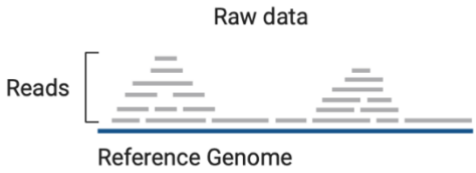
UNIVERSIDADE DA CORUÑA

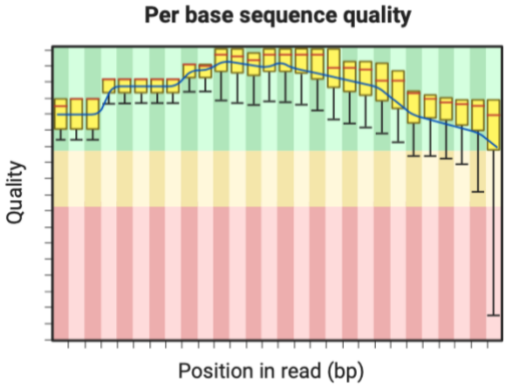
FACULTAD DE INFORMÁTICA

MUBICS

Cuaderno de prácticas

Estructuras de datos y algoritmia para secuencias biológicas





Pedro Sánchez García

Curso 2021-2022

Profesor:
Dr. Fernando Silva Coira

PRÁCTICA 4. HERRAMIENTAS PARA EL ENSAMBLAJE DE SECUENCIAS

OBJETIVO

Conocer y utilizar herramientas existentes para ensamblaje y visualización de alineamientos de secuencias.

TAREAS (1 punto)

Realiza la práctica siguiendo los siguientes pasos:

1.1 Analiza la calidad de los ficheros de lecturas (FASTQC)

Se evalúa la calidad de dos ficheros con formato FASTQ que se han generado mediante tecnología NGS a través del programa FASTQC. En primer lugar, cabe destacar que los ficheros constan de lecturas (*reads*) generadas a partir de muestras biológicas. En concreto, estos datos pueden corresponder a las secuencias de ADN obtenidas tras la amplificación de un fragmento de un determinado gen y posterior secuenciado NGS por un sistema como los de Illumina u Oxford Nanopore.

Si abrimos con un editor de texto los ficheros, apreciamos que para cada *read* se sigue la "sintaxis" característica del formato FASTQ:

```
@24960/1
AATGTTGTCACTTGGATTCAAATGACATTTTAAATCTAATTATTCATGAATCGAACTAGTACGAAATGCAATGAGCATCTTGTCTAGTTCGATTTTTTAATGTCTAAAAATGTCGTATATGTAATCAGAGTAGAAAGTGTTGAGGCGTTT
+
5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHIIHIIAHDHIIHIG#IIHIFHHHFGIII*IHHHIHFIIHGICIHHIHFFFHHHIIIIIHIHDHIIIAHHH?GHHHHHF@HGGH6GGHEEGBGGGGGFGFE6FGFEFE?GF5EGGEEAC
```

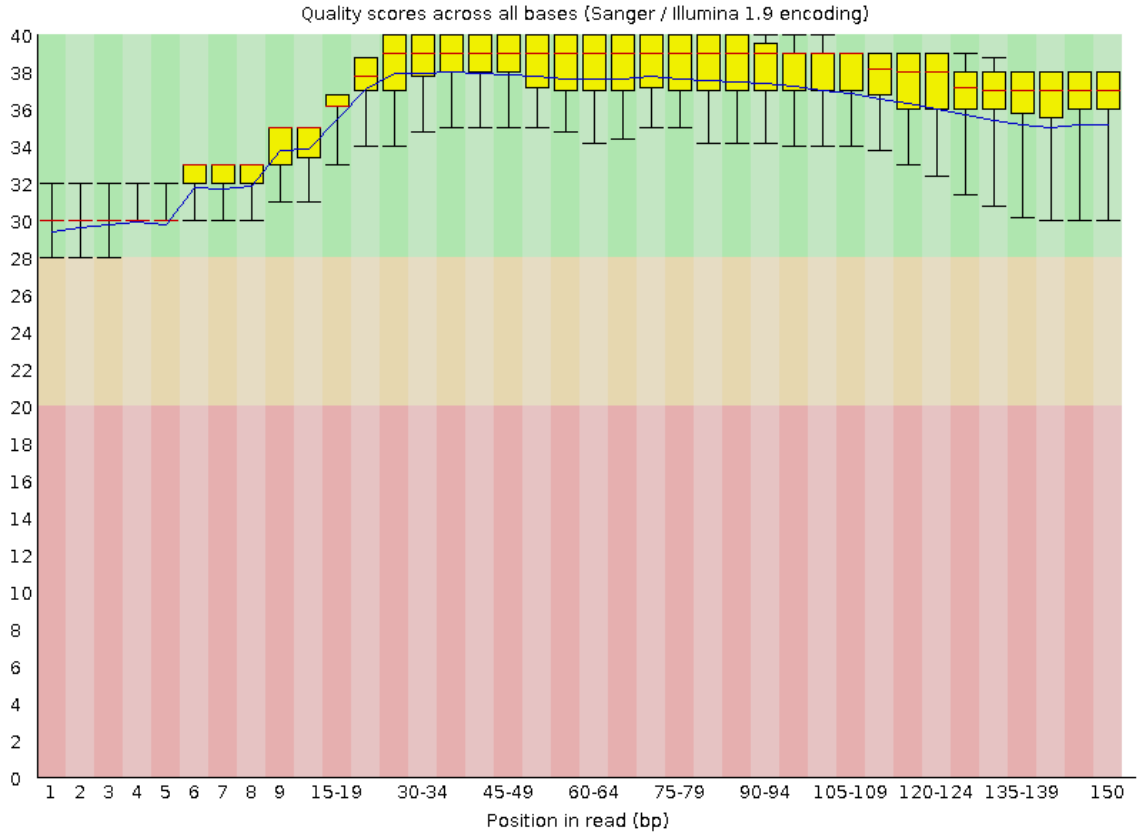
Tal y como se muestra en la imagen anterior, se distinguen 4 campos o líneas, donde la primera corresponde con la identificación de la secuencia, es decir, el nombre que le proporciona el secuenciador. La segunda línea es la secuencia de nucleótidos que ha leído el secuenciador, mientras que la tercera línea es un símbolo + que actúa como espaciador. Por último, la cuarta línea hace referencia a diferentes símbolos, números y/o letras (tantos como nucleótidos se han secuenciado), de tal forma que cada uno de ellos muestra el valor de la fiabilidad de la lectura en un nucleótido concreto: dicho de otro modo, indican la cobertura de este. Los símbolos o códigos en la calidad de lectura se denominan *Phred score (Q)* y se basan principalmente en un logaritmo que evalúa la probabilidad de lectura errónea en cada base concreta.

Cabe destacar que FASTQC detecta automáticamente la codificación de *Phred* en el fichero. De este modo, lo que nos interesa en investigaciones y proyectos es la consideración de lecturas fiables, por lo que centraremos la atención en un valor superior a 28 puntos.

A través de la ejecución de FASTQC en Galaxy, se alcanza el siguiente resultado para el fichero *reads1.fastq*, donde se puede apreciar un resumen adicional al resultado:

Basic Statistics

Measure	Value
Filename	reads1_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	12480
Sequences flagged as poor quality	0
Sequence length	150
%GC	33

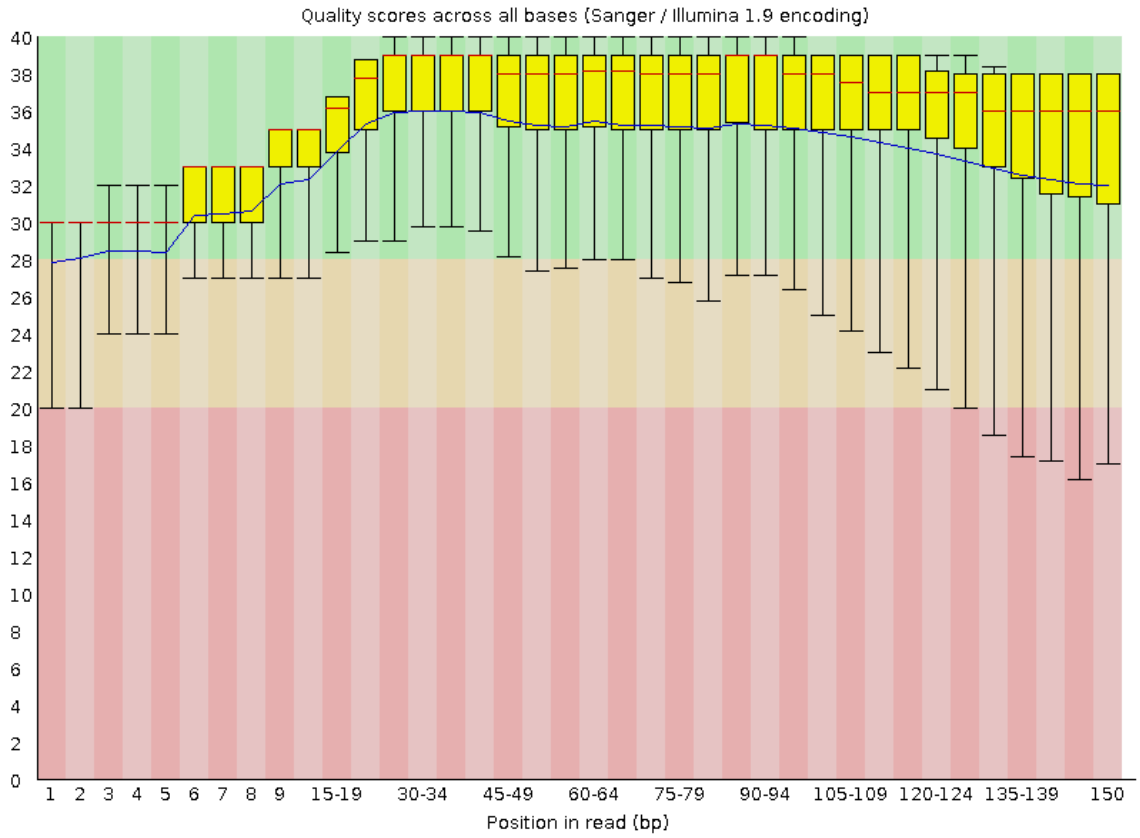


Teniendo en cuenta el apartado *basic statistics*, podemos observar que el fichero contiene 12.480 *reads*, donde su longitud de secuencia es de 150 pares de bases (bp). Además, se refleja que la plataforma con las que se han generado las *reads* es un sistema con la codificación Sanger / Illumina 1.9. Pasando al resultado gráfico, se puede apreciar que la totalidad de los nucleótidos presenta un *Quality score* elevado, ubicándose en la franja verde que comprende valores de 28 a 40. Si nos fijamos en los 5 primeros nucleótidos, estos no presentan una dispersión en los datos como sucede en el resto. Esto se podría asociar fundamentalmente a los adaptadores empleados por el sistema para el protocolo de la secuenciación. Además, en general, la mediana de Q se mantiene más elevada en las posiciones de 6 a 15-19, con mayor uniformidad en las posiciones restantes de las *reads*.

En lo que respecta al fichero *reads2.fastq*, se alcanzan los siguientes resultados:

Basic Statistics

Measure	Value
Filename	reads2_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	12480
Sequences flagged as poor quality	0
Sequence length	150
%GC	33



Visualizando el resultado gráfico, se detectan cambios notables en comparación con el fichero *reads1*, pues a pesar de que la totalidad de los nucleótidos presenta un *Quality score* elevado, la dispersión en los datos es notablemente superior. Además, incluso en el caso de los 5 primeros nucleótidos, estos presentan una dispersión en los datos similar al resto. De nuevo, la mediana de Q se mantiene más elevada en las posiciones de 6 a 15-19, con una mayor uniformidad en las posiciones restantes de las *reads*.

Cabe destacar que en este tipo de tecnología *paired-end sequencing*, la tendencia con los *reverse reads* es la que se ha obtenido en este caso con el fichero *reads2* que las contiene. Esto se debe al tiempo que el ADN pasa en el sistema hasta finalizar la secuenciación.

1.2 Realiza un ensamblaje *de novo* con las lecturas (SPAdes)

Mediante la selección del tipo *paired-end sequencing*, se selecciona el fichero *reads1.fastq* como el que contiene las *reads forward* y el fichero *reads2.fastq* como el que presenta las *reads reverse.* Con respecto a los parámetros que se pueden modificar, dado que no se especifica nada al respecto en el enunciado, se opta por mantenerlos por defecto.

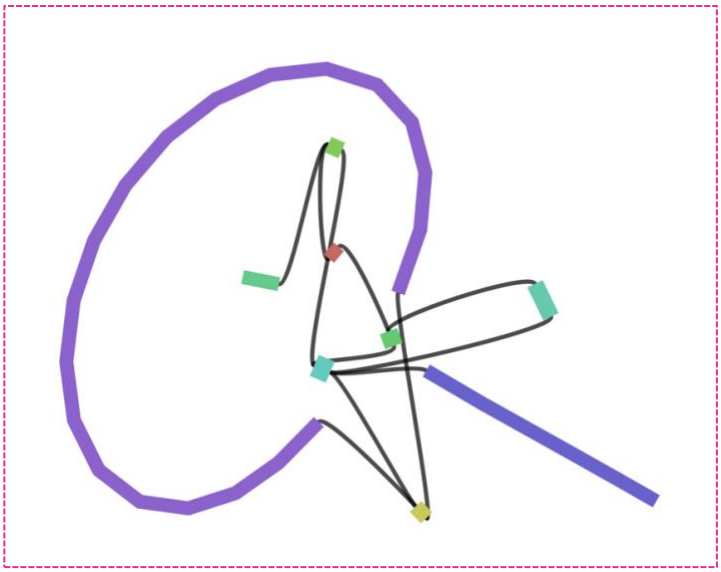
De esta forma, se ha planteado la siguiente instrucción en la línea de comandos:

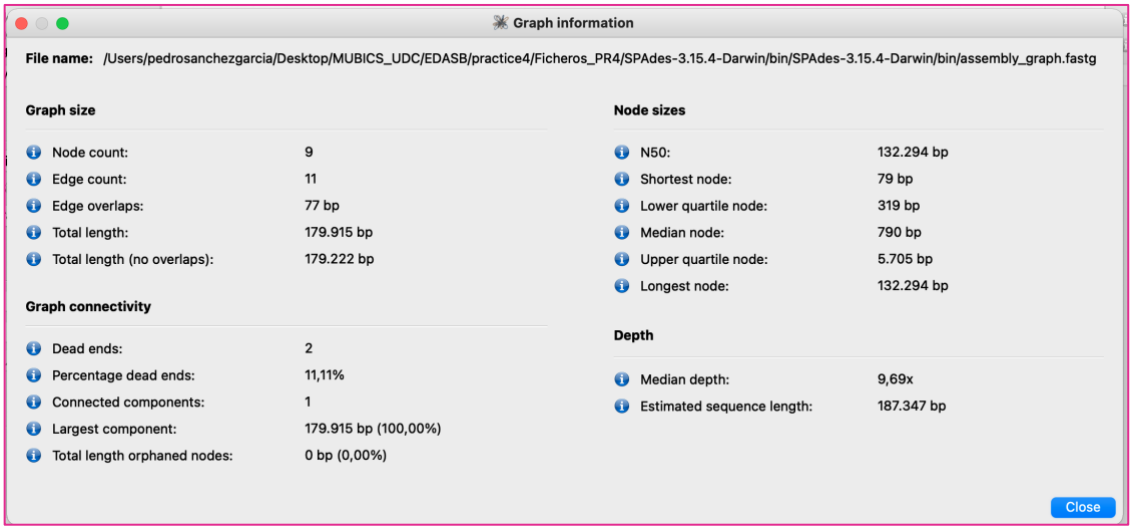
./spades.py -1 reads1.fastq -2 reads2.fastq -o SPAdes-3.15.4-Darwin/bin/

Como resultado, se generan unas salidas por separado correspondientes a los *scaffolds* y *contigs* generados, grafo de ensamblaje y el grafo de ensamblaje con *scaffolds*.

1.3 Visualiza el grafo de ensamblaje (Bandage)

Mediante la apertura del archivo en formato *fastg* del grafo de ensamblaje obtenido en SPAdes, se alcanzan los siguientes resultados, correspondientes a la visualización del grafo y al resumen general de este:





En base a lo anterior, podemos destacar la existencia de 9 nodos, con una longitud total de 179.915 pb. Atendiendo a la longitud de estos, cabe destacar que el más corto presenta 79 pb, mientras que el de longitud intermedia posee 790 pb. Por su parte, el más largo posee una longitud de 132.294 pb.

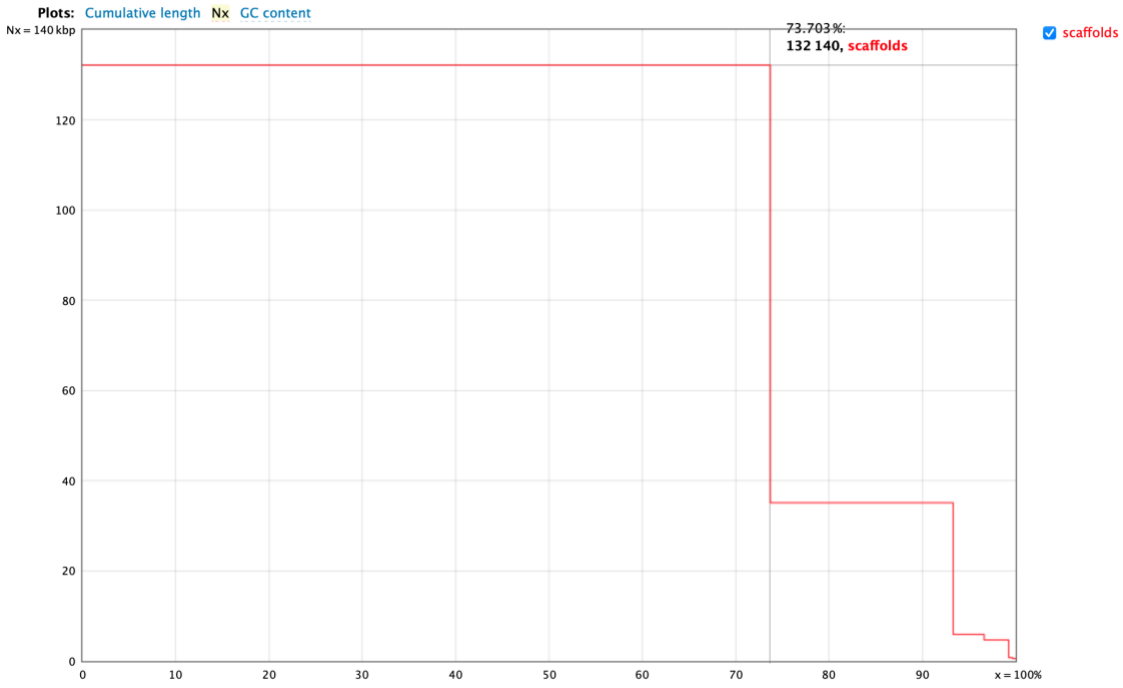
Por tanto, de acuerdo con la teoría expuesta en clase, el fundamento es que se trata de un grado donde se distingue 1 nodo para cada *k-mer* en los patrones de solapamiento. Posteriormente, siguiendo una aproximación como la del camino Hamiltoniano, grafo de De Bruijn o el camino Euleriano, se trataría la reconstrucción del genoma que se esté analizando. De este modo, el coste y la eficiencia del algoritmo en la resolución implicará una variabilidad del tiempo en cada caso.

1.4 Analiza la calidad del ensamblaje (QUAST)

Mediante la interfaz web de QUAST, se emplea el archivo correspondiente a los *scaffolds* obtenidos en el ejercicio anterior con SPAdes. Manteniendo los parámetros por defecto, se alcanza un resumen inicial similar al comentado anteriormente:

Statistics without reference	scaffolds
# contigs	6
# contigs (>= 0 bp)	7
# contigs (>= 1000 bp)	4
# contigs (>= 5000 bp)	3
# contigs (>= 10000 bp)	2
# contigs (>= 25000 bp)	2
# contigs (>= 50000 bp)	1
Largest contig	132 140
Total length	179 288
Total length (>= 0 bp)	179 607
Total length (>= 1000 bp)	177 919
Total length (>= 5000 bp)	173 211
Total length (>= 10000 bp)	167 264
Total length (>= 25000 bp)	167 264
Total length (>= 50000 bp)	132 140
N50	132 140
N75	35 124
L50	1
L75	2
GC (%)	33.59
Mismatches	
# N's	0
# N's per 100 kbp	0

Posteriormente, centramos la atención en el *N50*, lo que nos dará una idea de la calidad en el ensamblaje:



Con respecto al resultado alcanzado, apreciamos que la longitud de los *scaffolds* necesaria para la reconstrucción del 50 % del genoma es de 132.140 pb. Por tanto, tenemos una primera idea de calidad cuya rigurosa interpretación depende de otros parámetros complementarios en este tipo de procesos, no proporcionados en este caso con QUAST. No obstante, la tendencia con el *N50* es que cuanto mayor sea este, mejora calidad tiene el ensamblaje, pues la mitad del genoma se reconstruye en *contigs* notablemente más largos, por lo que hay una menor incertidumbre de *gaps* entre los *contigs*, lográndose una representación "1 *contig* – 1 cromosoma" para el genoma analizado.

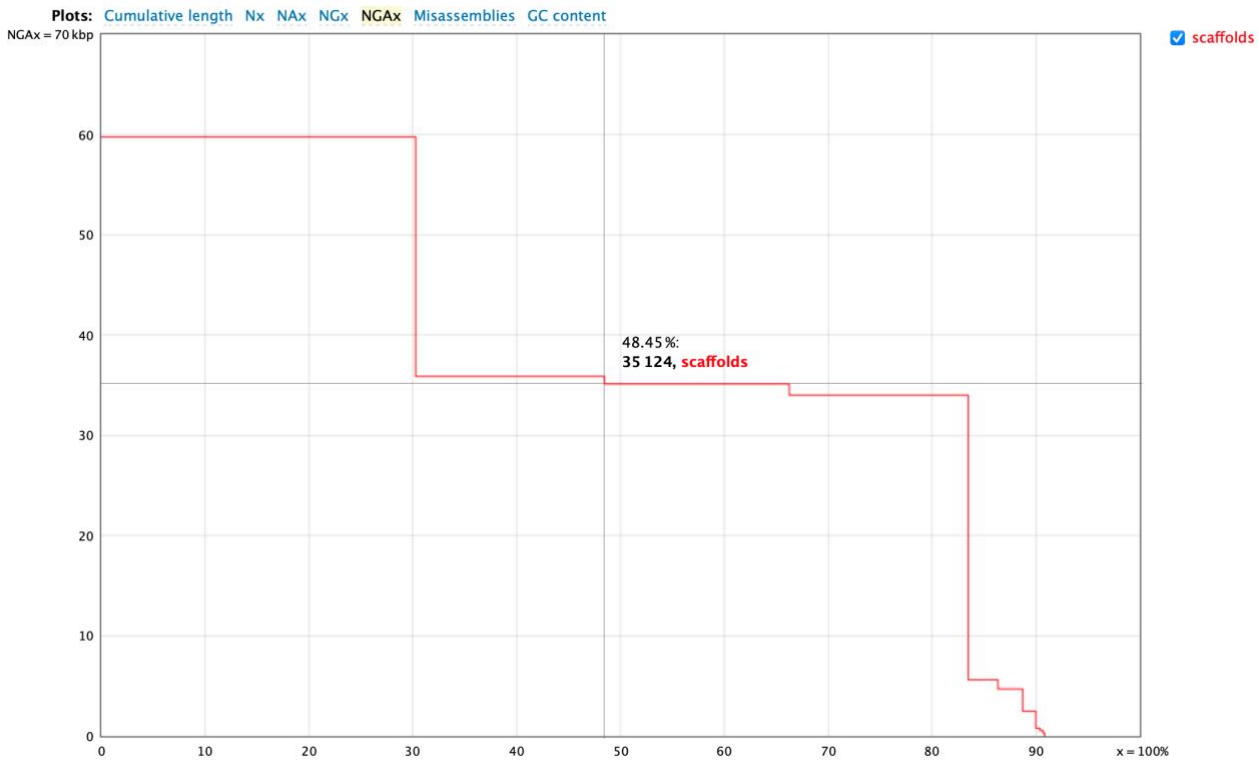
1.5 Analiza la calidad del ensamblaje utilizando el genoma de referencia (QUAST)

Mediante el uso del genoma de referencia proporcionado para la práctica, es posible alcanzar otro punto de vista con respecto a la calidad del ensamblaje. Como diferencia con respecto al apartado anterior, simplemente seleccionamos la sección *Genome*, con el fin de subir el archivo *reference.fna* correspondiente. Se ha alcanzado el siguiente resultado, con el resumen inicial y gráfico en torno al *N50*:

Genome statistics	scaffolds
Genome fraction (%)	90.751
Duplication ratio	1.001
Largest alignment	59 758
Total aligned length	179 288
NGA50	35 124
LGA50	3
Misassemblies	
# misassemblies	4
Misassembled contigs length	138 087
Mismatches	
# mismatches per 100 kbp	7.82
# indels per 100 kbp	3.91
# N's per 100 kbp	0
Statistics without reference	
# contigs	6
Largest contig	132 140
Total length	179 288
Total length (>= 1000 bp)	177 919
Total length (>= 10000 bp)	167 264
Total length (>= 50000 bp)	132 140

Extended report

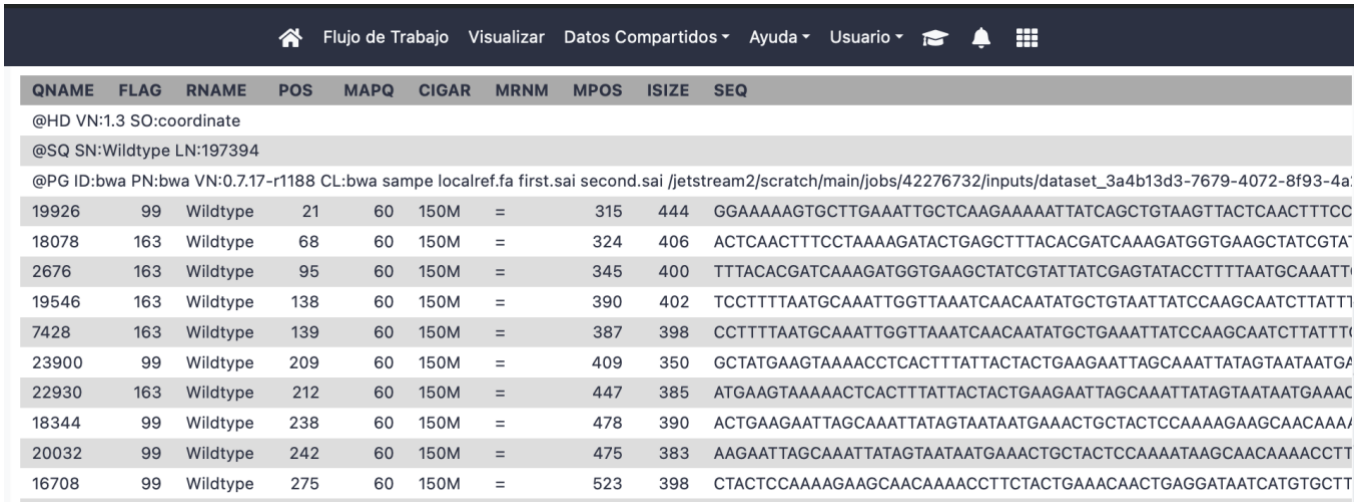
En general, se puede apreciar una nueva referencia en torno a parámetros que nos indican la calidad del ensamblaje. De esta forma, centramos la atención en el *NGA50*, que consiste en una aproximación similar al *N50* mencionado anteriormente, cuando se proporciona un genoma de referencia. Dado que considera los *mismatches*, así como los *contigs* generados, podemos concluir en este caso que la calidad del ensamblaje es adecuada a pesar de que se reduce notablemente en comparación con el valor de *N50*. Gráficamente, tendremos una clara visualización del *NGA50*, con la idea que se ha comentado:



Si se compara con el resultado gráfico del *N50*, vemos claramente que hay una misma tendencia en torno a la evaluación de la calidad del ensamblaje.

1.6 Indexa el genoma de referencia y obtén los alineamientos del genoma con las lecturas (BWA)

La resolución de este ejercicio se ha efectuado mediante línea de comandos y en Galaxy. Para este último caso, seleccionamos el archivo de referencia *reference.fna* como el genoma de referencia, el cual se indexa con BWA, obteniendo los alineamientos con las lecturas *forward* y *reverse* de los archivos *reads1.fastq* y *reads2.fastq*. En general, apreciamos la salida en un formato tabular:



QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ
@HD VN:1.3 SO:coordinate									
@SQ SN:Wildtype LN:197394									
@PG ID:bwa PN:bwa VN:0.7.17-r1188 CL:bwa sampe localref.fa first.sai second.sai /jetstream2/scratch/main/jobs/42276732/inputs/dataset_3a4b13d3-7679-4072-8f93-4a									
19926	99	Wildtype	21	60	150M	=	315	444	GGAAAAAGTGCTTGAAATTGCTCAAGAAAAATTATCAGCTGTAAGTTACTCAACTTTCC
18078	163	Wildtype	68	60	150M	=	324	406	ACTCAACTTTCCTAAAAGATACTGAGCTTTACACGATCAAAGATGGTGAAGCTATCGTA
2676	163	Wildtype	95	60	150M	=	345	400	TTTACACGATCAAAGATGGTGAAGCTATCGTATTATCGAGTATACCTTTTAATGCAAATT
19546	163	Wildtype	138	60	150M	=	390	402	TCCTTTTAATGCAAATTGGTTAAATCAACAATATGCTGTAATTATCCAAGCAATCTTATTT
7428	163	Wildtype	139	60	150M	=	387	398	CCTTTTAATGCAAATTGGTTAAATCAACAATATGCTGAAATTATCCAAGCAATCTTATTT
23900	99	Wildtype	209	60	150M	=	409	350	GCTATGAAGTAAAACCTCACTTTATTACTACTGAAGAATTAGCAAATTATAGTAATAATGA
22930	163	Wildtype	212	60	150M	=	447	385	ATGAAGTAAAAACTCACTTTATTACTACTGAAGAATTAGCAAATTATAGTAATAATGAAAC
18344	99	Wildtype	238	60	150M	=	478	390	ACTGAAGAATTAGCAAATTATAGTAATAATGAAACTGCTACTCCAAAAGAAGCAACAAA
20032	99	Wildtype	242	60	150M	=	475	383	AAGAATTAGCAAATTATAGTAATAATGAAACTGCTACTCCAAAATAAGCAACAAAACCTT
16708	99	Wildtype	275	60	150M	=	523	398	CTACTCCAAAAGAAGCAACAAAACCTTCTACTGAAACAACTGAGGATAATCATGTGCTT

En lo que se refiere a la línea de comandos, mediante la ejecución de la instrucción *bwa index reference.fna*, se procede a una primera fase en donde se indexa el genoma de referencia:

```
[curso314@fs6804 practice4_EDASB]$ bwa index reference.fna
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.03 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Pack forward-only FASTA... 0.00 sec
[bwa_index] Construct SA from BWT and Occ... 0.02 sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa index reference.fna
[main] Real time: 0.284 sec; CPU: 0.056 sec
[curso314@fs6804 practice4_EDASB]$ ls
reads1.fastq  reads2.fastq  reference.fna  reference.fna.amb  reference.fna.ann  reference.fna.bwt  reference.fna.pac  reference.fna.sa
```

Posteriormente, se lleva a cabo el alineamiento del genoma con las lecturas. Para ello, se ejecuta la instrucción *bwa mem reference.fna reads1.fastq reads2.fastq > alineamiento_prueba.sam.*

Visualizamos la cabecera del archivo *alineamiento_prueba.sam* generado (misma salida que la de Galaxy), donde la estructura que presenta pone de manifiesto el nombre de la *read* en un primer campo, el tipo (wildtype) o cromosoma en el que se ha mapeado, la posición inicial de la *read*, el valor de *mapping quality* (60 en este caso), la cadena CIGAR que nos informa del número de *matches* (150 *matches* debido al 150M que se aprecia) y finalmente, se centra la atención tanto en la secuencia como en la cadena de calidad correspondiente, recogidas en los últimos campos:

```
[curso314@fs6804 practice4_EDASB]$ head alineamiento_prueba.sam
@SQ     SN:Wildtype     LN:197394
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa mem reference.fna reads1.fastq reads2.fastq
24960   99      Wildtype        101455  60      150M    =       101710  405     AATGTTGTCACTTGGATTCAAATGACATTTTAAATCTAATTATTCATGAATCGAACTAGTACGAAATGCAATGAGCATCTTGTCTAGTTCG
ATTTTTTAATGTCTAAAAATGTCGTATATGTAATCAGAGTAGAAAGTGTTGAGGCGTTT     5??A9?BBBDDDBEDDBFF+FGHHIIHHHEIHIIHIIAHDHIIHIG#IIHIFHHHFGIII*IHHHIHFIIHGICIHHIHFFFHHHIIIIIHIHDHIIIAHHH?GHHH
HHF@HGGH6GGHEEGBGGGGGFGFE6FGFEFE?GF5EGGEEAC     NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i:0
24960   147     Wildtype        101710  60      150M    =       101455  -405    ATCAACGTATATAAGTTAAAATTGGTTTGGATAAAATGATATCTATCGTTGTGTATTGTTTGTTTTTATAGTTCGCGACGACGTCCAGCTA
ATAACGCTGCACCTAACGCTAATGATAATCCACCAAATACAGTTGTACCGATGAATGGA     -GE*G;;DGEFE-E-EG*DHDGHGGHDH;FF*HEHFFGHEF@FHH+-HHH,GIIHHHHH<H/AFIEH=?HIHHFHHEGIEH5I?CHEH?GID=IHHACCF8H9H9CI
IIIFHICIHHHHIF>HEHFHIAFEFFGD<B@DBD@BBB9????     NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i:0
24958   99      Wildtype        183317  60      150M    =       183563  396     CAAAGTCGTTGGTCATATAAAAAACCGCGTACAGTCAACTATAGATACAATCAAGATAAACTCATGCACAGATTGGGAGATATTTTAGTGC
AATATGGAATTCAACATGACACAGGGTTATTACCACATGAATGGCATTATCACATTTCG     ?A????@?DDDABDE9FGGGFGICFHIIIBGHIIIGICHHIFH=IHAFIHIHHHHIFCIIEIHAIFGIHIDDIHEIIFIIIIHHHICIIIHIH=@HFHEHBFHIHHH
H-HHHFCBHHDGHHEHGEH<GDEEEFHDEGGDGE*FGBGG,GD     NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i:0
24958   147     Wildtype        183563  60      150M    =       183317  -396    GATAAAGAACCAGTTGGTATTGATATCGAAAAGATATCACTACGTTTAGACTGGCGTACATTAGTGACGTGTTTCTCTACAAACGAAGCAC
ATCAAATATGTAGTTTAAATGATTTTTATCAAATATGGACACAAAAAGAAAGTTTTACA     2FC*DE:'GGFEG);EGGGD=GH9FFF(GFF+EHHEGDEF,HFFDDHGHIF-CFHHHIHHDH@ICH(CCH-HF6HHHH5HH.FHHH-DHIGDGFIIGEBHEIHHHAH
II+DAHIHEICIHF+EHIHII+CGGFFDDBDDDDDB?B?????     NM:i:1  MD:Z:40A109     MC:Z:150M       AS:i:145        XS:i:0
24956   99      Wildtype        37289   60      150M    =       37605   466     TATAAATTCAACTTTGCAACAGAACCATCTAATCTTCAACAAACTGGCCCGTTTGTTGAACTACTCTTTAATAAAATAATTTTTCCGTTCC
CAATTCCACATTGCAATAATAGAAAATCCATCTTCATCGGCTTTTTCGTCATCATCTGT     ?????BBADD5DDDDDGFGCFFEECFBBICIII,IIHIICIHIIFHHHHHIHIIIIIIIAHHIHHH5FHDHHHHFIIHFGF5IIIHIIHHI@<IHHHHHDHDHFHH
IGGEHHFHGEEFBHH?HH*FEGBHGH@EDG?C*FEGBGEGGGG     NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i:26
24956   147     Wildtype        37605   60      150M    =       37289   -466    ACAGGCTATTTTGCAGTTTCGTCAACTGCCGATTGTATATCCGATGTATAATTATTTTTCGGTCGAATCATTTGAACTTTTACATTTGGAT
CATAGTCTAATTTCATTGCCTTTTTCCAAAATTGAATCCATTGTTTTTGATTCACGTAG     *G*EE'FC2EEG;GEDEEF8@?1D=F7;G9DFF+H4HCHCEFFHG+HDCF+I,GC+IIHEHEHF@HCHIFHGIF-EIFDEFHH+HCDHFHHEAHCFHHIFGEFHHH-
H.HHF+HIIIHI/FI>H/GFFEGGGGGD9DBDDD-B7B9,??A     NM:i:4  MD:Z:5A19T19T4T99        MC:Z:150M       AS:i:130        XS:i:0
24954   83      Wildtype        137039  60      150M    =       136760  -429    TTTTAAATCTTTCAAGCTTTTAATTTTTGAGTCTTTTTTAACAAGAATTTCTGATTTATAACTATCTACAAGTTCTTTACTTGCTGAACCA
TCTTCTTTTACACCAAAACGTTGTGCTTGTAATAATAAATCAGCTGCTTTTTGATCATG     GEA:B,EEFGGGGCGE:;FGGHAGGFH9EFGHEBHHGGHH<HIHCHHGFHHFFHHHCHHHFHIHHHIHHCIHHGIIHFI+HIHHHHDHF>HHIHHIIIIIHHIIIHI
HI9HIFFHCFC/HFIBIIIHEFAGGGFDDDDDDDBB?<?=?A<     NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i:0
24954   163     Wildtype        136760  60      150M    =       137039  429     AGAAATTGTGTCATTCGGAATAGCTTGTGTTAATTTTAAAATTCGTGTGTCTTTAAATACATTTGGTTGGTCTTTTTTCACAGTATTACGT
GCATCGTTAAATACAGCTGCAGCATCTACATCTCCATTTAATAATGAGATAACTGCTTG     ????BA<D@DBDEBDGFGFCFIHIIII.CHHIIGHEIHIIH9HHFHEHFH/IHHHFHHGHECH5HEFFIBFHD+@HHFFI5GIHDH>HBHHGHEHGHHHH+D=DDD
HHHEHHHHE*DFHE=FHHGEEECDHCF+FGCEGGGFFEC5EEE     NM:i:0  MD:Z:150        MC:Z:150M       AS:i:150        XS:i:0
```

Debemos tener en cuenta que el manejo de ficheros SAM es frecuente en este ámbito, aunque habitualmente se emplean más los ficheros BAM. Esto se debe a que los últimos conforman una versión binaria de SAM (*tab delimited ASCII columns*), sobre la que trabajan la mayoría de los algoritmos de diversas herramientas. Es preciso destacar que en estos ficheros hay una asignación de *reads* a posiciones específicas de un genoma de referencia, por lo que hablamos de un fichero de ordenamiento de las *reads*, es decir, contiene la misma información sobre las *reads* y su mapeo que en el caso de ficheros SAM.

1.7 Convierte el fichero SAM a BAM ordenado (SAMtools)

Con el fin de generar un fichero BAM, se hace uso de la herramienta SAMtools. Para ello, en línea de comandos, mediante la instrucción *samtools view -S -b alineamiento_prueba.sam > alineamiento_prueba.bam*, se genera el nuevo archivo con la especificación solicitada en formato BAM ordenado:

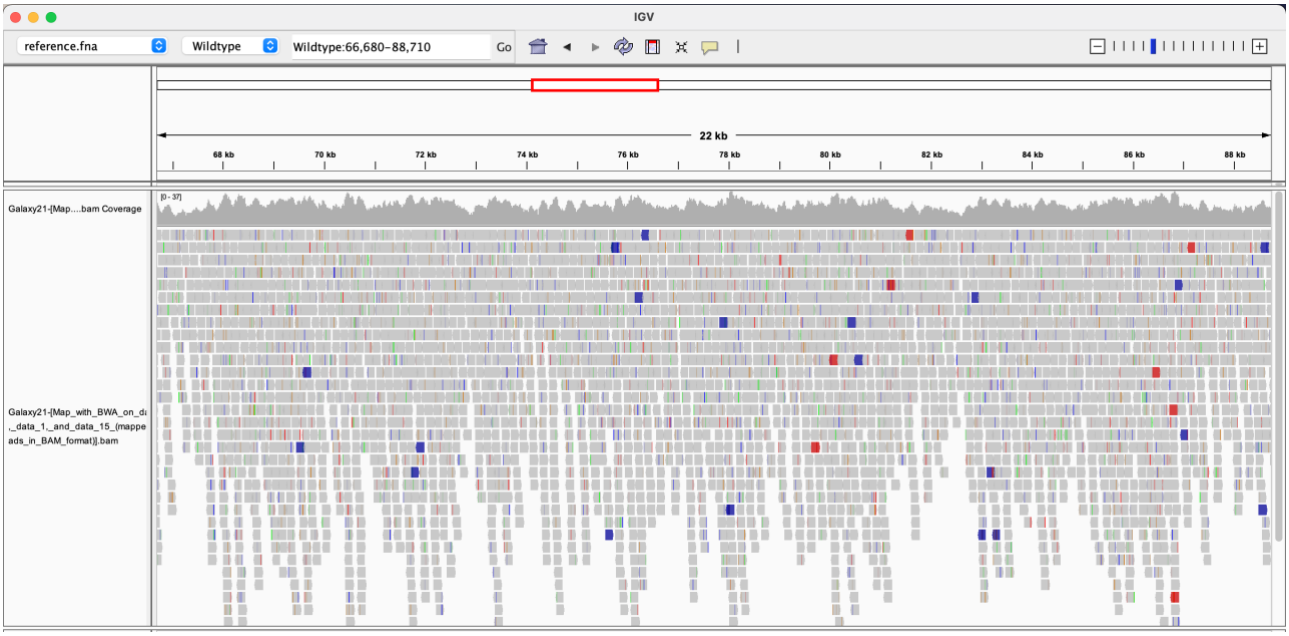
```
[curso314@fs6804 practice4_EDASB]$ ls
alineamiento_prueba.bam  reads1.fastq  reference.fna      reference.fna.ann  reference.fna.pac
alineamiento_prueba.sam  reads2.fastq  reference.fna.amb  reference.fna.bwt  reference.fna.sa
```

Se verifica que se ha generado correctamente, visualizando la cabecera, donde veremos la compleja codificación que presenta este nuevo fichero. Resulta evidente que esta versión resulta incomprensible para nosotros, pues se trata de un formato adaptado para multitud de algoritmos de diversas herramientas.

Por otro lado, es preciso destacar que para el ejercicio 1.8, finalmente se ha optado por los 2 archivos correspondientes a las salidas en Galaxy, que son del ensamblaje en formato BAM ordenado y un índice, por separado, del genoma de referencia (formato bai).

1.8 Visualiza los alineamientos ordenados de las lecturas (IGV)

En este último ejercicio, con el *Integrative Genomics Viewer* se pueden visualizar los alineamientos ordenados en formato BAM correspondientes a las *reads*. Con esto, tendremos un panorama gráfico como el siguiente, donde se muestran las *reads* enfrentadas al genoma de referencia contenido en *reference.fna*:



En este tipo de navegadores genómicos, la interfaz consiste habitualmente en una representación de la información del genoma de referencia (parte superior), cobertura (parte intermedia) y pistas con una serie de colores o símbolos que proporcionan información sobre las pistas con las que estemos trabajando, en este caso, las *reads* de partida. En proyectos de investigación relacionados con cuestiones a nivel de genes, mutaciones y ciertas regiones cromosómicas, esta visualización ayuda notablemente a su avance, así como a la postulación de modificaciones de fases o presentación de resultados.

Por otra parte, las herramientas de este tipo como el IGV muestran, tal y como se ha mencionado, la cobertura, es decir, a la representación del número de veces que una posición nucleotídica (del genoma de referencia) está presente en diferentes *reads*. Dicha cobertura se obtiene mediante el cálculo del promedio de todas las posiciones secuenciadas, pues en una misma secuenciación, la cobertura de las posiciones nucleotídicas nunca es igual. En resumen, a mayor cobertura, tendremos una mayor calidad de los datos, así como la posibilidad de discriminar si unos cambios de nucleótidos en una región conforman una determinada lesión o se trata de una variante sobre la que centrarse.