Tarea 2 – Ensamblaje de genomas

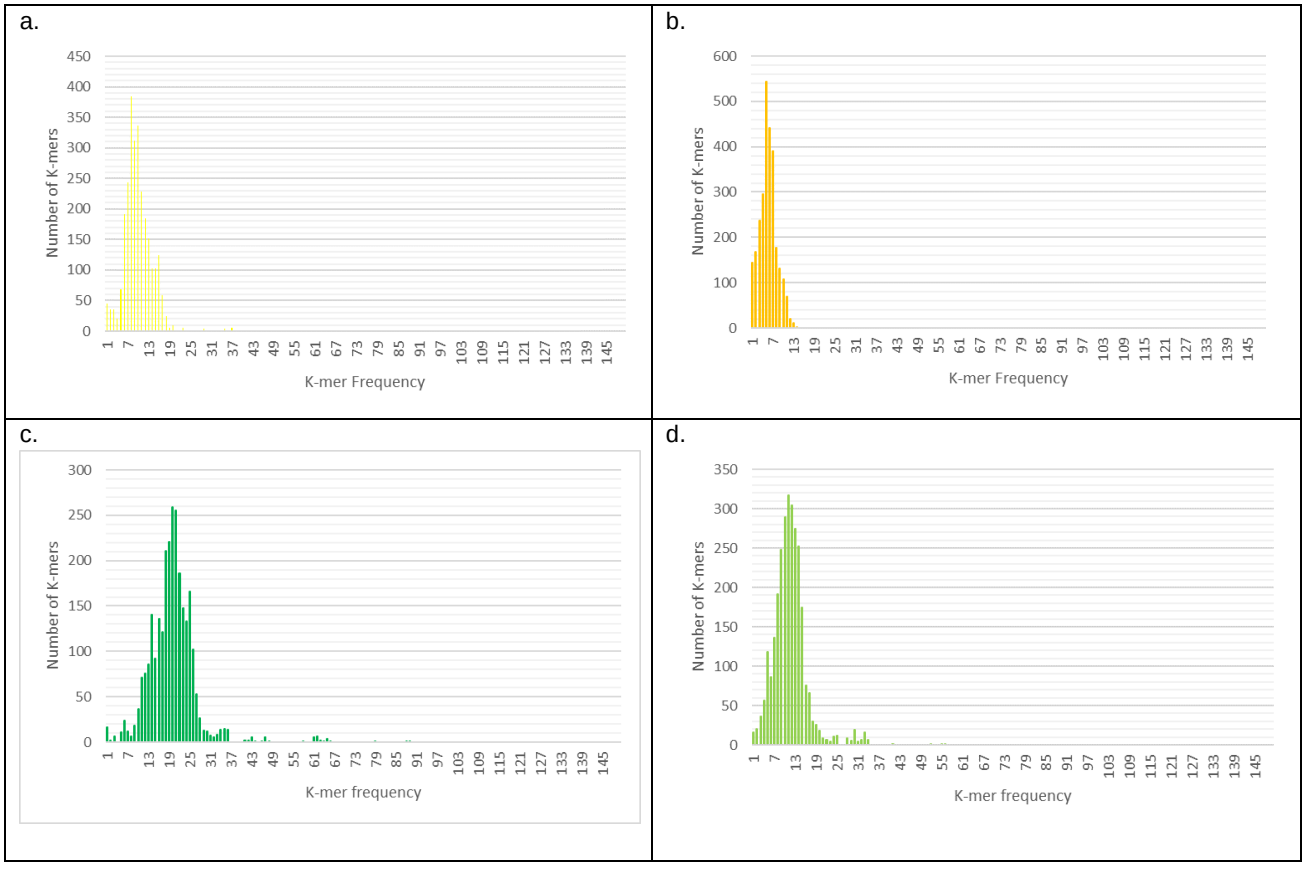
Ana Sofía Villa Benavides (201923361), Luisa Sacristán (201016005)

Nota: en la carpeta de entrega/ también se adjunta un documento de Excel con los datos completos de cada punto, en este informe incluimos gráficas y tablas que los resumen.

1. Luego de implementar todos los métodos de la clase “KmersTable” y probar el programa cada archivo fastq (10X,20X,50X,100X) de ejemplo y tamaños de k-mer (5, 10, 20, 50 y 75) se obtuvieron los resultados contenidos en la Tabla 1.

	10 X			20 X			50 X			100 X		
Tamaño de k-mer	Cantidad de k-mers	Abundancia media	Tiempo de ejecución (ms)	Cantidad de k-mers	Abundancia media	Tiempo de ejecución (ms)	Cantidad de k-mers	Abundancia media	Tiempo de ejecución (ms)	Cantidad de k-mers	Abundancia media	Tiempo de ejecución (ms)
5	2152	13,4	207	2177	25,9	204	2175	66,0	246	2179	130,2	354
10	2703	10,1	200	2741	19,9	219	2739	49,8	226	2743	99,5	263
20	2758	8,8	204	2796	17,4	220	2794	43,4	233	2798	86,8	298
50	2735	5,6	186	2773	11,5	193	2771	27,6	231	2775	55,1	241

Tabla 1 *Implementación de la clase KmersTable.*



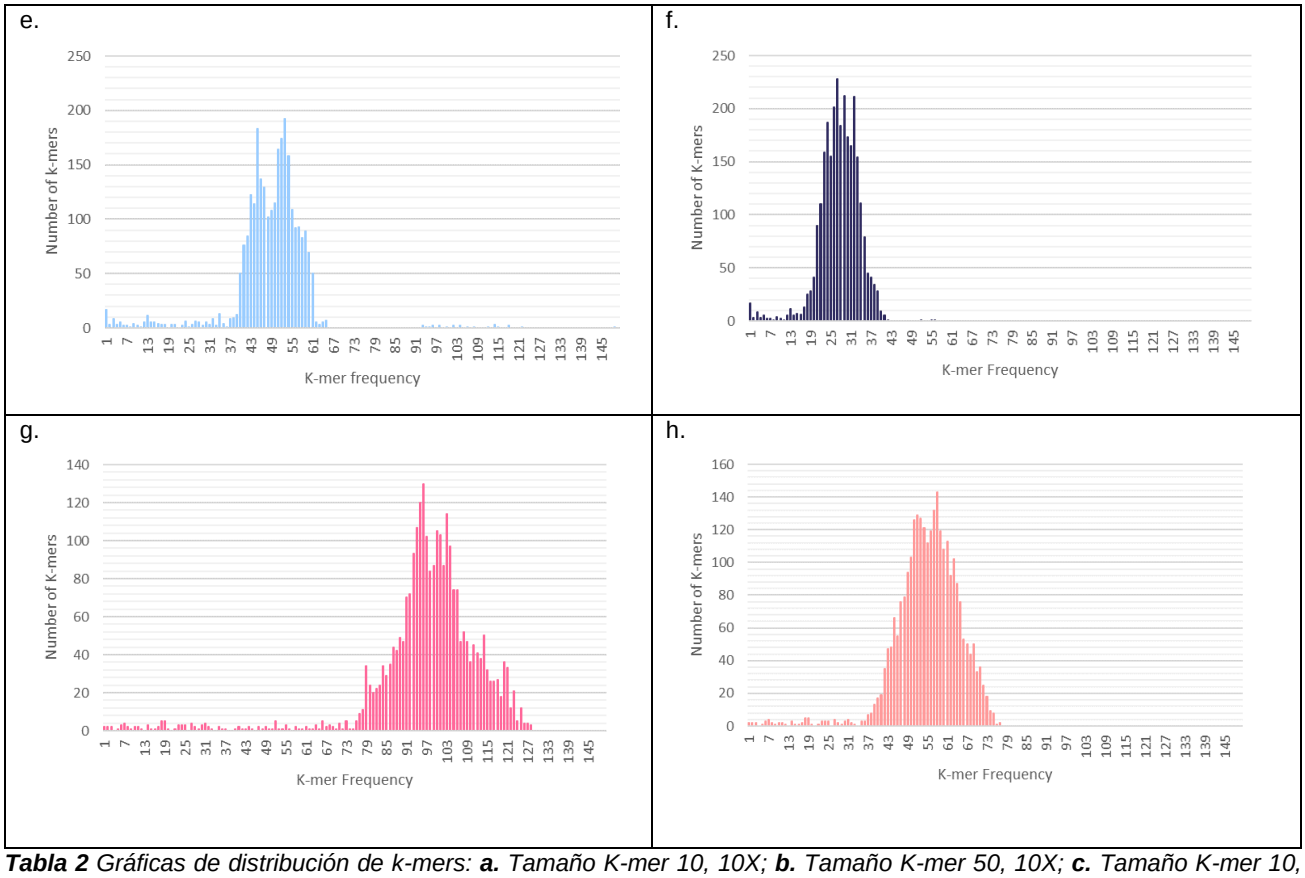


Tabla 2 *Gráficas de distribución de k-mers:* ***a.*** *Tamaño K-mer 10, 10X;* ***b.*** *Tamaño K-mer 50, 10X;* ***c.*** *Tamaño K-mer 10, 20X;* ***d.*** *Tamaño K-mer 50, 20X;* ***e.*** *Tamaño K-mer 10, 50X;* ***f.*** *Tamaño K-mer 50, 50X;* ***g.*** *Tamaño K-mer 10, 100X;* ***h.*** *Tamaño K-mer 50, 100X;*

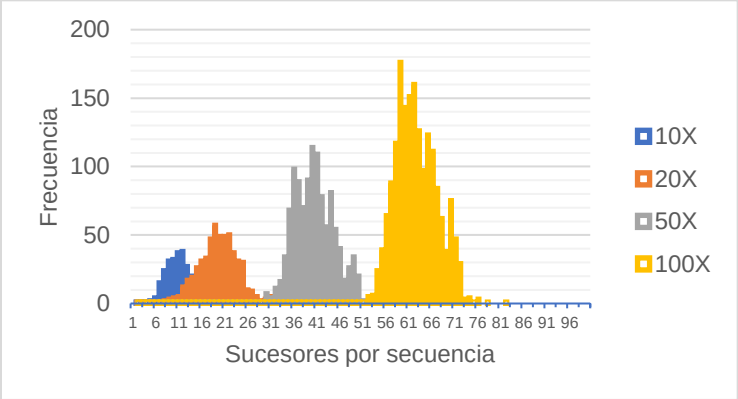
En primer lugar, la cantidad de k-mers se mantuvo relativamente constante, en un rango entre 2000 y 2800 esto tiene sentido puesto que este valor depende del tamaño total de la secuencia final a ensamblar y en este ejercicio el texto base a ensamblar es técnicamente el mismo. La variabilidad se debe a que las lecturas a diferentes coberturas puede que no cubran la totalidad de secuencias por lo cual puede que a baja cobertura ciertos kmers no estén representados. Se percibe también una relación entre la abundancia media y la cobertura promedio del archivo, esto tiene sentido ya que se espera que la media de la abundancia se encuentre. Finalmente, en cuanto al tiempo de ejecución, este mantiene un valor relativamente constante alrededor de 200 ms, esto indica que la fabricación de la tiene una complejidad baja o constante.
Nota: para calcular la abundancia promedio, se multiplico a partir de la distribución la abundancia por su frecuencia, estos valores se sumaron y esto se dividió sobre el número total de Kmers. Este cálculo se encuentra en el Excel entrega/Datos_Resultados

2. Luego de completar la clase “OverlapGraph” hasta el método “calculateOverlapDistribution”, se probó el programa con las distintas lecturas dadas (10X, 20X, 50X y 100X) a continuación se encuentra la distribución de abundancias de secuencias, la distribución de sucesores por secuencia y el tiempo de ejecución de cada prueba.

Abundancia	**10X**	**20X**	**50X**	**100X**
1	270	487	874	1012
2	15	52	233	546
3	0	3	44	197
4	0	0	7	62
5	0	0	0	9
6	0	0	0	2
7	0	0	0	0
8	0	0	0	0
Tiempo de ejecución (ms)	8	13	38	59

Tabla 2 *Distribución de abundancias de secuencias*

Figura1. *Distribución de sucesores por secuencia*



3. A continuación, se muestran las secuencias ensambladas a partir de los distintos archivos (10X, 20X, 50X, y 100X) utilizando el valor por defecto de sobre lape mínimo (10):

	SECUENCIA ENSAMBLADA
10X	l peloton de fusilamiento el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevo a conocer el hielo. Macondo era entonces una aldea de veinte casas de barro y canhabrava construidas a la orilla de un rio de aguas diafanas que se precipitaban por un lecho de piedras pulidas blancas y enormes como huevos prehistoricos. El mundo era tan reciente que muchas cosas carecian de nombre y para mencionarlas habia que senhalarlas con el dedo. Todos los anhos por el mes de marzo una familia de gitanos desarrapados plantaba su carpa cerca de la aldea y con un grande alboroto de pitos y timbales daban a conocer los nuevos inventos. Primero llevaron el iman. Un gitano corpulento de barba montaraz y manos de gorrion que se presento con el nombre de Melquiades hizo una truculenta demostracion publica de lo que el mismo llamaba la octava maravilla de los sabios alquimistas de Macedonia. Fue de casa en casa arrastrando dos lingotes metalicos y todo el mundo se espanto al ver qu
20X	chos anhos despues frente al peloton de fusilamiento el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevo a conocer el hielo. Macondo era entonces una aldea de veinte casas de barro y canhabrava construidas a la orilla de un rio de aguas diafanas que se precipitaban por un lecho de piedras pulidas blancas y enormes como huevos prehistoricos. El mundo era tan reciente que muchas cosas carecian de nombre y para mencionarlas habia que senhalarlas con el dedo. Todos los anhos por el mes de marzo una familia de gitanos desarrapados plantaba su carpa cerca de la aldea y con un grande alboroto de pitos y timbales daban a conocer los nuevos inventos. Primero llevaron el iman. Un gitano corpulento de barba montaraz y manos de gorrion que se presento con el nombre de Melquiades hizo una truculenta demostracion publica de lo que el mismo llamaba la octava maravilla de los sabios alquimistas de Macedonia. Fue de casa en casa arrastrando dos lingotes metalicos y todo el mundo se espanto al ver que los calderos las pailas las tenazas y los anafes se caian de su sitio y las maderas crujian por la desesperacion de los clavos y los tornillos tratando de desenclavarse y aun los objetos perdidos desde hacia mucho tiempo aparecian por donde mas se les habia buscado y se arrastraban en desbandada turbulenta detras de los fierros magicos de Melquiades. Las cosas tienen vida propia - pregonaba el gitano con aspero acento - todo es cuestion de despertarles el anima. Jose Arcadio Buendia cuya desaforada imaginacion iba siempre mas lejos que el ingenio de la naturaleza y aun mas alla del milagro y la magia penso que era posible servirse de aquella invencion inutil para desentranhar el oro de la tierra. Melquiades - que era un hombre honrado - le previno: Para eso no sirve. Pero Jose Arcadio Buendia no creia en aquel tiempo en la honradez de los gitanos asi que cambio su mulo y una partida de chivos por los dos lingotes imantados. Ursula Iguaran su mujer que contaba con aquellos animales para ensanchar el desmedrado patrimonio domestico no consiguio disuadirlo. Muy pronto ha de sobrarnos oro para empedrar la casa replico su marido. Durante varios meses se empenho en demostrar el acierto de sus conjeturas. Exploro palmo a palmo la region inclusive el fondo del rio arrastrando los dos lingotes de hierro y recitando en voz alta el conjuro de Melquiades. Lo unico que logro desenterrar fue una armadura del siglo XV con todas sus partes soldadas por un cascote de oxido cuyo interior tenia la resonancia hueca de un enorme calabazo lleno de piedras. Cuando Jose Arcadio Buendia y los cuatro hombres de su expedicion lograron desarticular la armadura encontraron dentro un esqueleto calcificado que llevaba colgado en el cuello un relicario de cobre con un rizo de muje
50X	os anhos despues frente al peloton de fusilamiento el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevo a conocer el hielo. Macondo era entonces una aldea de veinte casas de barro y canhabrava construidas a la orilla de un rio de aguas diafanas que se precipitaban por un lecho de piedras pulidas blancas y enormes como huevos prehistoricos. El mundo era tan reciente que muchas cosas carecian de nombre y para mencionarlas habia que senhalarlas con el dedo. Todos los anhos por el mes de marzo una familia de gitanos desarrapados plantaba su carpa cerca de la aldea y con un grande alboroto de pitos y timbales daban a conocer los nuevos inventos. Primero llevaron el iman. Un gitano corpulento de barba montaraz y manos de gorrion que se presento con el nombre de Melquiades hizo una truculenta demostracion publica de lo que el mismo llamaba la octava maravilla de los sabios alquimistas de Macedonia. Fue de casa en casa arrastrando dos lingotes metalicos y todo el mundo se espanto al ver que los calderos las pailas las tenazas y los anafes se caian de su sitio y las maderas crujian por la desesperacion de los clavos y los tornillos tratando de desenclavarse y aun los objetos perdidos desde hacia mucho tiempo aparecian por donde mas se les habia buscado y se arrastraban en desbandada turbulenta detras de los fierros magicos de Melquiades. Las cosas tienen vida propia - pregonaba el gitano con aspero acento - todo es cuestion de despertarles el anima. Jose Arcadio Buendia cuya desaforada imaginacion iba siempre mas lejos que el ingenio de la naturaleza y aun mas alla del milagro y la magia penso que era posible servirse de aquella invencion inutil para desentranhar el oro de la tierra. Melquiades - que era un hombre honrado - le previno: Para eso no sirve. Pero Jose Arcadio Buendia no creia en aquel tiempo en la honradez de los gitanos asi que cambio su mulo y una partida de chivos por los dos lingotes imantados. Ursula Iguaran su mujer que contaba con aquellos animales para ensanchar el desmedrado patrimonio domestico no consiguio disuadirlo. Muy pronto ha de sobrarnos oro para empedrar la casa replico su marido. Durante varios meses se empenho en demostrar el acierto de sus conjeturas. Exploro palmo a palmo la region inclusive el fondo del rio arrastrando los dos lingotes de hierro y recitando en voz alta el conjuro de Melquiades. Lo unico que logro desenterrar fue una armadura del siglo XV con todas sus partes soldadas por un cascote de oxido cuyo interior tenia la resonancia hueca de un enorme calabazo lleno de piedras. Cuando Jose Arcadio Buendia y los cuatro hombres de su expedicion lograron desarticular la armadura encontraron dentro un esqueleto calcificado que llevaba colgado en el cuello un relicario de cobre
100X	Muchos anhos despues frente al peloton de fusilamiento el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevo a conocer el hielo. Macondo era entonces una aldea de veinte casas de barro y canhabrava construidas a la orilla de un rio de aguas diafanas que se precipitaban por un lecho de piedras pulidas blancas y enormes como huevos prehistoricos. El mundo era tan reciente que muchas cosas carecian de nombre y para mencionarlas habia que senhalarlas con el dedo. Todos los anhos por el mes de marzo una familia de gitanos desarrapados plantaba su carpa cerca de la aldea y con un grande alboroto de pitos y timbales daban a conocer los nuevos inventos. Primero llevaron el iman. Un gitano corpulento de barba montaraz y manos de gorrion que se presento con el nombre de Melquiades hizo una truculenta demostracion publica de lo que el mismo llamaba la octava maravilla de los sabios alquimistas de Macedonia. Fue de casa en casa arrastrando dos lingotes metalicos y todo el mundo se espanto al ver que los calderos las pailas las tenazas y los

> anafes se caian de su sitio y las maderas crujian por la desesperacion de los clavos y los tornillos tratando de desenclavarse y aun los objetos perdidos desde hacia mucho tiempo aparecian por donde mas se les habia buscado y se arrastraban en desbandada turbulenta detras de los fierros magicos de Melquiades. Las cosas tienen vida propia - pregonaba el gitano con aspero acento - todo es cuestion de despertarles el anima. Jose Arcadio Buendia cuya desaforada imaginacion iba siempre mas lejos que el ingenio de la naturaleza y aun mas alla del milagro y la magia penso que era posible servirse de aquella invencion inutil para desentranhar el oro de la tierra. Melquiades - que era un hombre honrado - le previno: Para eso no sirve. Pero Jose Arcadio Buendia no creia en aquel tiempo en la honradez de los gitanos asi que cambio su mulo y una partida de chivos por los dos lingotes imantados. Ursula Iguaran su mujer que contaba con aquellos animales para ensanchar el desmedrado patrimonio domestico no consiguio disuadirlo. Muy pronto ha de sobrarnos oro para empedrar la casa replico su marido. Durante varios meses se empenho en demostrar el acierto de sus conjeturas. Exploro palmo a palmo la region inclusive el fondo del rio arrastrando los dos lingotes de hierro y recitando en voz alta el conjuro de Melquiades. Lo unico que logro desenterrar fue una armadura del siglo XV con todas sus partes soldadas por un cascote de oxido cuyo interior tenia la resonancia hueca de un enorme calabazo lleno de piedras. Cuando Jose Arcadio Buendia y los cuatro hombres de su expedicion lograron desarticular la armadura encontraron dentro un esqueleto calcificado que llevaba colgado en el cuello un relicario de cobre con un rizo de mujer

Como se puede ver, manteniendo el valor mínimo de sobre lape por defecto, únicamente se puede ver como el tamaño y calidad del ensamble va mejorando mediante aumenta la cobertura. Por ejemplo, con la cobertura 10X solo se puede ensamblar un fragmento y es evidente que no comienza en el inicio de la frase dada, en cambio cuando se tiene cobertura 100X el ensamble abarca mucho más contenido y comienza en el inicio de la frase.
Luego, se realizaron distintos intentos con distintos valores de sobre lape mínimo para cada archivo. En el caso del archivo con cobertura de 10X, no se logró ensamblar la secuencia de manera completa, en comparación a los otros ensamblajes de mayor cobertura es evidente que la secuencia resultante es mucho más corta. Al cambiar el parámetro de sobre lape mínimo, se obtuvo un ensamblaje de tamaño máximo con sobre lape mínimo de **4** y se mantuvo el tamaño de la secuencia mostrada con el valor por defecto (~1100 car) entre los rangos de **5-36**, al aumentar el sobre lape mínimo a 37 la secuencia resultante disminuía en tamaño. Para la cobertura 20X el rango en el cual el ensamblaje mantuvo su longitud sin disminuir sustancialmente fue de **5-27** (~2800). Para la cobertura de 50X el ensamblaje mantuvo un tamaño de ~2800 caracteres entre **7-80** de sobre lape mínimo. Finalmente, para la cobertura 100X,
De este ejercicio se puede concluir entonces que la cobertura es esencial para poder ensamblar una secuencia. Con coberturas menores a 100X no fue posible llegar al texto dado ni siquiera modificando los tamaños de sobre lape mínimo. También es importante aclarar que entre mayor sea el sobre lape mínimo más certeza habrá en que la secuencia ensamblada este correcta, sin embargo, para aumentar el sobre lape mínimo y seguir obteniendo un ensamblaje significativo es necesario aumentar la cobertura también.

4. Se completó el script “SimpleReadsSimulator” para generar lecturas aleatorias de una secuencia de COVID ubicada en la carpeta data (EPI_ISL_3491834.fasta) de ~29 Kbp de tamaño. De este proceso resultaron 20 archivos en los cuales se prueban los tamaños de secuencia 50, 100, 200, 500 y profundidades promedio de 5X, 10X, 20X, 50X y 100X. A continuación, se encuentran reportados los tiempos de ejecución de los algoritmos Overlap y Kmers en cada caso:
(se mantuvo el arg[2] por defecto en ReadAnalyzerExample)

	RL = 50			**RL = 100**		
	Tiempo K-mers (ms)	Tiempo Overlap Graph (ms)	Tiempo Ensamblaje (ms)	Tiempo K-mers (ms)	Tiempo Overlap Graph (ms)	Tiempo Ensamblaje (ms)
5X	212	16654	50	239	10916	22
10X	242	61414	220	321	40416	409
20X	315	204649	211	403	141937	389
50X	486	754879	17183	583	649371	5373
100X	677	679059	25742	876	927408	4950

	RL = 200			**RL= 500**		
	Tiempo K-mers (ms)	Tiempo Overlap Graph (ms)	Tiempo Ensamblaje (ms)	Tiempo K-mers (ms)	Tiempo Overlap Graph (ms)	Tiempo Ensamblaje (ms)
5X	256	6769	10	263	5258	11
10X	319	25737	84	336	17493	39
20X	451	92340	75	490	64784	59
50X	687	504725	6138	694	380195	710
100X	1019	1658776	18757	1061	1367576	1473

De acuerdo con los tiempos reportados, el algoritmo de construir la tabla de k-mers tiene una complejidad constante. Aunque si se presentan ciertas variaciones que indican una tendencia de aumento de tiempo de ejecución cuando aumenta el tamaño de lectura (RL) y la cobertura, ningún tiempo de ejecución de este algoritmo pasa de **1.5** segundos. En cuanto al algoritmo de ensamblaje de la secuencia, este presenta tiempos de ejecución mayores, sin embargo, ninguno de ellos pasa del minuto por lo cual se podría decir que este algoritmo sigue teniendo una complejidad baja. Por otro lado, el algoritmo de construcción del grafo de sobre lapes tiene una complejidad mucho mayor, llegando a tener un tiempo de ejecución máximo mayor a 2 minutos. Todos los algoritmos corrieron con todos los archivos simulados.

Comando	Cantidad de lecturas máxima que puede procesar

Construcción tabla de K-mers	~59000 lecturas (RD=100X, RL=50)
Construcción grafo de sobre lape	~59000 lecturas (RD=100X, RL=50)
Ensamblaje de secuencia	~59000 lecturas (RD=100X, RL=50)

Nota: dado que en la consigna de este punto daban como parámetro para probar distintas profundidades, modificamos el código de tal manera que el arg[3] “SimpleReadsSimulator” en fuese la profundidad deseada y no el número de reads. Como consecuencia el número de read se calcula dentro del método tomando en cuenta el tamaño de la secuencia fasta, la cobertura y el tamaño de los reads.

5. [10%] Mejorar el script reads simulator incluyendo un parámetro nuevo que permita simular una tasa de error aleatoria. Simular datos con un tamaño de lectura de 50bp y profundidad promedio de 20X. Visualizar y comparar la distribución de abundancia de k-mers en cada caso.

Al realizar el análisis, el número de kmers disminuye consedirablemente, por lo tanto no se pudo realizar la gráfica de distribución de kmers:

```
Time building k-mers table(ms): 89
Total number of k-mers: 4
The k-mer CAGGTAACCTACAAAGCAACCATGATCTGTATTGTCAAGTCCATGGTAAT is present 1 times
The k-mer TTTTGAAATTAAATTGGCAAAGAAATTTGACACCTTCAATGGGGAATGTC is present 1 times
The k-mer TGCAGATTTGGATGATTTCTCCAAACAATTGCAACAATCCATGAGCAGTG is present 1 times
The k-mer AAACAGGGTAATTTCAAAAATCTTAGTGAATTTGTGTTTAAGAATATTGA is present 1 times
K-mer distribution
1       4
2       0
3       0
4       0
5       0
6       0
7       0
8       0
9       0
10      0
11      0
12      0
13      0
14      0
15      0
```