Universidad de los Andes

BCOM4102-Ecologia Microbiana

Programa de Maestría en Biología Computacional

**Taller Semana 9: Ensamblaje de Genomas y Metagenomas**

Abril 08 de 2021

**Taller Semana 9**

Inicie una sesión en el cluster (magnus.uniandes.edu.co), recuerde hacerlo vía ssh (desde Terminal o MobaXterm), su login es bcom4102 y el password es 202110bcom4102.

**Objetivos:** Después de ejecutar este tutorial, usted debería poder:

**1.** Describir los pasos generales involucrados en el ensamblaje de genomas y metagenomas a partir de secuencias de lectura corta.

2. Evaluar la calidad de un ensamblaje y distinguir las características de un buen ensamblaje a nivel de comunidad y de genomas individuales.

3. Mapear lecturas de la comunidad a los contigs generados en un ensamblaje de metagenomas y determinar otros aspectos de su calidad a partir de ello.

4. Considerar los efectos de longitud de *k*-mero en un ensamblaje**.**

**Datos:** Localización de los datos crudos:

/hpcfs/home/ciencias/biologia/cursos/bcom4102/Datasets/Taller\_Semana9

**Taller nuestro:** /hpcfs/home/ciencias/biologia/cursos/bcom4102/patiño\_leon\_diaz/Taller8/Taller\_Semana9

**- Genoma:**

Datos de Escherichia coli str. K-12 substr. MG1655 número de accesión [SAMN04147053](https://www.ebi.ac.uk/ena/browser/view/SRR2627175) obtenidos a partir de secuenciación con Illumina MiSeq PE250.

*SRR2627175\_1.fastq.gz*

*SRR2627175\_2.fastq.gz*

- M**etagenoma**:

Comunidad artificial (*mock*) obtenida a partir del recurso [mockrobiota](http://mockrobiota.caporasolab.us/). La comunidad *mock* a utilizar corresponde a mock-17, y en [este enlace](https://github.com/caporaso-lab/mockrobiota/tree/master/data/mock-17) se puede obtener una idea general de la taxonomía esperada de la comunidad bacteriana.

*mock-forward.fastq.gz*

*mock-reverse.fastq.gz*

**Recursos:** A lo largo del documento encontrará para su consulta enlaces a los manuales de cada una de las herramientas a utilizar en el desarrollo del taller. Allí encontrará cada uno de los comandos detallando las posibles opciones a usar.

**INDICACIONES**

**Sección I:** Genomas individuales

Como podrá verificar, los archivos de secuencia para el genoma de *E. coli* contienen más de 2 millones de lecturas. Solo usaremos un subconjunto del conjunto de datos original para este taller.

**1.** Cree un directorio de datos en su carpeta de grupo que tenga el nombre “ensamblaje\_genoma”

2. Cargue el módulo de [Seqtk](https://github.com/lh3/seqtk) “seqtk/1.2” en Magnus y trabaje con una muestra del 25% de las secuencias de estos archivos. Busque el comando en [Seqtk](https://github.com/lh3/seqtk) para realizar esta labor y nombre los archivos de salida como “Ecoli\_0.25\_1.fastq” y “Ecoli\_0.25\_2.fastq”, y ubíquelos en el directorio que creó en el paso anterior.

**seqtk sample -s100** SRR2627175\_2.fastq.gz **0.25** > Ecoli\_0.25\_2.fastq

**seqtk sample -s100** SRR2627175\_1.fastq.gz **0.25** > Ecoli\_0.25\_1.fastq

La calidad del ensamblaje depende en gran medida del tamaño de *k*-mero utilizado. El tamaño ideal de *k*-mero depende de la longitud de lectura y la profundidad de secuenciación y la diversidad de la muestra secuenciada.

3. Cargue el módulo de [SPAdes](https://github.com/ablab/spades/blob/spades_3.15.2/README.md#sec3.2) “spades/3.14.0” disponible en Magnus y realice dos ensamblajes, uno con longitud de *k*-mero 61 y otro con longitud 91. Nombre los archivos de modo que reflejen la diferencia en el parámetro. Adicional a la especificación de k, utilice la opción “--only-assembler” en su comando. Recuerde también especificar que los archivos de entrada son pareados.

**spades.py –k** 61,91 **-1** Ecoli\_0.25\_1.fastq **-2** Ecoli\_0.25\_2.fastq **--only-assembler -o** Ecoli\_0.25\_ensamblados

/hpcfs/home/ciencias/biologia/cursos/bcom4102/patino\_leon\_diaz/Taller8/Taller\_Semana9/ensamblaje\_genoma/Ecoli\_0.25\_ensamblados/K61

4. Ahora evalúe la calidad de los dos ensamblajes. Utilice [QUAST](http://quast.sourceforge.net/quast), a través del módulo “quast/5.0.2” en Magnus. Utilice como referencia el genoma de *E. coli* strain K-12 substr. MG1655 disponible [aquí](https://www.ncbi.nlm.nih.gov/assembly/GCF_000005845.2).

**quast.py** final\_contigs.fasta **-r** ref\_gen.txt

**Para entregar:** Reportes de QUAST y una explicación que resulte de comparar los ensamblajes realizados con las dos longitudes de *k*-mero.

**REPORTES**

**Tabla 1.** Reporte de métricas de ensamblaje usando QUAST

|  |  |
| --- | --- |
| **QUAST K-61** | **QUAST K-91** |
| **Figura 1.** Nx del ensamblaje realizado  con k-meros= 61 | **Figura 11.** Nx del ensamblaje realizado  con k-meros= 91 |
| **Figura 2.** NGx del ensamblaje realizado  con k-meros= 61 | **Figura 12.** NGx del ensamblaje realizado  con k-meros= 91 |

Con respecto al Nx, la longitud más larga de un contig, es mayor cuando se escogen Kmeros de mayor tamaño y la cantidad de contigs con dicho tamaño es mayor

**Tabla 2.** Reporte de métricas de ensamblaje usando QUAST

|  |  |
| --- | --- |
| **QUAST K-61** | **QUAST K-91** |
| **Figura 3.** Longitud acumulada del ensamblaje realizado con k-meros= 61 | **Figura 13.** Longitud acumulada del ensamblaje realizado con k-meros= 91 |
| **Figura 4.** Contenido de GC del ensamblaje realizado con k-meros= 61 | **Figura 14.** Contenido de GC del ensamblaje realizado con k-meros= 91 |

Con respecto a 3 y 13, es evidente como cuando se hacen los ensamblajes con contigs con un número mayor de Kmeros, el número de bases es igual con una menor cantidad de contigs lo cual tiene sentido. Con respecto al porcentaje de GC, resulta evidente que este no cambia con el K.

**Tabla 3.** Reporte de métricas de ensamblaje usando QUAST

|  |  |
| --- | --- |
| **QUAST K-61** | **QUAST K-91** |
| **Figura 5.** Histograma de cobertura del ensamblaje realizado con k-meros= 61 | **Figura 15.** Histograma de cobertura del ensamblaje realizado con k-meros= 91 |
| **Figura 6.** Número de translocaciones presentes en el ensamblaje realizado con k-meros= 61 | **Figura 16.** Número de translocaciones presentes en el ensamblaje realizado con k-meros= 91 |

Con respecto al número de ensamblajes erróneos, resulta evidente que cuando se usan kmeros de mayor longitud, se aumenta dicho valor. Con respecto al histograma de cobertura, es evidente que dependiendo de la longitud de kmero usada, la uniformidad en la cobertura de secuenciación detectada cambia, siendo un poco más homogénea con kmeros más largos.

**Tabla 4.** Reporte de métricas de ensamblaje usando QUAST

|  |  |
| --- | --- |
| **QUAST K-61** | **QUAST K-91** |
| **Figura 7.** Número de translocaciones presentes en el ensamblaje realizado con k-meros= 61 | **Figura 17.** Número de translocaciones presentes en el ensamblaje realizado con k-meros= 91 |
| **Figura 8.** NAx d el ensamblaje realizado con k-meros= 61 | **Figura 18.** NAx d el ensamblaje realizado con k-meros = 91 |

**Tabla 5.** Reporte de métricas de ensamblaje usando QUAST

|  |  |
| --- | --- |
| **QUAST K-61** | **QUAST K-91** |
| **Figura 9.** NGAx del ensamblaje realizado con k-meros= 61 | **Figura 19.** NGAx del ensamblaje realizado con k-meros= 91 |
| **Figura 10.** Longitud acumulada del ensamblaje realizado con k-meros= 61 | **Figura 20.** Longitud acumulada del ensamblaje realizado con k-meros= 91 |

**Tabla 6.** Reporte de métricas más importantes luego de realizar el ensamblaje usando QUAST

|  |  |  |
| --- | --- | --- |
| **Métrica calculada** | **Contigs finales con K=61** | **Contigs finales con K=91** |
| **# contigs (>= 0 bp)** | 562 | 195 |
| **# contigs (>= 1000 bp)** | 141 | 73 |
| **# contigs (>= 5000 bp)** | 106 | 56 |
| **# contigs (>= 10000 bp)** | 89 | 51 |
| **# contigs (>= 25000 bp)** | 59 | 41 |
| **# contigs (>= 50000 bp)** | 32 | 27 |
| **Total length (>= 0 bp)** | 4585926 | 4583272 |
| **Total length (>= 1000 bp)** | 4514581 | 4544757 |
| **Total length (>= 5000 bp** | 4428472 | 4508882 |
| **Total length (>= 10000 bp)** | 4290665 | 4469853 |
| **Total length (>= 25000 bp)** | 3803506 | 4299141 |
| **Total length (>= 50000 bp)** | 2819989 | 3790466 |
| **# contigs** | 162 | 83 |
| **Largest contig** | 169244 | 414033 |
| **Total length** | 4529468 | 4551166 |
| **Reference length** | 4091782 | 4091782 |
| **GC (%)** | 50.74 | 50.74 |
| **Reference GC (%)** | 51.83 | 51.83 |
| **N50** | 60350 | 148541 |
| **NG50** | 73698 | 173413 |
| **N75** | 35580 | 66400 |
| **NG75** | 42739 | 95078 |
| **L50** | 23 | 10 |
| **LG50** | 19 | 9 |
| **L75** | 46 | 21 |
| **LG75** | 38 | 17 |
| **# misassemblies** | 3793 | 3865 |
| **# misassembled contigs** | 136 | 66 |
| **Misassembled contigs length** | 4493543 | 4524077 |
| **# local misassemblies** | 1 | 3 |
| **# scaffold gap ext. mis.** | 0 | 0 |
| **# scaffold gap loc. mis.** | 0 | 0 |
| **# unaligned mis. contigs** | 1 | 2 |
| **# unaligned contigs** | 4 + 81 part | 3 + 51 part |
| **Unaligned length** | 503889 | 550840 |
| **Genome fraction (%)** | 96.467 | 96.731 |
| **Duplication ratio** | 1.020 | 1.011 |
| **# N's per 100 kbp** | 0.00 | 0.00 |
| **# mismatches per 100 kbp** | 0.46 | 2.20 |
| **# indels per 100 kbp** | 0.13 | 0.30 |
| **Largest alignment** | 7077 | 7077 |
| **Total aligned length** | 3945872 | 3957212 |
| **NA50** | 1066 | 1068 |
| **NGA50** | 1161 | 1167 |
| **NA75** | 648 | 648 |
| **NGA75** | 780 | 786 |
| **LA50** | 1396 | 1400 |
| **LGA50** | 1200 | 1194 |
| **LA75** | 2729 | 2738 |
| **LGA75** | 2269 | 2257 |

**Tabla 7.** Reporte de métricas más importantes luego de realizar el ensamblaje usando QUAST

|  |  |  |
| --- | --- | --- |
| **Métrica calculada** | **Contigs finales con K=61** | **Contigs finales con K=91** |
| **# misassemblies** | 3793 | 3865 |
| **# contig misassemblies** | 3793 | 3865 |
| **# c. relocations** | 0 | 0 |
| **# c. translocations** | 3793 | 3865 |
| **# c. inversions** | 0 | 0 |
| **# scaffold misassemblies** | 0 | 0 |
| **# s. relocations** | 0 | 0 |
| **# s. translocations** | 0 | 0 |
| **# s. inversions** | 0 | 0 |
| **# misassembled contigs** | 136 | 66 |
| **Misassembled contigs length** | 4493543 | 4524077 |
| **# local misassemblies** | 1 | 3 |
| **# scaffold gap ext. mis.** | 0 | 0 |
| **# scaffold gap loc. mis.** | 0 | 0 |
| **# unaligned mis. contigs** | 1 | 2 |
| **# mismatches** | 18 | 87 |
| **# indels** | 5 | 12 |
| **# indels (<= 5 bp)** | 5 | 12 |
| **# indels (> 5 bp)** | 0 | 0 |
| **Indels length** | 6 | 13 |

**Tabla 8.** Reporte de métricas más importantes luego de realizar el ensamblaje usando QUAST

|  |  |  |
| --- | --- | --- |
| **Métrica calculada** | **Contigs finales con K=61** | **Contigs finales con K=91** |
| # fully unaligned contigs | 4 | 3 |
| Fully unaligned length | 5348 | 5246 |
| # partially unaligned contigs | 81 | 51 |
| # partially unaligned contigs | 498541 | 545594 |
| # N's | 0 | 0 |

Al comparar los resultados obtenidos en el ensamblaje del genoma con diferentes valores de k-mero podemos concluir que:

* El largo de los contigs cambia en un factor mayor a 2x cuando comparamos el ensamblaje realizado con k-mero 61 con el realizado con el k-mero 91.
* La longitud acumulada es similar entre ambos ensamblajes, sin embargo, varían bastante al compararlos con el ensamblaje de referencia. Esto se podría atribuir al hecho de que estamos trabajando con un sub-conjunto del 25% con respecto al genoma original.
* La cobertura del ensamblaje con k-mero 91 es mucho más alta con respecto a la cobertura con k-mero 61, sin embargo, la profundidad de la cobertura es menor.

Finalmente, no se observan diferencias representativas en los apartados de: número de translocaciones; NGAx; NAx; longitud acumulada de contigs alineados; y contenido de GC del ensamblaje.

**Sección II: Metagenomas**

Estos datos ya han sido limpiados y filtrados por calidad. De nuevo, solo usaremos un subconjunto del conjunto de los datos originales para esta sección**.**

**/hpcfs/home/ciencias/biologia/cursos/bcom4102/patino\_leon\_diaz/Taller8/Taller\_Semana9/ensamblaje\_genoma/Ecoli\_0.25\_k91/K61/final\_contigs.fasta**

**/hpcfs/home/ciencias/biologia/cursos/bcom4102/patino\_leon\_diaz/Taller8/Taller\_Semana9/ensamblaje\_genoma/Ecoli\_0.25\_k91/contigs.fasta**

1. Cree un directorio de datos en su carpeta de grupo que tenga el nombre “ensamblaje\_metagenoma”

2. Cargue el módulo de [Seqtk](https://github.com/lh3/seqtk) “seqtk/1.2” en Magnus y trabaje con una muestra del 25% de las secuencias de estos archivos. Busque el comando en Seqtk para realizar esta labor y nombre los archivos de salida para reflejar este procesamiento y ubíquelos en el directorio que creó en el paso anterior.

**seqtk sample -s100** mock-forward.fastq.gz **0.25** > metF.fastq

**seqtk sample -s100** mock-reverse.fastq.gz **0.25** > metR.fastq

3. Realice el ensamblaje del metagenoma con los parámetros por defecto para librerías pareadas usando el módulo “megahit/1.1.1” en Magnus del software [Megahit](https://github.com/voutcn/megahit).

megahit -1 metF.fastq -2 metR.fastq -o out

**Para entregar:**

**¿Cuáles son los parámetros por defecto del programa? Consulte.**

De acuerdo con el flag de ayuda (-h) para el programa instalado en el clúster de la Universidad de los Andes, los valores por defecto de los flags se colorearán de verde, tal y como se muestra a continuación:

**megahit [options]** **-1** mock-forward.fastq.gz **-2** mock-reverse.fastq.gz **-o mock\_assembly**

-**-min-count**  minimum multiplicity for filtering (k\_min+1)-mers **[2]**

**--k-list**  comma-separated list of kmer size

all must be odd, in the range 15-255, increment <= 28);

**[21,29,39,59,79,99,119,141]**

**--k-min**  minimum kmer size (<= 255), must be odd number **[21]**

**--k-max** maximum kmer size (<= 255), must be odd number **[141]**

**--k-step** increment of kmer size of each iteration (<= 28), must be even number **[12]**

**--bubble-level** intensity of bubble merging (0-2), 0 to disable **[2]**

**--merge-level** merge complex bubbles of length <= l\*kmer\_size and similarity >= s

**[20,0.95]**

**--prune-level** strength of low depth pruning (0-3) **[2]**

**--prune-depth** remove unitigs with avg kmer depth less than this value **[2]**

**--low-local-ratio** ratio threshold to define low local coverage contigs **[0.2]**

**--mem-flag** SdBG builder memory mode 0: minimum; 1: moderate; others:

use all memory specified by '-m/-- memory' **[1]**

**¿Qué parámetros considera que sería útil modificar? Justifique su respuesta.**

Consideramos que sería útil modificar parámetros relacionados con el k-mero teniendo en cuenta que se puede obtener la información sobre la longitud de la lectura y la profundidad de la secuenciación, lo que nos facilitaría escoger un valor adecuado para obtener finalmente una expresión de genes óptima en el proceso de ensamblaje del metagenoma.

4. De nuevo utilice QUAST (módulo indicado en la Sección I) para revisar los estadísticos del ensamblaje.

quast.py final\_contigs.fa -R ref\_met.txt

**Reporte de QUAST.**

**Tabla 9.** Reporte de métricas más importantes luego de realizar el ensamblaje del metagenoma usando QUAST

|  |  |
| --- | --- |
| **Métrica calculada** |  |
| **# contigs (>= 0 bp)** | 30667 |
| **# contigs (>= 1000 bp)** | 1824 |
| **# contigs (>= 5000 bp)** | 7 |
| **# contigs (>= 10000 bp)** | 0 |
| **# contigs (>= 25000 bp)** | 0 |
| **# contigs (>= 50000 bp)** | 0 |
| **Total length (>= 0 bp)** | 16677799 |
| **Total length (>= 1000 bp)** | 2641994 |
| **Total length (>= 5000 bp** | 41034 |
| **Total length (>= 10000 bp)** | 0 |
| **Total length (>= 25000 bp)** | 0 |
| **Total length (>= 50000 bp)** | 0 |
| **# contigs** | 12332 |
| **Largest contig** | 6817 |
| **Total length** | 9588217 |
| **Reference length** | 167681 |
| **GC (%)** | 46.66 |
| **Reference GC (%)** | 53.41 |
| **N50** | 752 |
| **NG50** | 3910 |
| **N75** | 600 |
| **NG75** | 3569 |
| **L50** | 4338 |
| **LG50** | 17 |
| **L75** | 7935 |
| **LG75** | 28 |

|  |  |
| --- | --- |
| **Figura 21.** Nx resultante del metagenoma ensamblado | **Figura 23.** NGx resultante del metagenoma ensamblado |
| **Figura 22.** Curva acumulatica resultante del metagenoma ensamblado | **Figura 24.** Contenido GC resultante del metagenoma ensamblado |

Del cuadro anterior resulta interesante ver como el contenido de GC presenta dos picos bien marcados a diferencia del metagenoma de referencia, diferencia que también resulta evidente en la gráfica de longitud acumulativa en la que resulta evidente que existe una discrepancia entre el metagenoma ensamblado y el metagenoma de referencia

**¿Qué le indican la longitud total, el número de contigs totales, el número de contigs > 1Kb, y las métricas de N50 y L50?**

**Particularmente en ensamblaje de metagenomas, es importante saber cuál es la contribución de las lecturas obtenidas al ensamblaje resultante. No es lo mismo que los contigs reflejen la información de 80% de las secuencias iniciales a que reflejen solo el 10% de las secuencias de una comunidad dada. Para revisar esto vamos a realizar un procedimiento de mapeo**.

Todos los métodos de ensamblaje se basan en la simple suposición de que fragmentos de ADN altamente similares se originan de la misma posición dentro del genoma. De esta manera, la similitud entre secuencias de ADN se usa para conectar fragmentos individuales en secuencias contiguas más largas, es allí donde corresponde el termino contigs y su numero total en el emsamblado (secuencias consenso obtenidas a partir del ensamblaje de los reads).

Un contig todavía se refiere a un contiguo de datos de secuencia creado por superposición de reads. Debido a que los fragmentos son de longitud conocida, se conoce la distancia entre las dos lecturas de los extremos de cada fragmento, siendo esta su longitud total Además, la longitud de los contigs tiene un efecto importante debido a que las características de cobertura y composición pueden ser mejor estimadas a partir de contigs largos, y así aproximar los valores para la especie microbiana en cuestión.Todo ello requiere Control de Calidad y de uso de allí de su criterios de longitud como el es caso >1Kb , se selección en longitud mayor a un (1) KB, así que se hará, control, limpieza y eliminar todos aquellos contigs que tengan menos de ello.

En biología computacional, N50 y L50 son estadísticas de un conjunto de longitudes de contig o estructura. El N50 es similar a una media o mediana de longitudes, pero tiene mayor peso dado a los contigs más largos. Se usa ampliamente en el ensamblaje del genoma, especialmente en referencia a longitudes de contig dentro de un ensamblaje de borrador. N50, L50 y estadísticas relacionadas -

N50 La estadística N50 define la calidad del ensamblaje en términos de contigüidad

Dado un conjunto de contigs, el N50 se define como la longitud de secuencia del contig más corto al 50% de la longitud total del genoma. Puede considerarse como el punto de la mitad de la masa de la distribución; el número de bases de todos los contigs más largos que el N50 estará cerca del número de bases de todos los contigs más cortos que el N50.

Por otro lado, un L50 Dado un conjunto de contigs, cada uno con su propia longitud, el recuento L50 se define como el número más pequeño de contigs cuya suma de longitud constituye la mitad del tamaño del genoma.

5. Utilice [Bowtie2](http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml), módulo “bowtie/2.4.1” en Magnus, para mapear las lecturas iniciales al archivo resultante “final.contigs.fa” producido por Megahit.

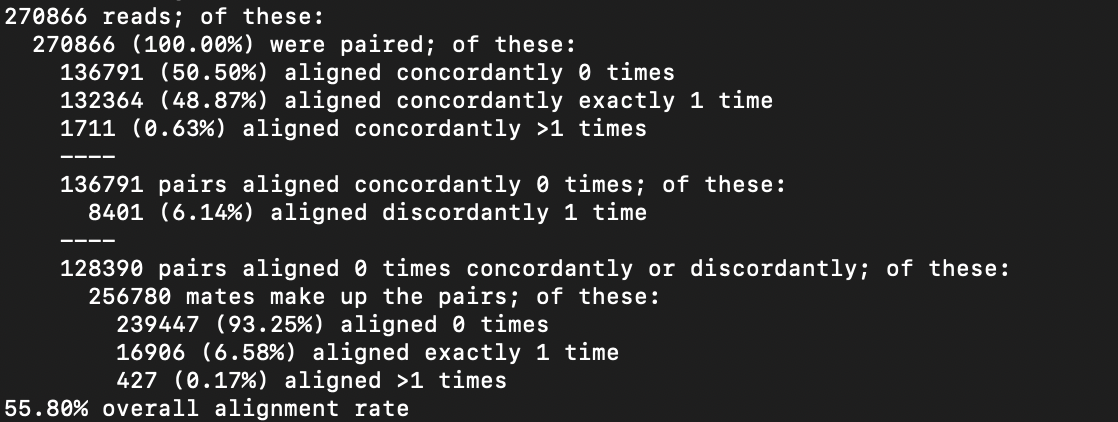
Primero cree el índice usando el comando *bowtie-build*. El comando requiere 2 argumentos. El primer argumento es la referencia FASTA (“final.contigs.fa”). El segundo argumento es el nombre de archivo "base" que se utilizará para los archivos de índice creados. Creará un montón de archivos que comienzan con *nombre\_elegido\_índice*\*.

bowtie2-build **final\_contigs.fa mapLec**

Luego, corra el mapeo como tal: solo le diste correr ¿= o como sacamos esto ¿

**bowtie2 -x** *nombre\_elegido\_índice* -1 archivo\_1.fastq **-2** archivo\_2.fastq **-S** mock.sam

**bowtie2 -x** mapLec**-1** metF.fastq **-2** metR.fastq **-S** mock.sam



**¿Qué porcentaje de lecturas pudo mapear a los contigs generados? ¿Qué conclusiones podría sacar?**

El porcentaje de lecturas que se pudo mapear es de 55.8%. Por lo anterior podríamos decir que los contigs reflejan la información de más de la mitad de las secuencias por lo que se podría concluir que el proceso de alineamiento de las lecturas fue relativamente exitoso y que se logró la correcta alineación de lecturas teniendo en cuenta el genoma. También nos parece importante resaltar que el 0.63% de lecturas fue alineado consistentemente más de una vez, lo que podría significar que existen repeticiones en el genoma de referencia (por ejemplo muchas copias de un gen).

**¿Cómo podría cambiar esta visión si solo se mantienen los contigs > 2Kb?**

Si se mantienen solo los contigs > 2Kb cambiaría el número de lecturas que son alineadas más de una vez, ya que usualmente este proceso sucede cuando las lecturas son pequeñas y se interpreta como repeticiones en el genoma de referencia. Esto resultaría útil ya que a veces es difícil identificar la procedencia de dichas repeticiones por lo que varios pipelines las ignoran por defecto.

**¿Para qué otras aplicaciones son útiles la técnica de mapeo a (meta)genomas de referencia?**

La técnica de mapeo es útil también para identificar la localización de mutaciones o rasgos específicos en un cromosoma; esta información es importante ya que cuanto más cerca están dos marcadores es más probable que se hereden juntos y no se separen en un cruce genético. De la misma manera, la frecuencia con la que se recombinan los marcadores genéticos se puede obtener a través de un mapeo y se mide en centiMorgans que representa la separación de marcadores genéticos en un experimento de cruce genético.

**REFERENCIAS**

1. Batut et al., 2018 Community-Driven Data Analysis Training for Biology Cell Systems [10.1016/j.cels.2018.05.012](https://doi.org/10.1016%2Fj.cels.2018.05.012)
2. N50, L50, and related statistics - Wikipedia. (2021). Retrieved 11 April 2021, from <https://es.qaz.wiki/wiki/N50,_L50,_and_related_statistics>
3. Visualización y curación de genomas ensamblados desde metagenomas (MAGs). (2021). Retrieved 11 April 2021, from http://www.castrolab.org/isme/mags/mags.html
4. Brown TA. Genomes. 2nd edition. Oxford: Wiley-Liss; 2002. Chapter 5, Mapping Genomes. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK21116/>
5. Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075. doi: 10.1093/bioinformatics/btt086
6. Sequencing Coverage for NGS Experiments. (2021). Retrieved 12 April 2021, from <https://www.illumina.com/science/technology/next-generation-sequencing/plan> experiments/coverage.html