAM: 40 GB

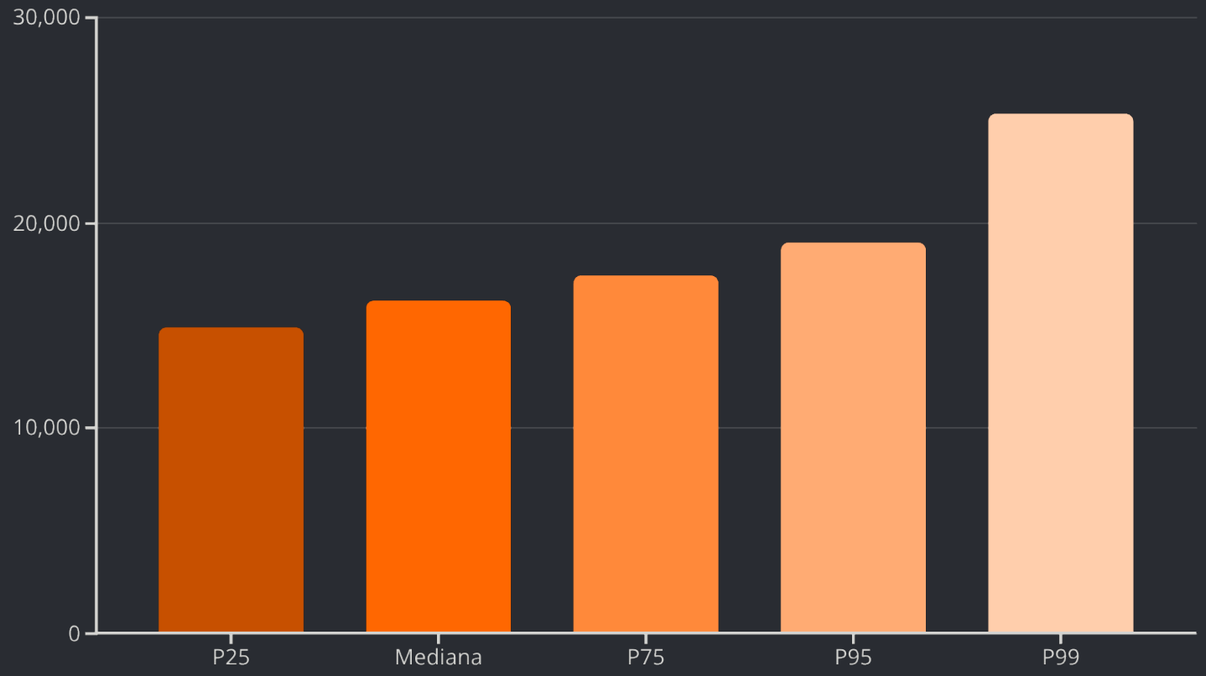
Total Reads: 495,903,424

Bins Analizados: 458,397 (tamaño 1kb)

Tiempo de Procesamiento: 7 minutos

Reads Mapeados: 98.38% (Excelente calidad)

Los cuantiles calculados con KLL Sketches revelan la distribución de la cobertura a lo largo del genoma, identificando regiones con variaciones significativas.



Detección de Variantes Estructurales (CNVs)

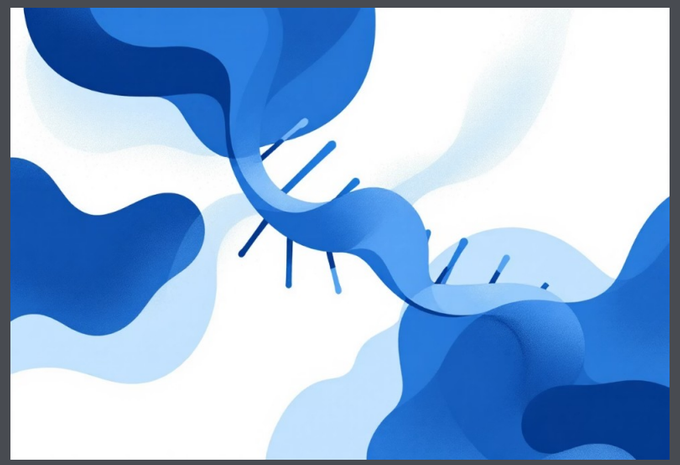
El análisis de cuantiles de cobertura, facilitado por KLL Sketches, permite identificar regiones con posibles deleciones o duplicaciones, indicativas de CNVs.

Deleciones

565 bins (1.96%) con baja cobertura (< 0.5× mediana).

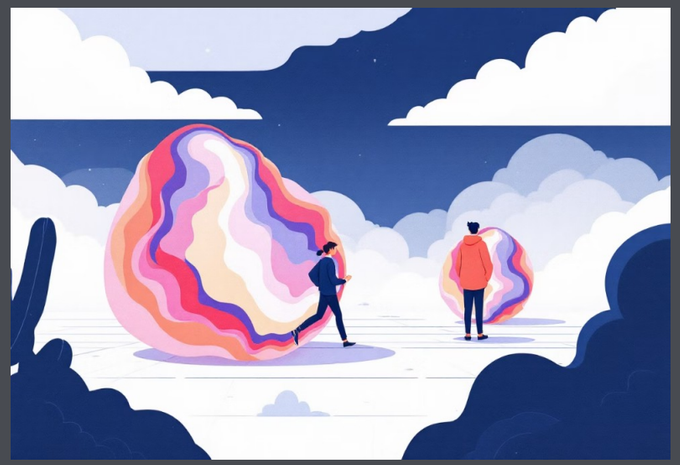
Duplicaciones

322 bins (1.12%) con alta cobertura (> 1.5× mediana).



Compresión

51× (de 458k bins a 563 valores), optimizando el almacenamiento.



Complejidad Computacional y Eficiencia

La comparación entre los métodos tradicionales de ordenamiento y KLL Sketches resalta la superioridad de estos últimos en términos de escalabilidad y uso de recursos.

Método	Tiempo	Memoria	Complejidad
Sort Tradicional (ej. Merge sort)	O(n log n)	Lineal	No escala
KLL Sketch	O(n)	Sublineal independiente de n	Escala infinitamente

La diferencia es crítica al trabajar con grandes volúmenes de datos genómicos, donde los KLL Sketches permiten una gestión de memoria constante.

Conclusiones Principales

Herramienta Complementaria: No reemplaza herramientas especializadas de CNV (CNVnator, DELLY), sino que complementa su uso.

Uso Ideal: Excelente para control de calidad (QC) y análisis exploratorio, no para diagnóstico final.

Impacto: Los sketches probabilísticos permiten un análisis escalable de datos genómicos masivos.

En estudios grandes: Crítico para estudios donde se deban procesar más de un genoma a la vez, donde los métodos tradicionales resultan inviables.