|  |  |  |
| --- | --- | --- |
| Parámetro | Resultado | Posible Causa |
| Per base sequence quality | Fail | Las bases de 96-97 es menor a 5 |
| Per tile sequence quality | Pass |  |
| Per sequence quality score | Pass |  |
| Per base sequence content | Fail | En A y T hay una diferencia de más del 20% |
| Per sequence GC content | Fail | La suma de las desviaciones de la distribución normal representa más del 30% de las lecturas. |
| Per base N content | Regular | La base 1 es >5% |
| Sequence length distribution | Pass |  |
| Sequence duplication levels | Pass |  |
| Overrepresented sequence | Fail | Hay una secuencia representada con el 13,8% |
| Adapter content | Pass |  |
| Basic statistics | Pass |  |
| Recomendación final |  | Repetir el análisis |

**Natalia Larrazábal y Clara Sicilia LAB IV**

**SECUENCIA READ1**

**SECUENCIA READ2**

|  |  |  |
| --- | --- | --- |
| Parámetro | Resultado | Posible cuasa |
| Per base sequence quality | pass |  |
| Per tile sequence quality | pass |  |
| Per sequence quality score | pass |  |
| Per base sequence content | fail | La diferencia entre A y T en la base 9 es mayor a 20% |
| Per sequence GC content | fail | la suma de las desviaciones de la distribución normal representa más del 30% de las lecturas. |
| Per base N content | pass |  |
| Sequence lengh distribution | pass |  |
| Sequence duplication levels | pass |  |
| Overrepresented sequence | regular | Dos secuencias representan más del 0,1% del total. (0,11%) |
| Adapter content | pass |  |
| Recomendación final |  | Solo hay dos fallos y sí que recomiendo usar esta secuencia |

**SECUENCIA READS 3**

|  |  |  |
| --- | --- | --- |
| Parámetro | Resultado | Posible cuasa |
| Per base sequence quality | Pass |  |
| Per tile sequence quality | Regular | La mediana de algunas bases entre los valores de 450-459 de menos de 25 |
| Per sequence quality score | Pass |  |
| Per base sequence content | Fail | Falla porque la diferencia entre A y T y la diferencia entre G y C en el intervalo de las bases 450-459 es mayor del 20% |
| Per sequence GC content | Pass |  |
| Per base N content | Pass |  |
| Sequence lence distribution | Regular | Hay una gran variación entre la longitud de los fragmentos |
| Sequence duplication levels | Pass |  |
| Overrepresented sequence | Regular | Los valores son mayores al 0.1% |
| Adapter content | Pass |  |
| Recomendación final |  | No se considera suficientemente fiable, por lo que se debería repetir el análisis |

**RAW GBS**

|  |  |  |
| --- | --- | --- |
| Parámetro | Resultado | Posible causa |
| Per base sequence quality | fail | A partir de la bp 68-69 los cuartiles son inferiores al límite de 5 que tiene el sistema para este parámetro. |
| Per sequence quality score | pass |  |
| Per base sequence content | fail | Desde la base 1 a la 5 la diferencia entre las bases complementarias excede el límite del 20%. |
| Per sequence GC content | regular | La suma de las dos desviaciones representa más del 15% de las lecturas. |
| Per base N content | fail | A partir de la bp 65 supera el 20% de contenido de N |
| Sequence lenght distribution | regular | Las secuencias no tienen todas la misma longitud. |
| Sequence duplication levels | regular | Las secuencias que no son únicas superan en más de un 20% a las secuencias totales. |
| Overrepresented sequence | fail | Hay alguna secuencia que representa más del 1% del total. |
| Adapter content | pass |  |
| Basic statistics | pass |  |
| Recomendación final |  | Repetir análisis. |

**SECUENCIA CLEAN**

|  |  |  |
| --- | --- | --- |
| Parámetro | Resultado | Posible causa |
| Per base sequence quality | Pass |  |
| Per sequence quality score | Pass |  |
| Per base sequence content | Fail | Desde la bp 1 a la 5 la  diferencia entre las bases  complementarias excede el  límite del 20%. |
| Per sequence GC content | Regular | La suma de las dos  desviaciones representa  más del 15% de las  lecturas. |
| Per base N content | Pass |  |
| Sequence lence distribution | Regular | Las secuencias no tienen todas la misma longitud. |
| Sequence duplication levels | Regular | Las secuencias que no son únicas superan en más de un 20% a las secuencias totales. |
| Overrepresented sequence | Fail | Hay alguna secuencia que representa más del 1% del total. |
| Adapter content | Pass |  |
| Basic statistics | Pass |  |
| Recomendación final |  | Repetir análisis |

**CUESTIONES**

**1.Investiga qué programas se suelen usar para continuar con el pre-procesamiento.**

En el preprocesamiento de datos, varios programas y lenguajes de programación son útiles dependiendo del tipo de datos y el objetivo.

Algunos son: Python, R, PySpark, Hadoop, SQL, FastUniq

**2.Los archivos Raw GBS y Clean GBS son la misma muestra antes (Raw GBS) y despues (Clean GBS) de hacer el pre-procesamiento ¿Que cambia de una con respecto a la otra?**

Los archivos Raw GBS contienen datos de secuenciación de ADN sin procesar, que incluyen lecturas de baja calidad, secuencias contaminantes, duplicados, y errores de secuenciación. Durante el preprocesamiento para generar los archivos Clean GBS, se eliminan estas lecturas de baja calidad, se filtran secuencias no deseadas, se descartan duplicados y se corrigen errores. Además, se realiza un filtrado para retener solo polimorfismos (SNPs) de alta calidad. Esto prepara los datos para análisis genómicos confiables y precisos.