Tarea 3.

David Figueroa Blanco - Sergio Canales.

Alineamiento de lecturas.

Se estudió la calidad de las lecturas de secuenciación utilizando NGSEPcore_3.2.0.jar. Se calculo el error mediante la razón entre el número de diferencias con el genoma de referencia y el total de alineamientos únicos, donde se obtuvo la siguiente gráfica :

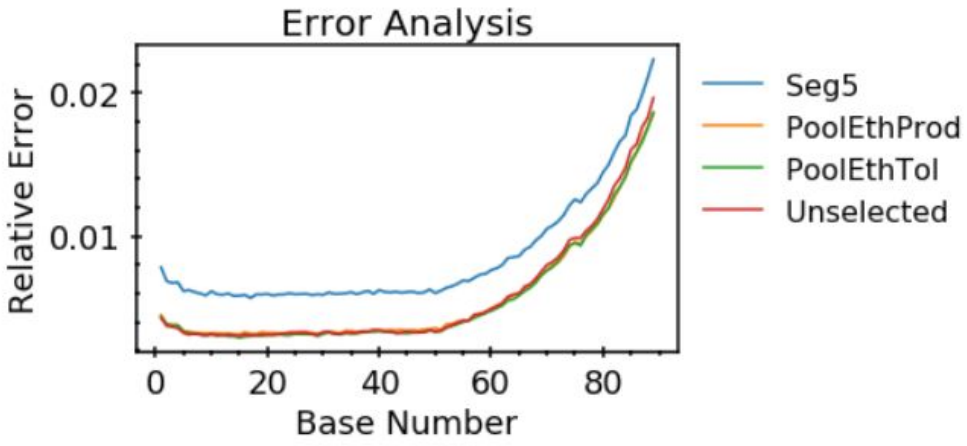
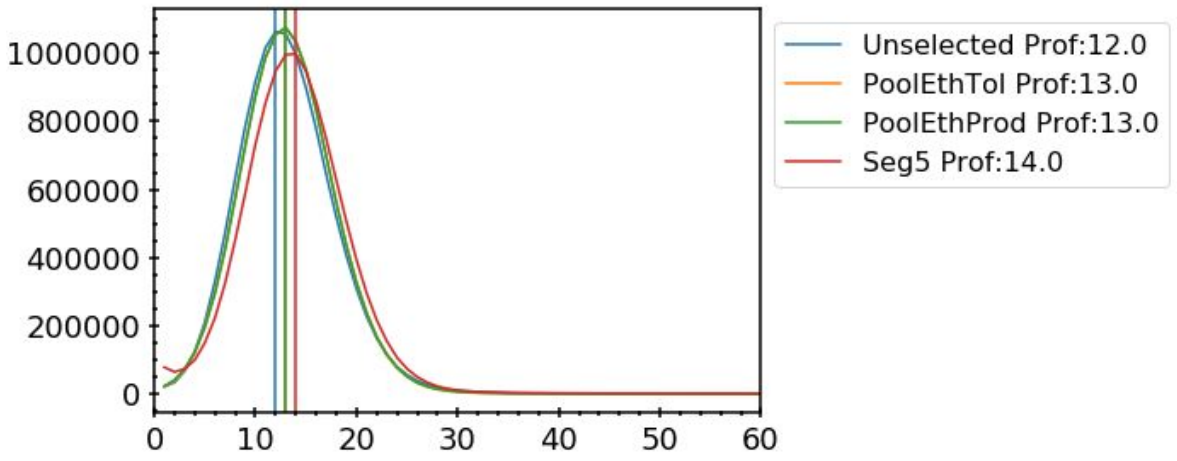


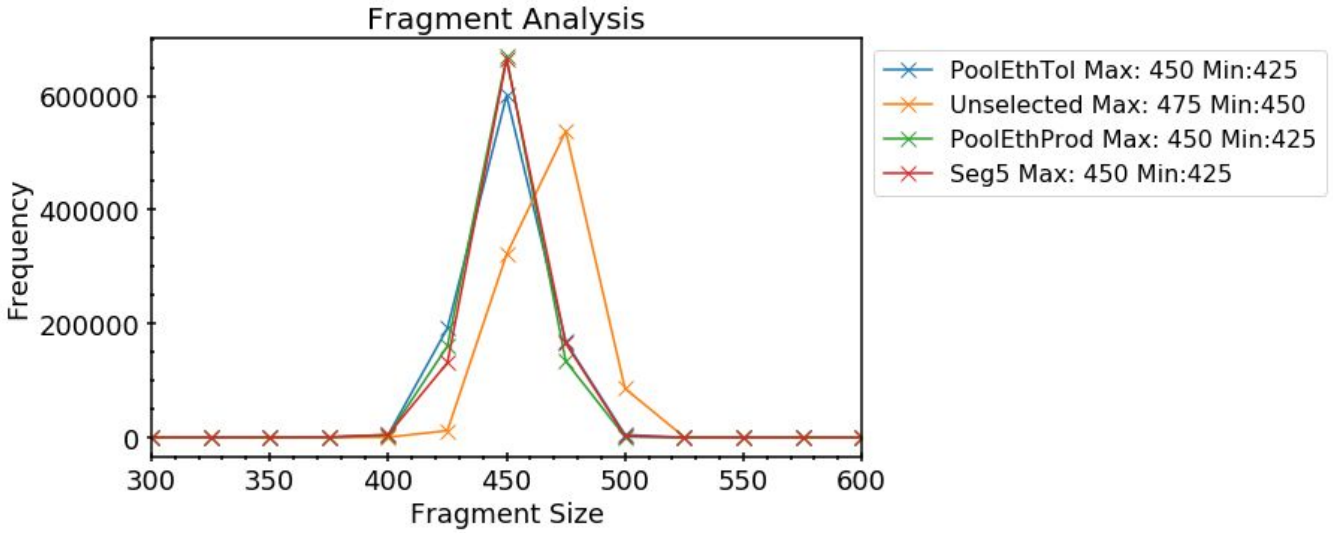
Figura 1. Tasa de error para cada una de las gráficas.

Es posible evidenciar que no se obtiene una distribución uniforme, y que el error aumenta considerablemente al final de la lectura. En todas las muestras se encontró un error máximo en la última base con valores de 0.0224, 0.0186, 0.0187 y 0.0197 para Seg5,PoolEthProd, PoolEthTol y Unselected respectivamente. Esto se debe a que la secuenciación de ILLUMINA presenta un error considerable al final de la lectura.

Para la estadísticas de cubrimiento se encontraron las siguientes distribuciones donde se muestra la moda como una línea vertical y se muestra su valor en cada caso como Prof :



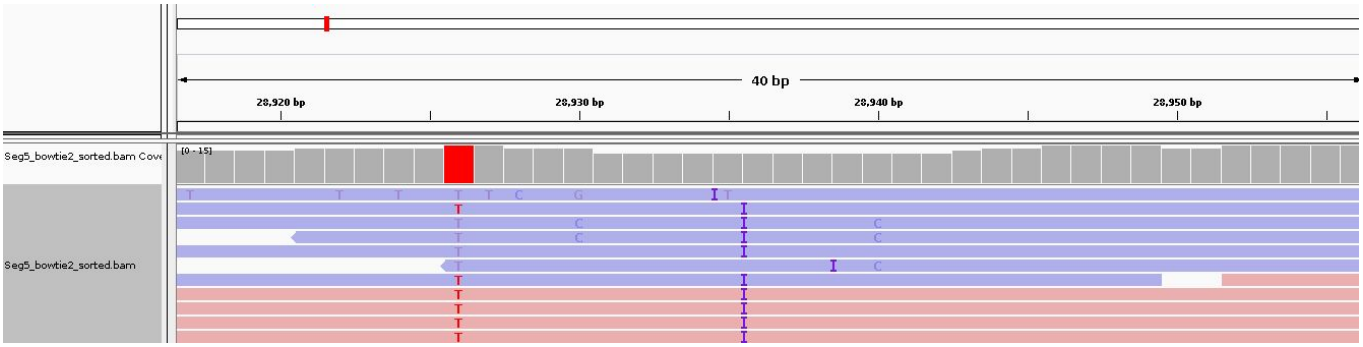
Finalmente para la distribución del tamaño del fragmento se obtuvo :



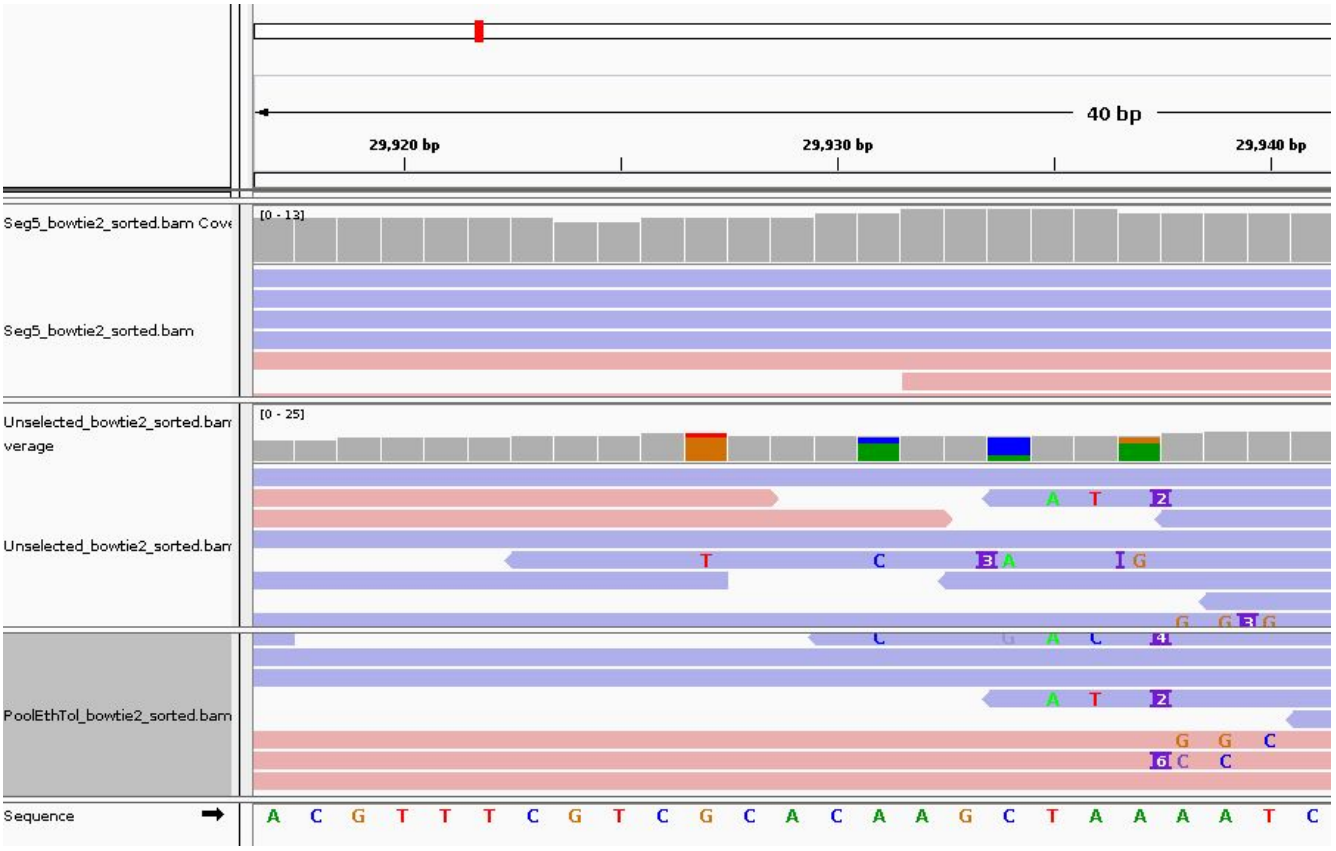
Ahora bien, a partir de los datos obtenidos es posible determinar que el tamaño mínimo puede ser 425 y el máximo de 475.

Visualización de alineamientos.

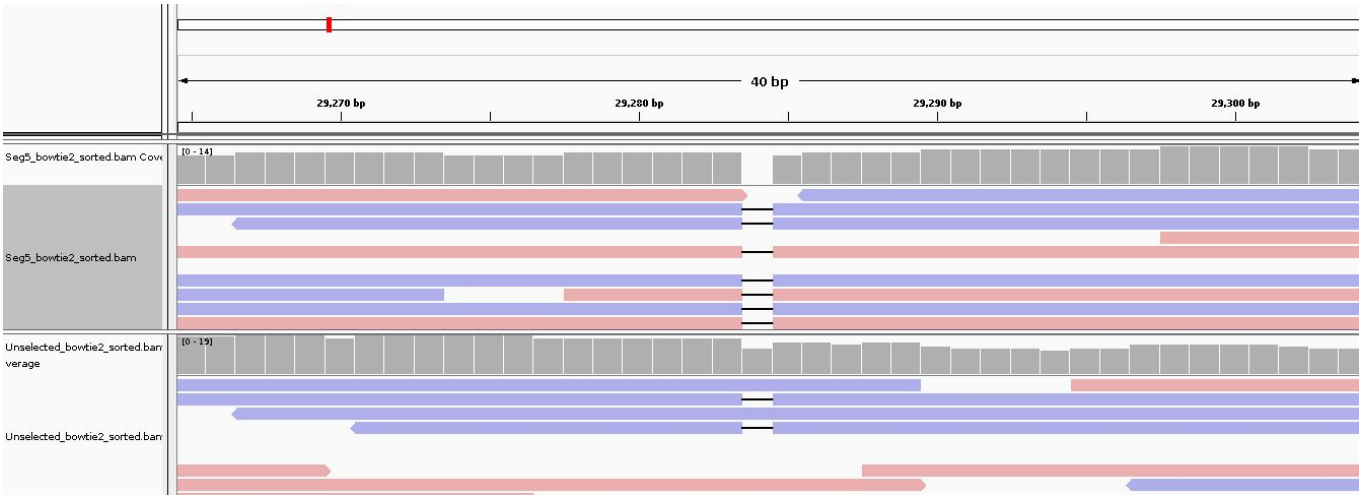
En la herramienta de visualización de IGV las inserciones detectadas son marcadas con color púrpura como se muestra a continuación. Cada hebra marcada con un color específico:



Como se evidencia en este caso ocurre una inserción en las dos hebras. Adicionalmente es posible identificar inserciones de varias bases en algunas muestras ("Unselected" & "PolrEthTool") pero no en Seg5. Coordenadas chrI.29.937:

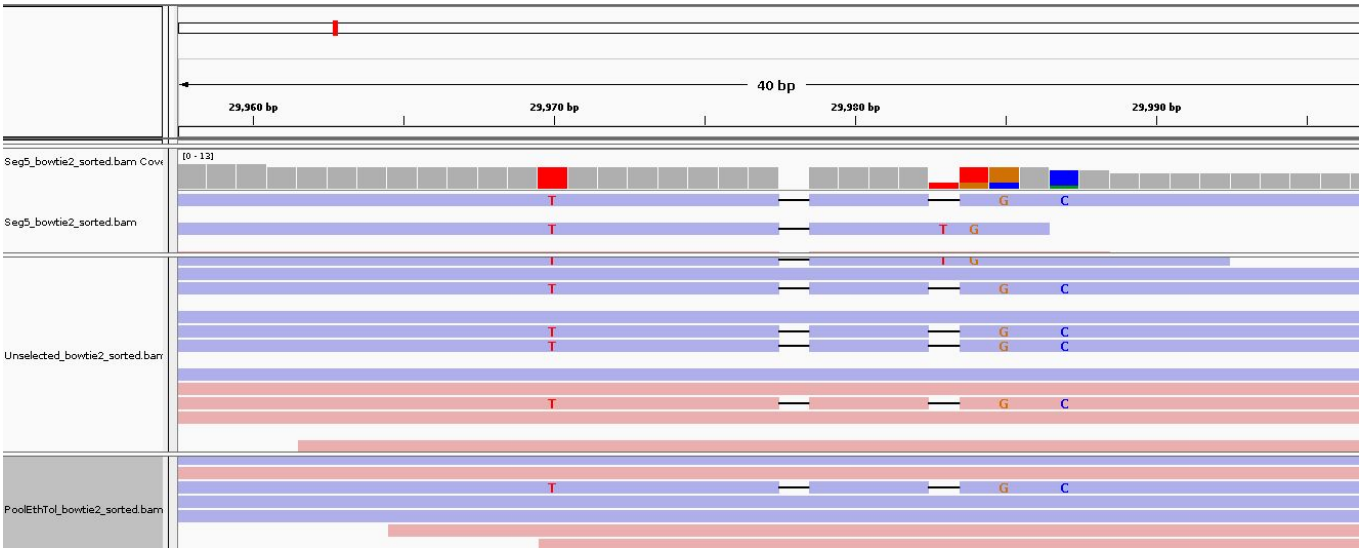


Paralelamente las eliminaciones de alguna base se marcan con "-" como se muestra a continuación (Coordenadas Chr1:29,284). Cabe resaltar que por ejemplo todas las lecturas de las muestra "Seg5" tienen esa eliminación sin embargo algunas lecturas de la muestra "Unselected" si contienen la base correspondiente.



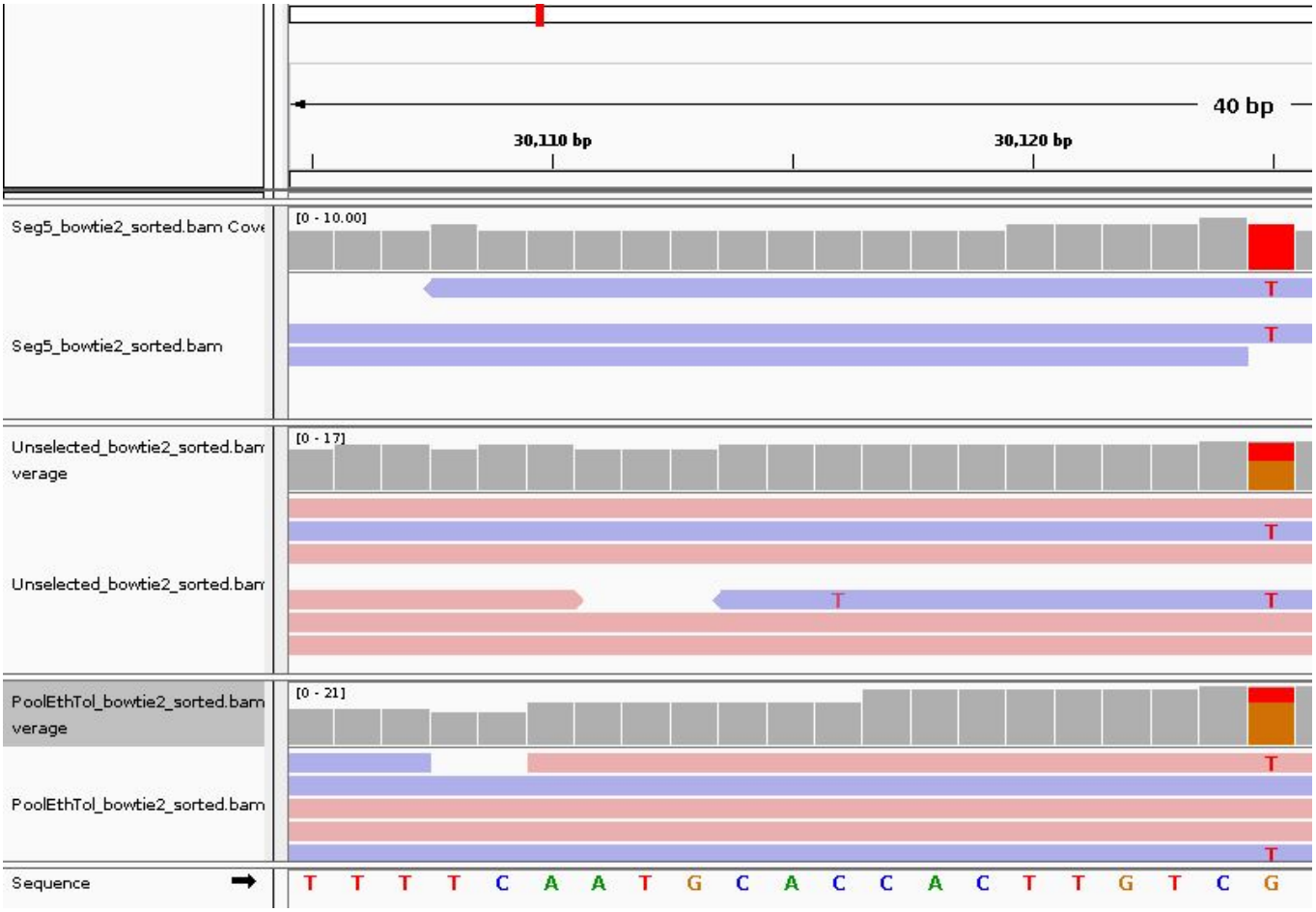
SNPs,

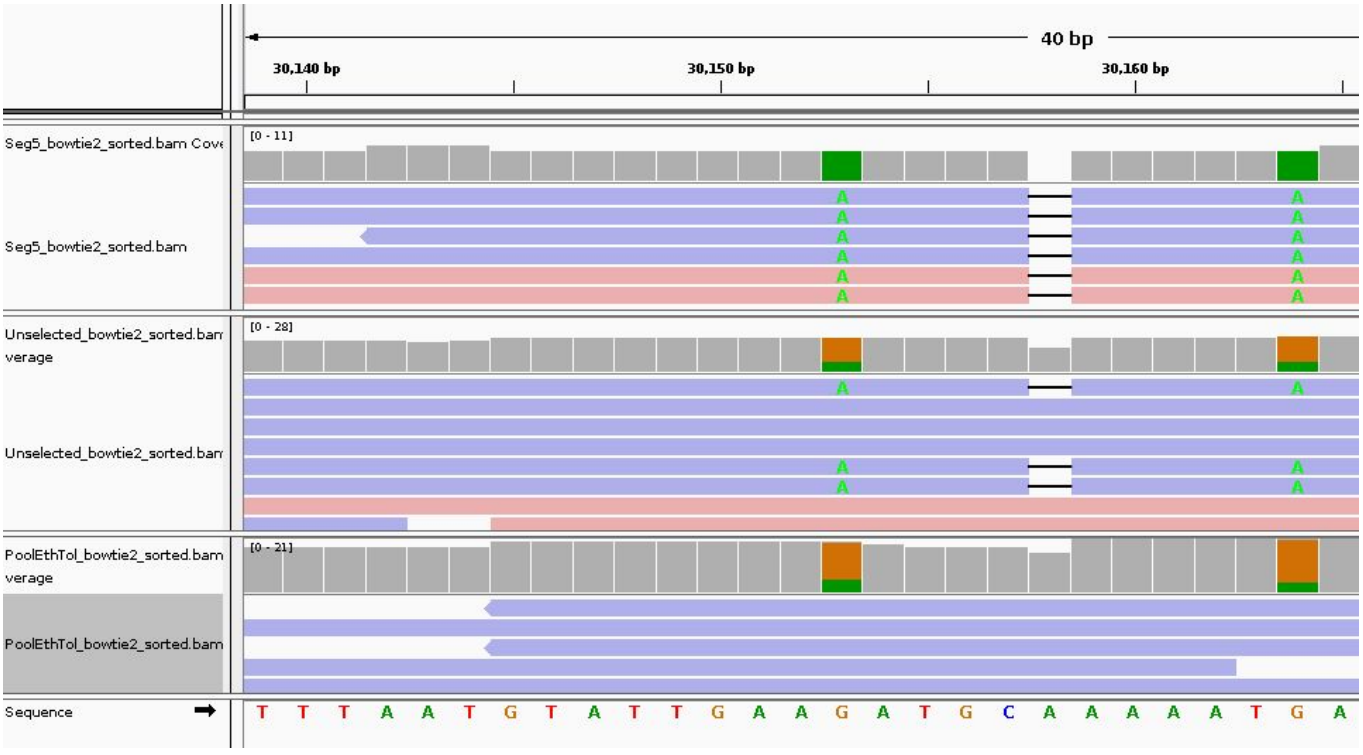
En la coordenada chrI:29,978 se evidencia el caso de un eliminacion en ambas lecturas dado paso a un genotipo homocigoto mientras que en las muestras de Unselected y PoolEthTol no se encuentra la eliminación en todas las lecturas:



SNPs.

Para la detección de SNPs se utilizó un porcentaje mayor al 20 % en variaciones y se encontraron SNPsen chrI:30,125, chrI:30,153 y chrI:30,164 en Seg5, Unselected y PoolEthTol como se muestra a continuación:





Para Seg5 se evidencia que se da un genotipo homocigoto debido a un SNP de A contra G mientras que para las muestras de “Unselected “ y “PoolEthTol” se encuentran genotipos heterocigotos debido a que se data una mezcla en los SNPS. en chrI:30,153 y chrI:30,164

Punto 3

El script para correr el tercer punto es *punto3.py*

Arreglo de sufijos

Para correr el arreglo de sufijos es necesario enviar 3 parámetros por consola.

param1: SUFFIX

param2: fasta file location

param3: fastq file location

los archivos se buscan bajo la carpeta T3

un ejemplo con datos ya existentes es

```
/T3$ python3 punto3.py SUFFIX COVID.fasta COVID_seqLen50_2.fastq
```

para los datos que no se encontraron, se imprime -1

BWT

Para correr la transformación de BW, es necesario ingresar 3 parámetros

param1: BWT

param2: fasta file location

param3: subcadena a buscar

un ejemplo con datos ya existentes es

```
/T3$ python3 punto3.py BWT COVID.fasta AGATTTTCAAGAAAACTGGAACACT
```

cuya respuesta es

```
AGATTTTCAAGAAAACTGGAACACT was found in the suffix array at the position: 730
```

Para intentar cualquier dato con ese archivo se puede abrirlo, copiar la cadena a buscar y pegarla junto con los comandos pedidos. Es importante recalcar que la posición se cuenta desde 1 y no desde 0.