

## Práctica de ensamble de genomas a distancia usando KBase

- ¿Qué es el ensamblaje de genomas?

Alinear pequeñas secuencias o lecturas para formar “contigs”, que al alinearse entre sí crean “scaffolds”, los cuales construyen cromosomas o un genoma en conjunto.

- ¿Cómo se ensamblan los genomas (usando KBase)?

Primeramente se ingresa a **KBase**, en el apartado de **Navigator** se crea una nueva **narrativa** en donde llevaremos a cabo el ensamblaje, por lo que, en esta importamos el archivo **SRR16201156** el cual contiene secuencias de lectura generadas por Illumina para convertirlas a un objeto de datos de lectura, con el cual, usando la aplicación del apartado de **Genome Assembly** llamada **Assemble reads with SPAdes** realizará el ensamblaje.

- **Análisis del ensamblaje:**

De primera mano nos despliega el siguiente reporte de **QUAST (Quality Assessment Tool for Genome Assemblies)** que contiene métricas que indican que tan

bueno fue el ensamblaje, las principales con buenos resultados son: número de contigs, bajo; tamaño de contigs, alto; N50, grande, GC(%), normal (depende de lo ensamblado), entre otras.

Habiendo analizado las métricas se descarga un archivo **SPAdes.contigs.FASTA.zip** que contiene un archivo fasta que utilizamos posteriormente para cargarlo en NCBI y hacer un BLAST (Basic Local Alignment Search Tool) de tipo nucleótidos, que compara, en este caso nuestro ensamblado con todas las secuencias de nucleótidos de NCBI, lo que finalmente despliega una lista de las secuencias con las que se obtuvo un alineamiento significativo, en este caso el ensamblado tuvo un “porcentaje de identidad” (fracción de nucleótidos idénticos entre las dos secuencias) y un “query cover” (porcentaje del ensamblado cubierto por la secuencia encontrada de la base de datos) del 99.99% y 99% respectivamente con SARS-CoV-2 (Severe acute respiratory syndrome coronavirus 2).

Statistics without reference	SPAdes.Assembly
# contigs	1
# contigs (>= 0 bp)	1
# contigs (>= 1000 bp)	1
# contigs (>= 10000 bp)	1
# contigs (>= 100000 bp)	0
# contigs (>= 1000000 bp)	0
Largest contig	29 809
Total length	29 809
Total length (>= 0 bp)	29 809
Total length (>= 1000 bp)	29 809
Total length (>= 10000 bp)	29 809
Total length (>= 100000 bp)	0
Total length (>= 1000000 bp)	0
N50	29 809
N75	29 809
L50	1
L75	1
GC (%)	37.97
<b>Mismatches</b>	
# N's	0
# N's per 100 kbp	0
<b>Predicted genes</b>	
# predicted genes (unique)	3
# predicted genes (>= 0 bp)	3 + 0 part
# predicted genes (>= 300 bp)	2 + 0 part
# predicted genes (>= 1500 bp)	1 + 0 part
# predicted genes (>= 3000 bp)	1 + 0 part

Descriptions	Graphic Summary	Alignments	Taxonomy					
Sequences producing significant alignments								
Download Select columns Show 100								
<input type="checkbox"/> select all 0 sequences selected								
<div>GenBankGraphicsDistance tree of resultsMSA Viewer</div>								
Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input type="checkbox"/> Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/FL-CDC-STM-000719919/20...	Severe acute res...	54911	54911	99%	0.0	99.99%	29890	OL535497.1
<input type="checkbox"/> Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/FL-CDC-STM-000709988/20...	Severe acute res...	54911	54911	99%	0.0	99.99%	29890	OL533301.1

En conclusión las secuencias ensambladas fueron de SARS-CoV-2.