Plataforma distribuída con Spark para el análisis de secuencias Next-generation

HPC

MSc. Vicente Machaca Arceda

November 19, 2021

Overview

Table of Contents

Problema

El análisis de secuencias de ADN obtenidas por *Next-generation sequencing* es una tarea vital en los estudios de Bioinformática. Pero lamentablemente, las herramientas tradicionales no lentas y no aprovechan el poder computacional de un sistema distribuído

Objetivos

Desarrollar una herramienta para el análisis de secuencias obtenidas de *Next-generation sequencing*.

En esta versión, se implemento la funcionalidad para analizar la calidad de las lecturas.

Table of Contents

Secuenciamiento de ADN

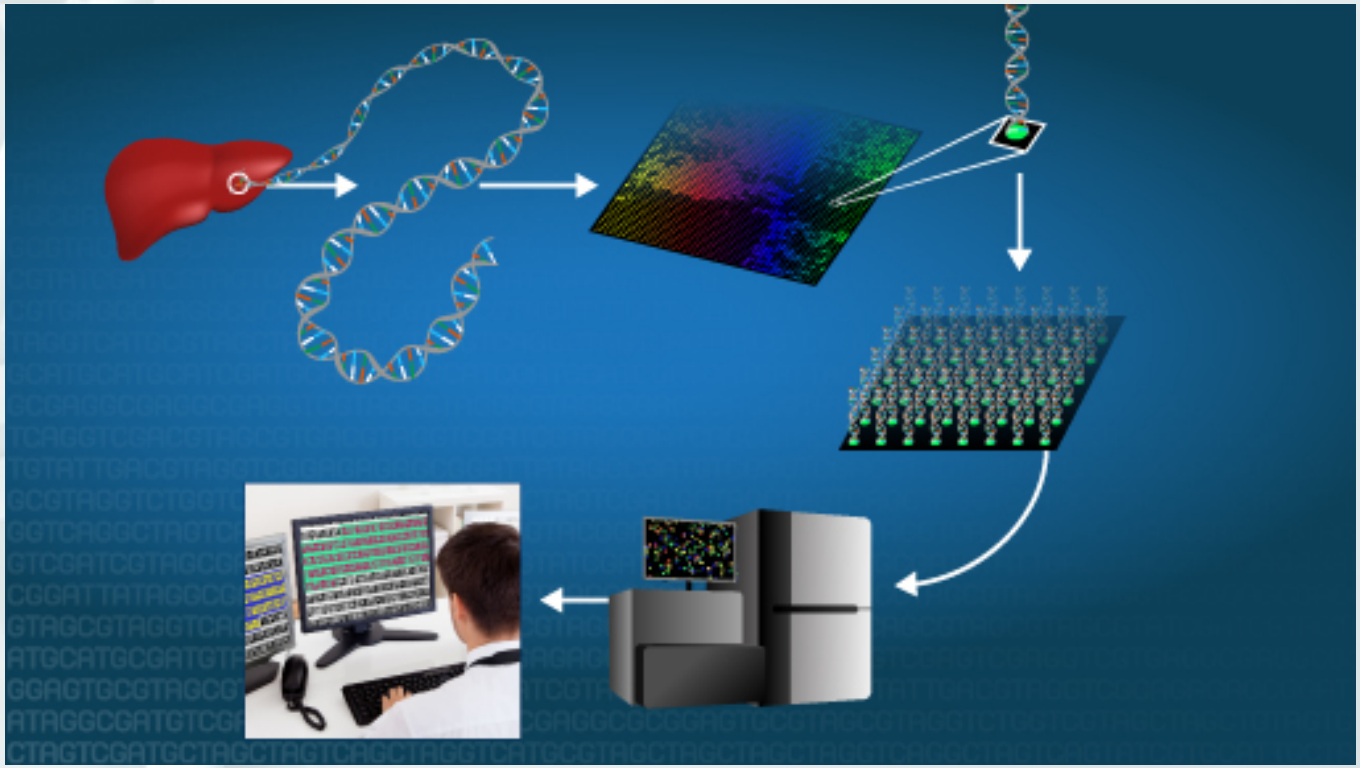

Figure: Secuenciamiento de ADN

Next-generation sequencing

Figure: Next-generation sequencing

Single cell RNAseq

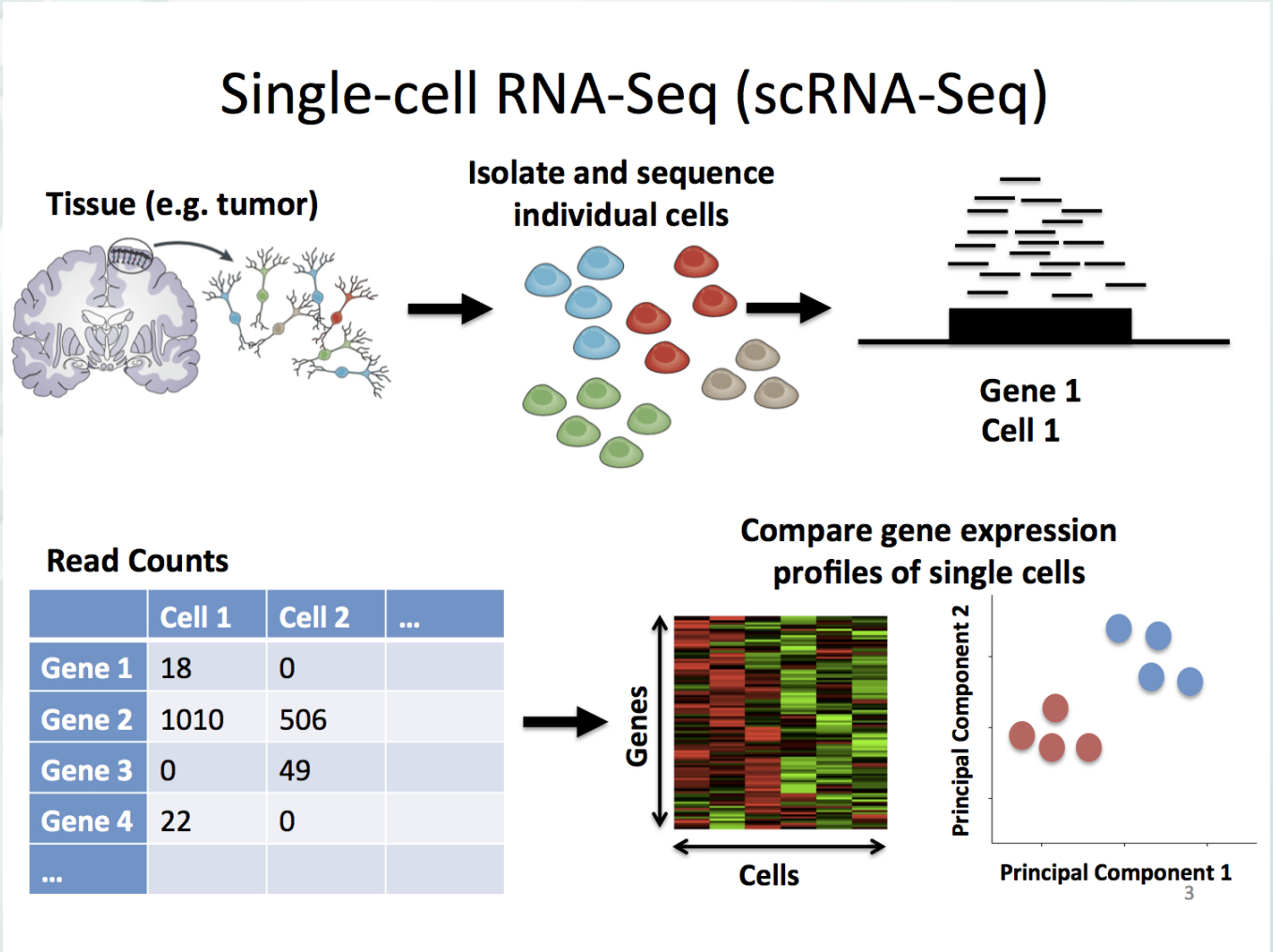

Figure: Single cell RNAseq

Table of Contents

Propuesta
Análisis de secuencias

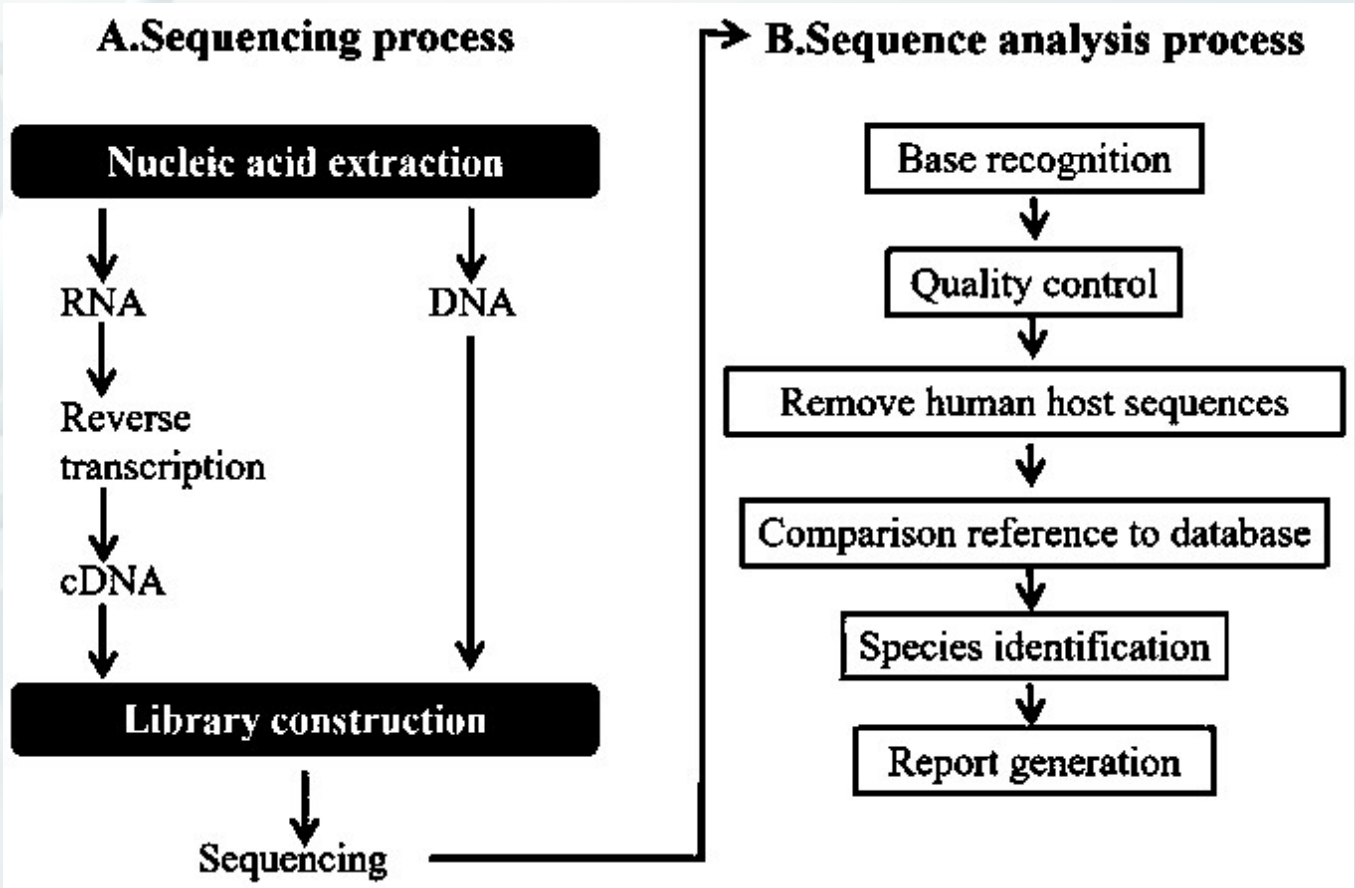

Figure: Phases comunes realizadas en un análisis de secuencias de ADN (Next-generation).

Table: Herramientas utilizadas para el proyecto

Herramienta	Version
Spark	3.2.0
Python	3.8.10
Pyspark	3.2.0

Table: Computadoras utilizadas en el sistema distribuído

Nombre de PC	Especificaciones
Desktop Asus	Procesador i7 de séptima generación y 8GB de memoria RAM. Sistema operativo Linux.
Laptop Asus	Procesador i5 de quinta generación y 4GB de memoria RAM. Sistema operativo Linux.

Propuesta
Hardware

Propuesta
Funcionalidades

- Conteo de la cantidad de secuencias.
- Conteo total de las bases nitrogenadas.
- Computo de la longitud de todas las secuencias.
- Computo del promedio de las longitudes de las secuenias.
- Computo de la ocurrencia de cada base nitrogenada.
- Análisis de contenido por base.

Table of Contents

Resultados

Para evaluar el desempeño de la propuesta se evaluó las secuencias con el código **ERR3014700**, estas fueron descargadas de NCBI.

Resultados
Estadísticas


```
{
    "bases": 1843156,
    "total_seqs": 462393,
    "seqs_len": [
        523,        600,        599,        600,        599,
        600,        600,        529,        600,        538
    ],
    "seqs_len_mean": 564,
    "bases_ocurrence": [
        [           "C",            71397200        ],
        [           "N",            4054        ],
        [           "D",            3004        ],
        [           "G",            70502420        ],
        [           "A",            59888213        ],
        [           "F",            16160       ],
        [           "T",            59010438        ],
        [           "B",            203         ],
        [           "E",            4144        ]
    ]
}
```


Figure: Estadísticas de las lecturas.

Resultados
Análisis de contenido por base

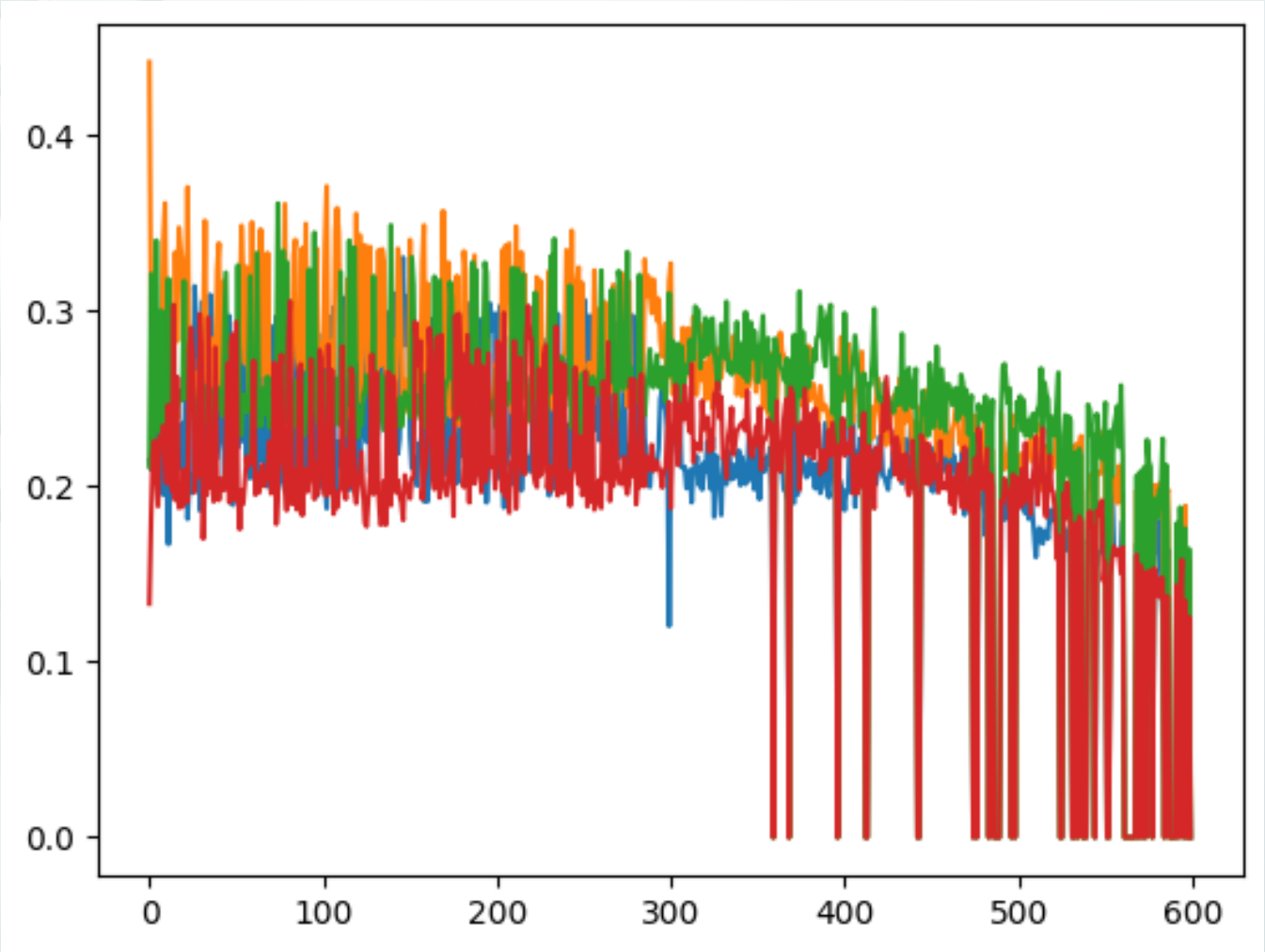

Figure: Análisis de contenido por base.

Table of Contents

Conclusiones

- En este proyecto se ha desarrollado una herramienta distribuída que permite hacer el análisis masivo de grandes cantidades de lecturas de ADN (Next-generation sequencing).
- El proyecto se enfoco en el análisis de calidad de las lecturas de ADN, este es un paso crucial en cualquier experimento de Bioinformática. Como resultado, la propuesta obtiene estadísticas referentes a las ocurrencias de las bases nitrogenadas y además se desarrollo un análisis de contenido por base.

References I