TECHNICAL REPORT

Bilateral Project: *Implementación de tecnologías de última generación para detección simultánea de múltiples patógenos a gran escala, aplicada a la vigilancia fitosanitaria y diagnóstico temprano de enfermedades en arándano y yuca (MIDAGRI-Peru)*

Bilateral project: *Establishing sustainable solutions to cassava diseases in mainland South-East Asia (ACIAR)*

Plant Health Initiative WP1: *Establishment of nanopore seq tools for pathogen detection (2022)*

Evaluación del sistema flongle (Oxford Nanopore) de bajo costo para secuenciación de patógenos

Ana Maria Leiva[1], Alejandra Gil[1], Ida Bartolini[2], Jose A. Olórtegui[2], Ricardo Velasquez[3], Wilmer J. Cuellar[1]*

[1]Cassava Program, Alliance of Bioversity International and CIAT. Cali, Colombia. [2]Servicio Nacional de Sanidad Agraria (SENASA), Lima, Perú. [3]Instituto Nacional de Innovación Agraria (INIA), Huaral, Perú.

*Para mas información: Wilmer Cuellar (w.cuellar@cgiar.org)

El secuenciador MinION de Oxford Nanopore Technologies (ONT) es una herramienta portátil que permite la secuenciación de AND en tiempo real a un precio bajo comparado con las opciones disponibles en laboratorios de capital reducido. Estos costos pueden ser aun mas reducidos si se secuencian mezclas de muestras, usan 'códigos de barra' para identificarlas luego. Una celda de flujo standard MinION puede producir de 10–20 Gb de data de secuencia en 48 horas; en la práctica una profundidad suficiente se puede obtener en unos pocos minutos de haber iniciado la secuenciación para obtener una secuencia consenso >99% nt idéntica a la secuencia original.

Recientemente, ONT ha puesto a disposición un Nuevo tipo de celda de flujo llamada "Flongle™" (Flow Cell Dongle FLO-FLG001), la cual es más pequeña y más económica que la celda standard (e.g. SpotON flow cell Mk I, R9.4.1). La celda, viene con adaptadores de tamaño para hacerla compatible con los equipos MinION y GridION, de manera que no se requiere adquirir equipos adicionales. Una Flongle contiene 126 canales de secuenciación en lugar del standard 512 y está diseñada para aplicaciones que requieren menos profundidad de secuenciación, por ejemplo, en el caso de genomas reducidos como aquellos de virus, bacterias y hongos. Por lo tanto, venimos implementando este sistema en laboratorios de diagnóstico, de bajo presupuesto para la identificación de patógenos emergentes y/o cuarentenarios. En este reporte presentamos como ejemplo resultados de la identificación y secuencia genomica de Cassava common mosaic virus (CsCMV), a partir de muestras de ARN total.

Muestras
Se procesaron muestras de arándano y yuca. La selección de las muestras de yuca se realizó con base en los resultados de PCR utilizando protocolos estándar de diagnóstico, con el objetivo de validar los resultados de secuenciación. Las muestras se conservaron y procesaron como se ha descrito en Jimenez et al. (2021), antes de obtener el ARN total.

Preparación de librerías
Se construyeron librerías de los sets seleccionados, utilizando 150 ng de RNA total. Para la construcción de la librería se usó el kit de secuenciación **kit SQK-DCS109** con barcodes **EXP-NBD104** para cDNA genómico de la compañía Oxford Nanopore Technology, la librería se llevó a cabo siguiendo los datos del fabricante.

Secuenciación
La secuenciación se desarrollo en la celda **R9.4** (**FLO-MIN106D**) durante 48 horas. Para mas detalles consultar el **Anexo 1**.

Pre procesamiento de lecturas y asignación taxonómica
El llamado de bases se realizó con Guppy v6 (ONT, 2022) empleando el algoritmo de alta precisión. Para el control de calidad se utilizó NanoFilt v2.8 (De Coster et al., 2018), donde se descartaron las secuencias con un valor Phred Score inferior a 10. Para filtrar las secuencias asociadas al hospedero, se alinearon las secuencias producidas a partir del tejido de yuca y arándano contra el genoma de referencia de *M. esculenta* (GCA_001659605.2) y *V. corymbosum* (GCA_014504835.1), respectivamente, usando Minimap2 v2.17 (Li, 2018); posteriormente se aislaron las lecturas no alineadas con SAMtools. Los archivos resultantes se subieron a la plataforma One Codex (Minot et al., 2015), donde se realizaron los análisis de composición correspondientes.

Ensamblaje del genoma de CsCMV aislado del set de yuca
Las lecturas de secuenciación se procesaron para hacer un ensamblaje completo del genoma de Cassava Common Mosaic Virus (CsCMV) utilizando diferentes ensambladores a partir de un genoma de referencia (KT002435.2): Minimap2 v2.17 (Li, 2018), Pilon y Medaka v1.2.0.

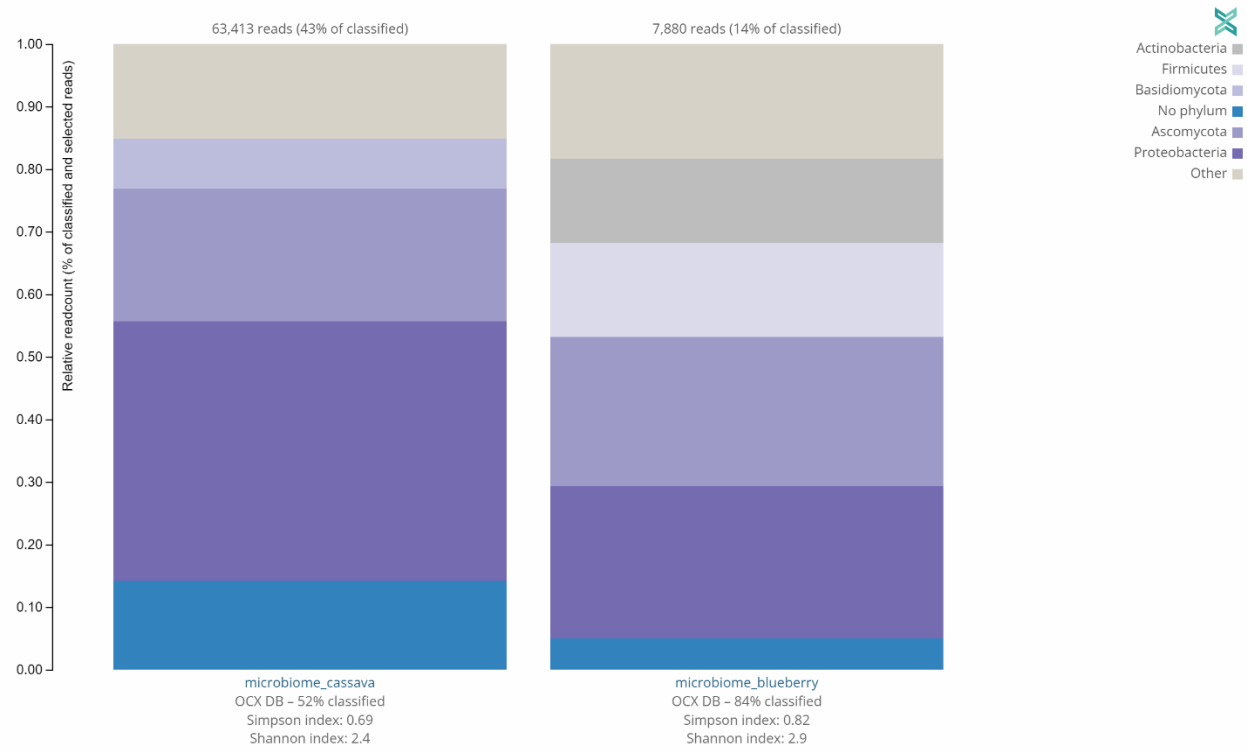
Base de datos utilizadas para comparación
Se utilizó la base de datos del *National Center for Biotechnology Information* (https://www.ncbi.nlm.nih.gov) para consultar e incluir los genomas de referencia empleados en el análisis.

Análisis filogenéticos complementares
Se realizó el análisis filogenético para la comparación de genomas de CsCMV disponibles empleando la herramienta Distance tree con el método Fast Minimum Evolution (max. Seq. difference: 0.75) del NCBI.

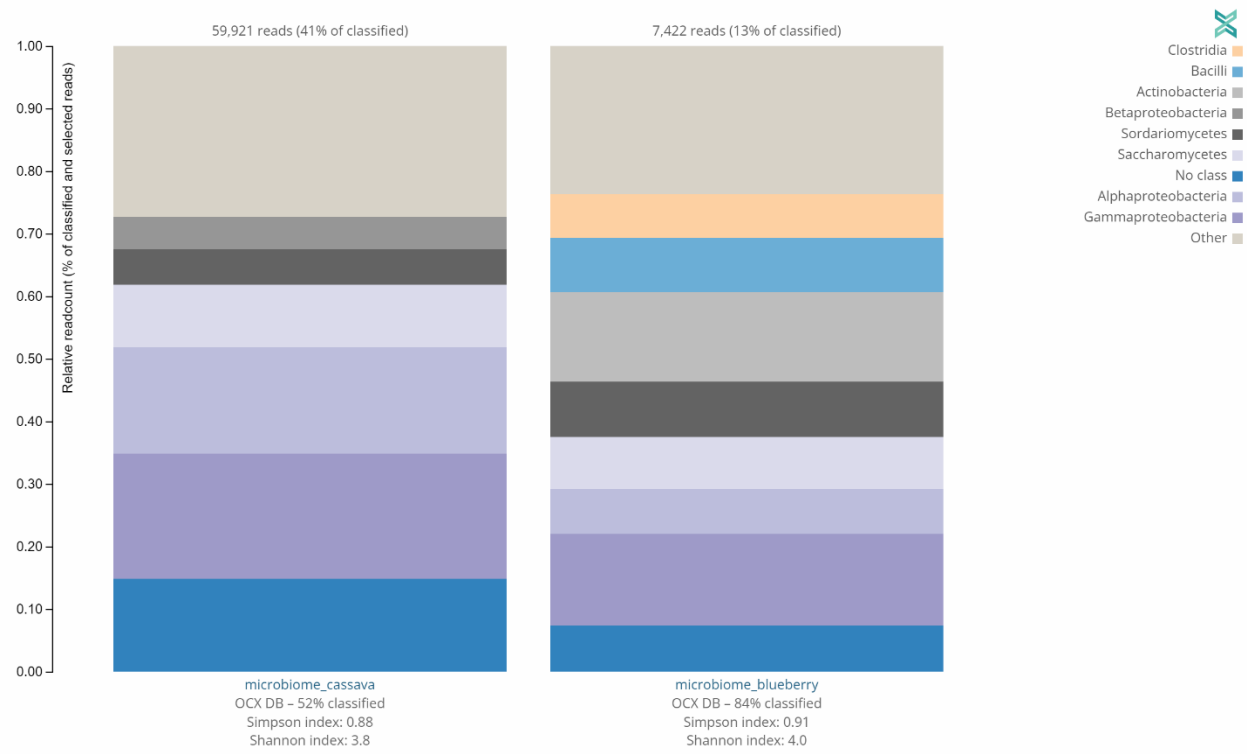
RESULTADOS

Perfil taxonómico del microbioma foliar

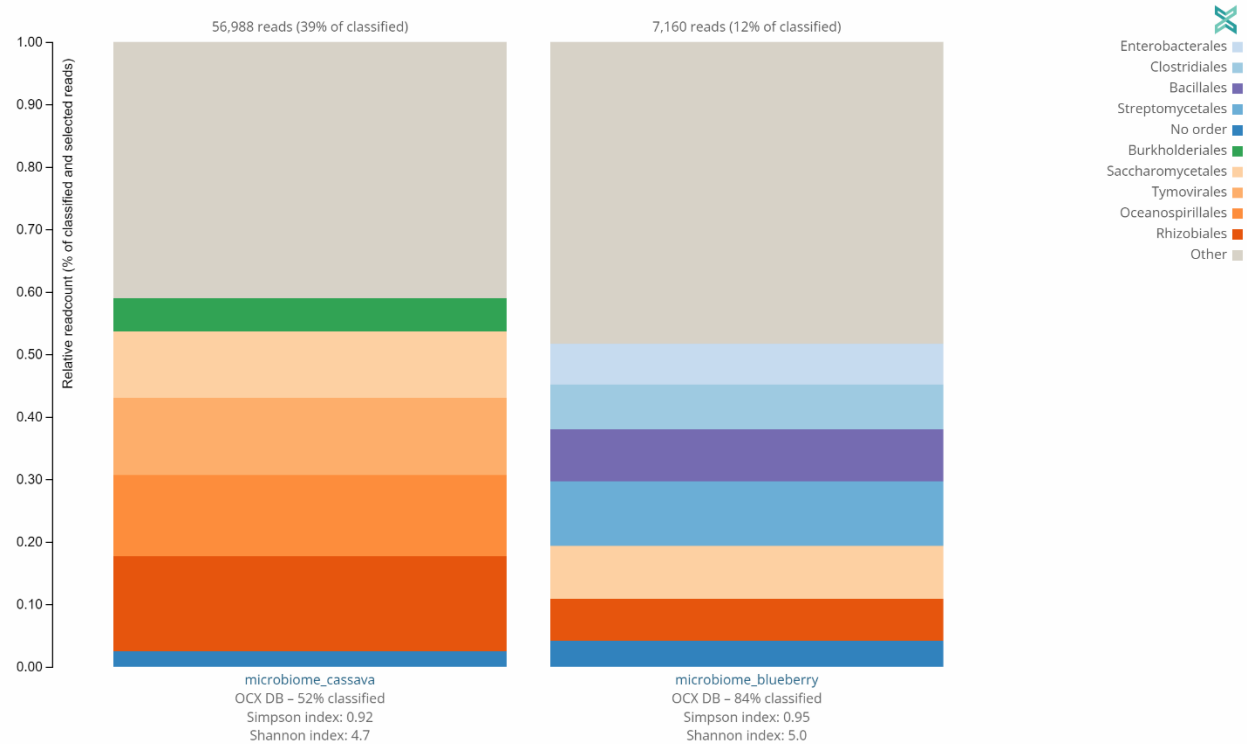
Tras procesar la información en la plataforma One Codex, se compararon las estadísticas de secuenciación y la composición taxonómica de las muestras. De la muestra de yuca y arándanos se obtuvieron 63413 y 7880 lecturas, respectivamente, de las que se clasificaron el 52 y el 84%, en cada caso. Al examinar las abundancias relativas de los diferentes rangos taxonómicos por muestra, se incluyeron en la descripción aquellos grupos representados por al menos el 5% de las lecturas identificadas. (**Figura 1**).



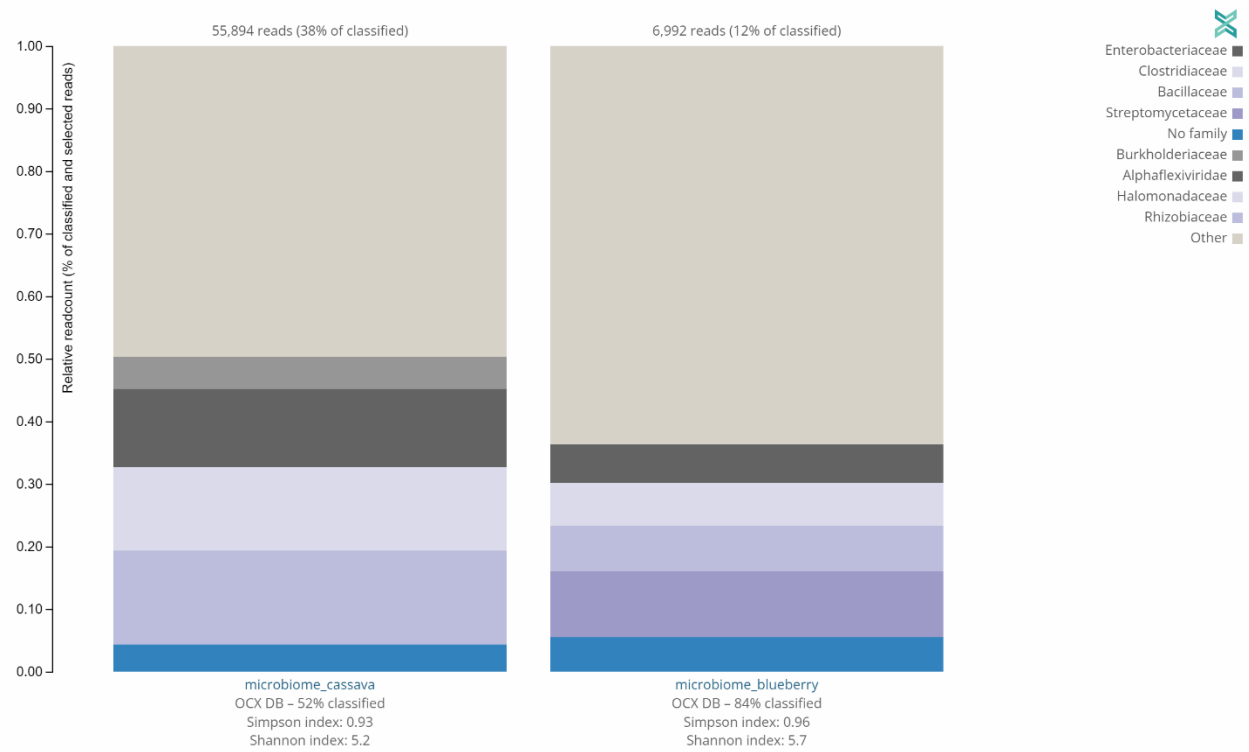
A


59,921 reads (41% of classified)
7,422 reads (13% of classified)
Relative readcount (% of classified and selected reads)
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Clostridia
Bacilli
Actinobacteria
Betaproteobacteria
Sordariomycetes
Saccharomycetes
No class
Alphaproteobacteria
Gammaproteobacteria
Other
microbiome_cassava
OCX DB – 52% classified
Simpson index: 0.88
Shannon index: 3.8
microbiome_blueberry
OCX DB – 84% classified
Simpson index: 0.91
Shannon index: 4.0

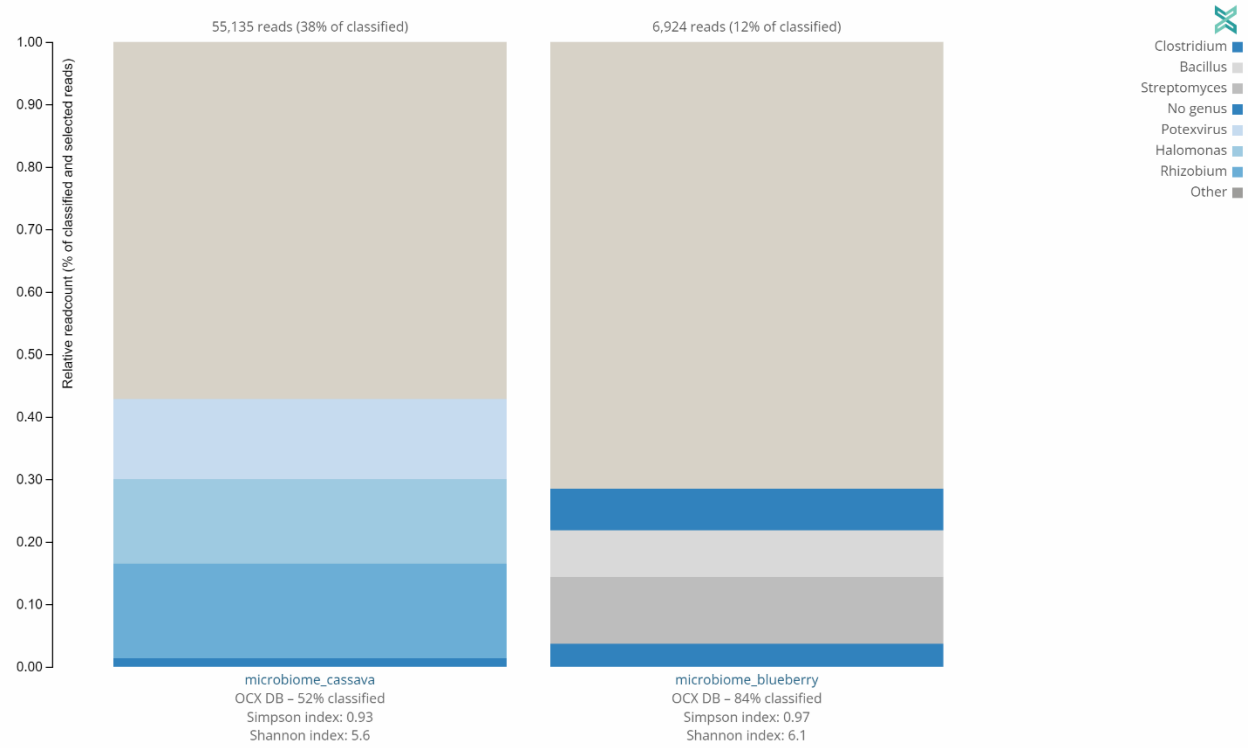
B


56,988 reads (39% of classified)
7,160 reads (12% of classified)
Relative readcount (% of classified and selected reads)
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Enterobacterales
Clostridiales
Bacillales
Streptomycetales
No order
Burkholderiales
Saccharomycetales
Tymovirales
Oceanospirillales
Rhizobiales
Other
microbiome_cassava
OCX DB – 52% classified
Simpson index: 0.92
Shannon index: 4.7
microbiome_blueberry
OCX DB – 84% classified
Simpson index: 0.95
Shannon index: 5.0

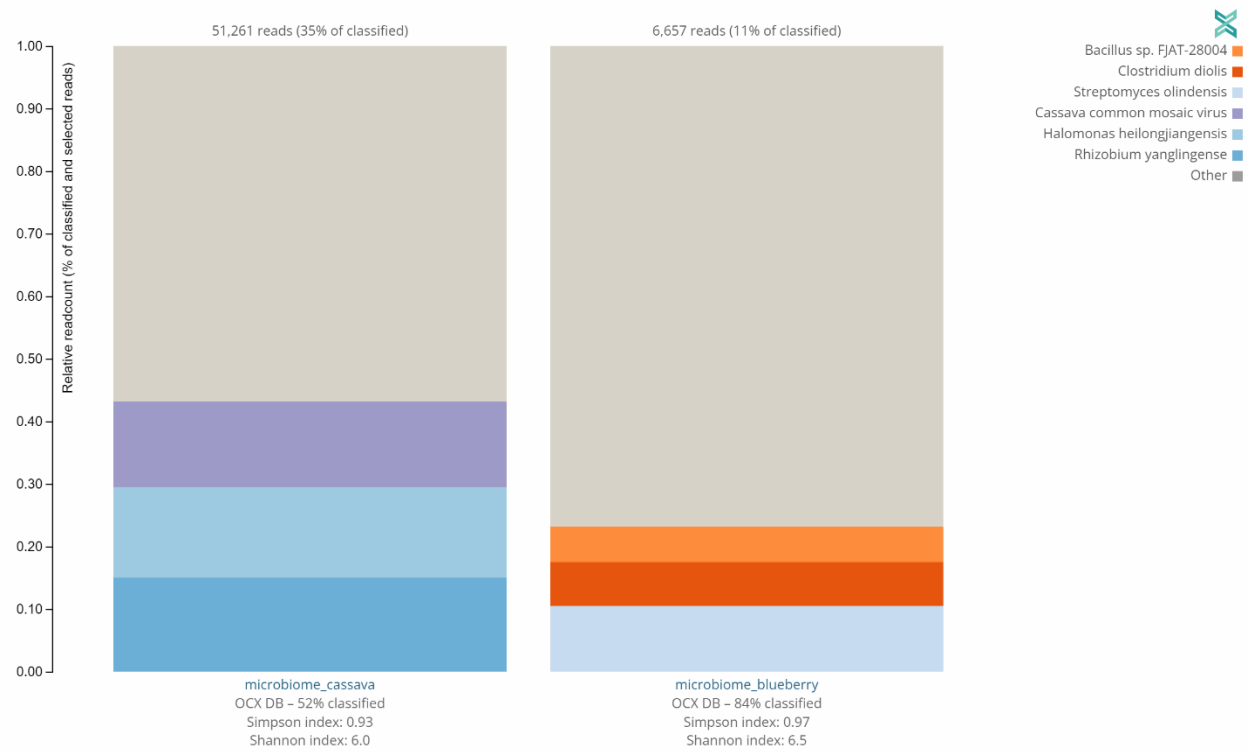
C


55,894 reads (38% of classified)
6,992 reads (12% of classified)
Relative readcount (% of classified and selected reads)
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Enterobacteriaceae
Clostridiaceae
Bacillaceae
Streptomycetaceae
No family
Burkholderiaceae
Alphaflexiviridae
Halomonadaceae
Rhizobiaceae
Other
microbiome_cassava
OCX DB – 52% classified
Simpson index: 0.93
Shannon index: 5.2
microbiome_blueberry
OCX DB – 84% classified
Simpson index: 0.96
Shannon index: 5.7

D


55,135 reads (38% of classified)
6,924 reads (12% of classified)
Relative readcount (% of classified and selected reads)
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Clostridium
Bacillus
Streptomyces
No genus
Potexvirus
Halomonas
Rhizobium
Other
microbiome_cassava
OCX DB – 52% classified
Simpson index: 0.93
Shannon index: 5.6
microbiome_blueberry
OCX DB – 84% classified
Simpson index: 0.97
Shannon index: 6.1

E



F

Figura 1. Abundancia relativa de phylum (A), clases (B), órdenes (C), familias (D), géneros (E) y especies (F) microbianas presentes en al menos 5% de las lecturas secuenciadas a partir de tejido foliar de *Manihot esculenta* (microbiome_cassava) y *Vaccinium corymbosum* (microbiome_blueberry).

En las muestras de arándano no se encontraron secuencias de virus, mientras que en las muestras de yuca se encontró el CsCMV, por lo que se procedió con su ensamblaje.

Ensamblaje de CsCMV y análisis filogenético

Las secuencias obtenidas presentaron calidad suficiente para realizar el ensamblaje del genoma. Se procesaron 3,578,998 lecturas de las cuales mapearon un total de 92,12% contra el genoma de referencia, reconstruyendo el genoma de CsCMV con una cobertura de 888,11X (Tabla 1).

Tabla 1. Estadísticas del ensamblaje del genoma *Cassava Common Mosaic* en Perú.

Summary

Globals	
Reference size	6,392
Number of reads	1,379,959
Mapped reads	12,364 / 0.9%
Unmapped reads	1,367,595 / 99.1%
Mapped paired reads	0 / 0%
Secondary alignments	4
Supplementary alignments	3,621 / 0.26%
Read min/max/mean length	0 / 26,974 / 722.2
Duplicated reads (estimated)	9,092 / 0.66%
Duplication rate	58.04%
Clipped reads	12,344 / 0.89%
ACGT Content	
Number/percentage of A's	1,690,424 / 30.62%
Number/percentage of C's	1,461,322 / 26.47%
Number/percentage of T's	1,209,636 / 21.91%
Number/percentage of G's	1,158,844 / 20.99%
Number/percentage of N's	0 / 0%
GC Percentage	47.46%
Coverage	
Mean	888.113
Standard Deviation	326.0763
Mapping Quality	
Mean Mapping Quality	51.38
Mismatches and indels	
General error rate	13.32%
Insertions	49,897
Mapped reads with at least one insertion	88.54%
Deletions	88,724
Mapped reads with at least one deletion	94.57%
Homopolymer indels	37.05%

Chromosome stats

Name	Length	Mapped bases	Mean coverage	Standard deviation
KT002435.2	6392	5676818	888.113	326.0763

En cuanto al análisis filogenético, se encontró que los genomas de CsCMV disponibles se agrupan en dos clados, donde la muestra sometida (Query_8905) se integra al clado conformado por muestras procedentes de Brasil y China.

CONCLUSIONES

Hasta el momento, no se detectaron patógenos asociados a enfermedades en plantas en la evaluación del microbioma foliar de arándano (representatividad> 5% lecturas). Por otra parte, el microbioma de yuca evidenció la presencia del virus patogénico CsCMV (3,296,997 lecturas), que permitieron el ensamblaje del genoma completo, con una profundidad de 888,11X. Como se evidencia en los registros fotográficos de las plantas (ver **Anexos**), estos resultados coinciden con la presencia de síntomas de mosaico.

El protocolo descrito se viene utilizando con una sub muestra de las colecciones de yuca y arándano de INIA, donde todo el proceso se ha validado en el laboratorio de sanidad Vegetal, de

Senasa en Lima (ver Anexo 1). El grupo de Senasa, con apoyo del equipo de Virologia de CIAT ha identificado otros microorganismos, y se está trabajando en la comunicación oficial en el caso de patógenos y plagas cuarentenarias. La secuenciación portátil se ha evaluado en otras regiones como parte de proyectos bilaterales complementarios, a partir de ARN y ADN total y también de productos de PCR. Al momento se ha confirmado fusarium TR4 en Perú, *Bemisia tabaci* Asia_II_1 y Asia_II_6 en el Laos (Leiva et al a, b y c) y recientemente hemos incluido virus de las familias Closteroviridae y Secoviridae en Africa y Las Americas y bacteria causante del marchitamiento de la yuca *Xanthomonas phaseoli* pv manihotis en Colombia (*no publicado*).

REFERENCIAS

1. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M., Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. Bioinformatics, 34(15), 2666-2669.
2. Jimenez, J., Leiva, A.M., Olaya C., Acosta-Trujillo, D., Cuellar, W.J. (2020) An optimized nucleic acid isolation protocol for virus diagnostics in cassava (*Manihot esculenta* Crantz.). MethodsX 8, 101496.
3. Leiva, A.M., Chittarath, K., Lopez-Alvarez, D., Vongphachanh, P., Gomez, M.I., Sengsay, S., Wang, X-W., Rodriguez, R., Newby, J., Cuellar, W.J. (2022a). Mitochondrial genetic diversity of *Bemisia tabaci* (Gennadius) (Hemiptera: Aleyrodidae) associated with cassava in Lao PDR. Insects. https://doi.org/10.3390/insects13100861
4. Leiva, A.M., Jimenez, J., Sandoval, H., Perez, S., Cuellar, W.J. (2022b). Complete genome sequence of a novel secovirid infecting cassava in the Americas. Archives of Virology. https://doi.org/10.1007/s00705-021-05325-2
5. Leiva, A.M., Rouard, M., Lopez-Alvarez, D., Cenci, D., Breton, C., Acuña, R., Rojas, J.C., Dita, M., Cuellar, W.J. (2022a). Draft Genome sequence of *Fusarium oxysporum* f. sp. cubense tropical race 4 from Peru, Obtained by Nanopore and Illumina Hybrid Assembly. Microbiology Resource Announcements. https://doi.org/10.1128/mra.00347-22
6. Li, H. (2018). minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094 –3100. https://doi.org/10.1093/bioinformatics/
7. Minot, S.S., Krumm, N., Greenfield, N.B. (2015). One codex: a sensitive and accurate data platform for genomic microbial identification. BioRxiv, 027607.
8. Oxford Nanopore Technologies. Analysis Solutions for Nanopore Sequencing Data. Available online: https://nanoporetech.com/support/nanopore-sequencing-data-analysis (accessed on 19 May 2022).
9. Vaser, R., Sovic´, I., Nagarajan, N., Šikic´, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27: 737–746. https://doi.org/10.1101/gr.214270.116.
10. Warmington, R.J., Kay, W., Jeffries, A., O'Neill, P., Farbos, A., Moore, K., et al. (2019). High-Quality Draft Genome Sequence of the Causal Agent of the Current Panama Disease Epidemic. Microbiol Resour Announc. 8:e00904-19.

ANEXO 1

Guía bioinformática disponible en Google Co-Laboratory

Laboratorio de Virologia - Centro Internacional de Agricultura Tropical (CIAT), Colombia

Resumen

Esta guía se preparó como parte del curso de capacitación en análisis bioinformático para la identificación de patógenos a partir de datos metagenómicos. El curso se organizó en CIAT (Cali) y en Senasa (Lima) en 2022 y comprendía tanto la parte de laboratorio (preparación de librerías) como el análisis bioinformático. Esta guía es abierta y se enfoca en el análisis de las secuencias obtenidas de cada librería y requiere la ejecución de 5 pasos, descritos brevemente a continuación:

Para poder instalar los programas bioinformáticos y correr nuestros datos, se necesitan dos pasos obligatorios: instalar conda (1) y conectar nuestro Drive al Google Colaboratory (2). Posteriormente, el procesamiento de lecturas incluye 3 fases: control de calidad (3), extracción de lecturas no eucariotas por mapeo (4), y asignación taxonómica (5).

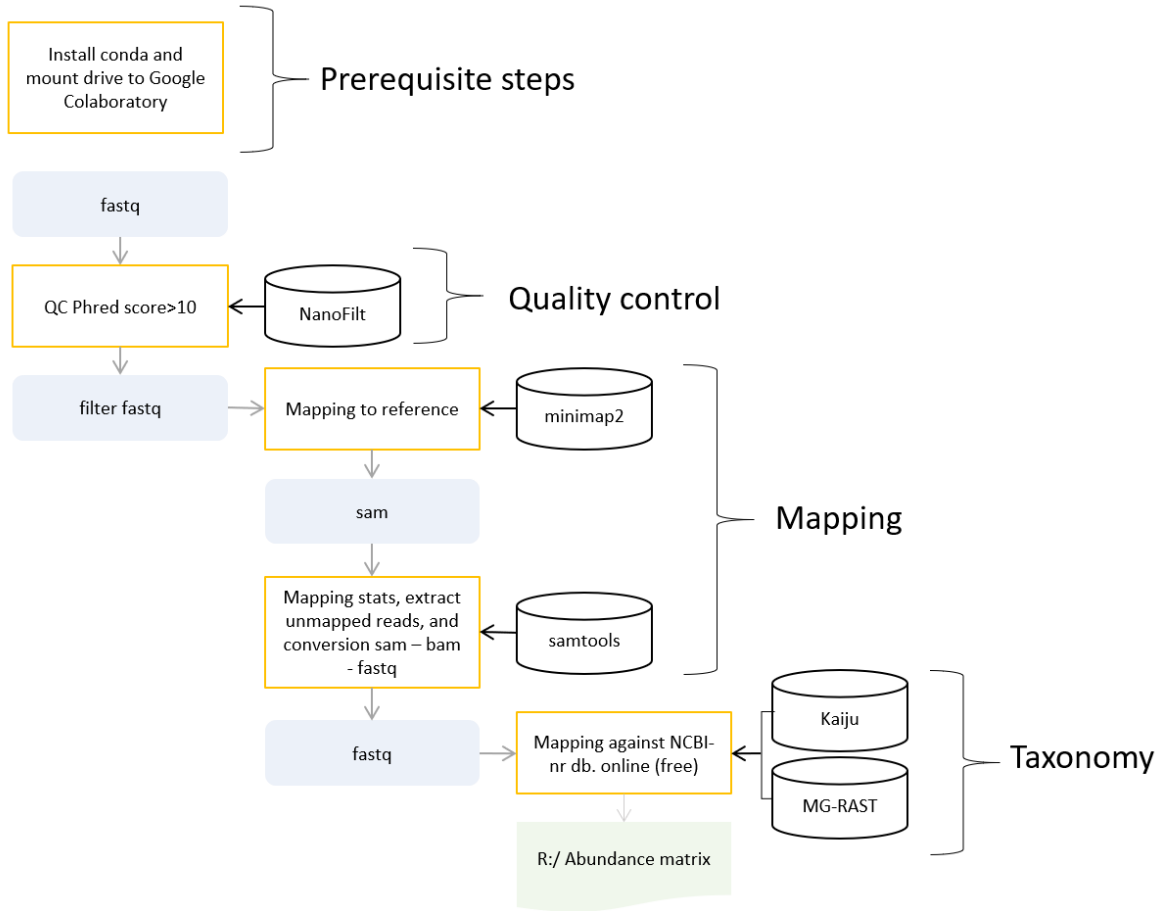


Figura 1. Guía esquemática del procesamiento informático propuesto.

Cómo correr un comando en Google Colaboratory (colab)

1. Instalar conda

```
%%bash
MINICONDA_INSTALLER_SCRIPT=Miniconda3-py37_4.9.2-Linux-x86_64.sh
MINICONDA_PREFIX=/usr/local
wget https://repo.anaconda.com/miniconda/$MINICONDA_INSTALLER_SCRIPT
chmod +x $MINICONDA_INSTALLER_SCRIPT
./$MINICONDA_INSTALLER_SCRIPT -b -f -p $MINICONDA_PREFIX
```

2. Conectar a Drive

Al montar su Google Drive, podrá cargar archivos fast5 que se pueden procesar y la salida se puede volver a escribir en la misma ubicación dentro de Drive.
El siguiente fragmento realiza el montaje. Se le pedirá que se autentique, solo siga las instrucciones y todo debería ir bastante bien.

```
from google.colab import drive
drive.mount('/content/gdrive', force_remount=True)
```

2.1. Movernos a la carpeta de trabajo

Con el comando cd nos movemos entre carpetas. cd gdrive/MyDrive/[CARPETA/DE/TRABAJO].
Ejemplo: `cd gdrive/MyDrive/CIAT/cassava`

```
cd gdrive/MyDrive/CIAT/pcr
```

3. Control de calidad: NanoFilt

NanoFilt permite realizar el filtrado por calidad y/o longitud de lectura.

IMPORTANTE: El criterio para realizar dicho cribado debe contemplar que, a diferencia de la calidad encontrada en lecturas cortas provenientes de tecnologías de secuenciación tipo Illumina (Phred Score ~ 30), las lecturas largas producidas por tecnologías de secuenciación de "tercera generación" como ONT o PacBio presentan una baja calidad general (Phred Score ~ 10). Por lo tanto, el umbral mínimo de calidad de lecturas largas suele ajustarse a un Phred Score de 10-12.

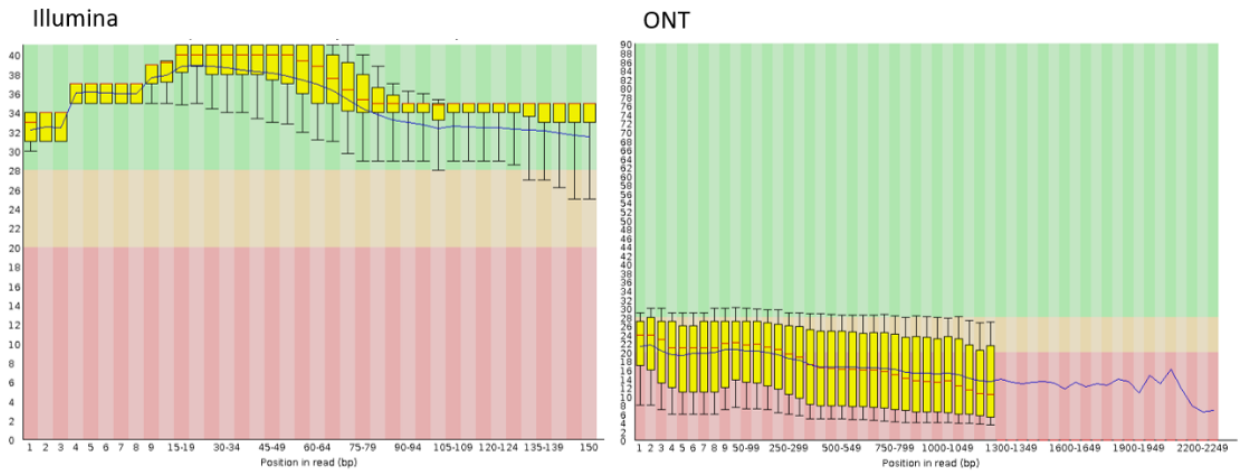


Figura 2. Comparación de Per base sequence quality (calidad por base en la secuencia) entre lecturas producidas por Illumina y Oxford Nanopore Technologies (ONT).

3.1. Instalar NanoFilt

```
!conda install -c bioconda nanofilt -y
```

3.2. Comandos intermedios

Con el comando `ls` listamos los archivos en la carpeta de trabajo.

```
!ls
```

IMPORTANTE: Si se produjo más de un archivo fastq durante la corrida, debe concatenar el archivo con el comando cat (facilita el análisis en adelante). De lo contrario, omita este paso.

```
!cat *fastq > cat.fastq.gz
```

Con el comando `gunzip` descomprimimos archivos que finalicen en `gz`.
Se debe reemplazar el nombre del archivo a descomprimir, por ejemplo: `FAR83207_pass_barcode01_9778d2ef_0.fastq.gz`

```
!gunzip FAR83207_pass_barcode01_9778d2ef_0.fastq.gz
```

Con el comando `mv` podemos cambiar de nombre del archivo. Para facilitar el análisis en adelante, cambiamos el nombre del archivo: ya sea el concatenado (cat.fastq.gz) o el nombre asignado por el MinION (ej: FAR83207_pass_barcode01_9778d2ef_0.fastq.gz) a pass.fastq.gz.

```
!mv fastq_runid_366cd4f875c6d45ac62ed9e0238ef0d345423e26_0_0.fastq pass.fastq
```

Ejecutamos NanoFilt, donde filtramos las secuencias con calidad Phred Score inferior a 10, y las almacenamos en nuevo archivo llamado `filterpass.fastq`.

```
!NanoFilt -q 10 pass.fastq > filterpass.fastq
```

4. Mapeo

4.1. Instalando programas mapeo

```
!conda install -c bioconda samtools conda=4.9.2 -y
```

```
!conda install -c bioconda minimap2 conda=4.9.2 -y
```

4.2. Minimap2

Minimap2 es un programa de alineación de propósito general para mapear el ADN o secuencias largas de ARNm contra una gran base de datos de referencia. Es muy bueno con genomas de referencia y secuencias producidas por ONT.

Cómo descargar un genoma de referencia del NCBI

Reemplazar link de su especie de interés. Sugerencia: NCBI (https://www.ncbi.nlm.nih.gov/genome). Ejemplo para descargar genoma de referencia de Manihot esculenta:

```
wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/659/605/GCF_001659605.2_M.esculenta_v8/GCF_001659605.2_M.esculenta_v8_genomic.fna.gz
```

```
!wget https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/659/605/GCF_001659605.2_M.esculenta_v8/GCF_001659605.2_M.esculenta_v8_genomic.fna.gz
```

Para facilitar el análisis en adelante, renombramos el genoma de referencia (mv), lo descomprimimos (gunzip) y pedimos que nos muestre el encabezado para corroborar la integridad del archivo (head).

```
!mv GCF_001659605.2_M.esculenta_v8_genomic.fna.gz referencia.fna.gz
```

```
!gunzip referencia.fna.gz
```

```
!head referencia.fna
```

Si se trata de un fragmento de PCR, descargar fasta manualmente y subir a directorio de trabajo.

```
!ls
```

```
!mv CMLV.fasta referencia.fna
```

```
!head referencia.fna
```

Este paso puede tomar algún tiempo

```
!minimap2 -ax map-ont referencia.fna filterpass.fastq > aln.sam
```

4.3. SAMtools

Formato SAM / BAM

El archivo SAM es un archivo de texto delimitado por tabulaciones que contiene información para cada lectura individual y su alineación con el genoma.

La versión binaria comprimida de SAM se denomina archivo BAM. Usamos esta versión para reducir el tamaño y permitir la indexación, lo que permite un acceso aleatorio eficiente de los datos contenidos en el archivo.

El archivo comienza con un encabezado, que es opcional. El encabezado se usa para describir la fuente de datos, secuencia de referencia, método de alineación, etc., esto cambiará dependiendo del alineador que se use. Después del encabezado está la sección de alineación. Cada línea que sigue corresponde a la información de alineación para una sola lectura. Cada línea de alineación tiene 11 campos obligatorios para información de mapeo esencial y un número variable de otros campos para información específica del alineador. A continuación, se muestra una entrada de ejemplo de un archivo SAM con los diferentes campos resaltados.

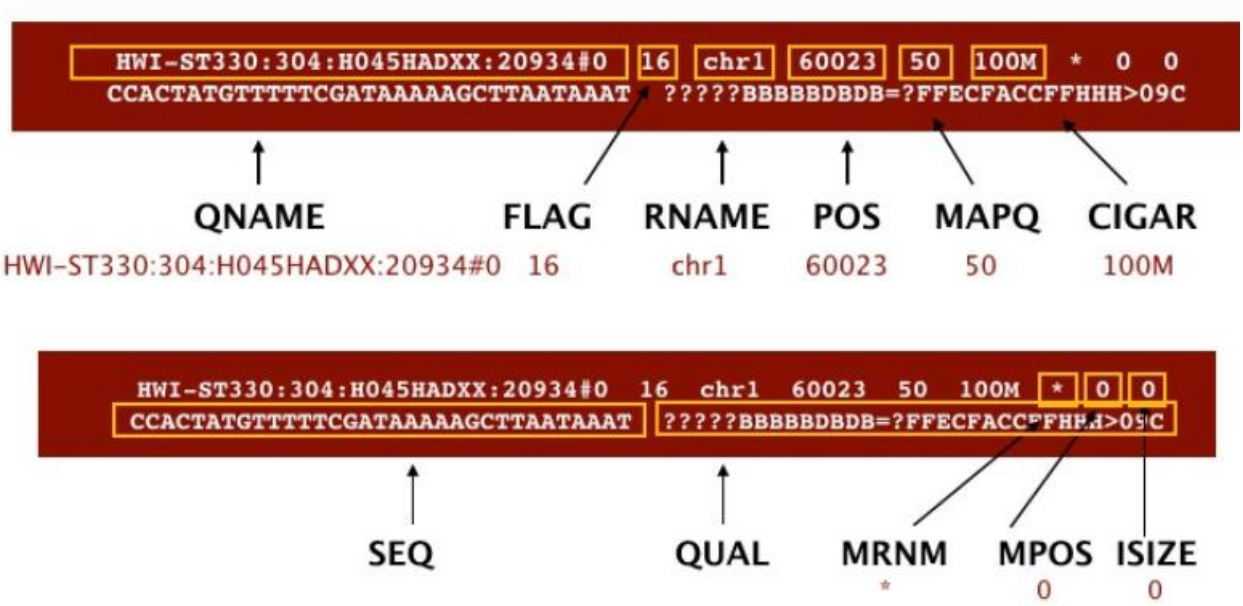


Figura 3. Formato SAM

Convertiremos el archivo SAM a formato BAM usando el programa samtools usando el argumento view y le diremos a este comando que la entrada está en formato SAM (-S) y que salga en formato BAM (-b).

```
%%bash
samtools view -S -b aln.sam > aln.bam
samtools sort -o aln_sort.bam aln.bam
samtools index aln_sort.bam
```

IMPORTANTE: El siguiente comando genera las estadísticas de alineamiento, incluido el número de lecturas mapeadas contra el genoma de referencia. Al trabajar con microorganismos secuenciados a partir del tejido del hospedero, dichas lecturas suelen constituir un porcentaje importante del total de secuencias procesadas (>90%).

```
!samtools flagstat aln_sort.bam
```

Con base en la información contenida en las etiquetas del archivo SAM / BAM, podemos guardar lecturas que compartan cierta característica. En este caso, nos interesa conservar las lecturas que no se hayan alineado contra el genoma de referencia (posiblemente pertenecientes al microbioma del hospedero). Con este paso, se genera un archivo más pequeño y depurado, agilizando así los análisis posteriores.

IMPORTANTE: Para conservar las lecturas mapeadas (mp) en lugar de las no mapeadas (unmp), sustituir -f por -F en la primera linea del comando.

```
%%bash
samtools view -b -f 4 aln_sort.bam > aln_unmp.bam
samtools sort -o sort_unmp.bam aln_unmp.bam
samtools index sort_unmp.bam
```

Código alterno

```
%%bash
samtools view -b -F 4 aln_sort.bam > aln_mp.bam
samtools sort -o sort_mp.bam aln_mp.bam
samtools index sort_mp.bam
```

Podemos verificar el cambio de tamaño ejecutando este comando de linux que "lista" los archivos de nuestra ruta de trabajo que terminan en bam.

```
!ls -lah *bam
```

Finalmente, convertimos el archivo bam a fastq, formato general para múltiples programas de asiganción taxonómica. Debido a que el ONT produce lecturas single-end, se utiliza el argumento -0, que indica la dirección de lectura. En lecturas paired-end, se indica -1 para forward y -2 para reverse.

```
!samtools fastq -0 unmp.fastq sort_unmp.bam
```

Código alterno

```
!samtools fastq -0 mp.fastq sort_mp.bam
```

5. Asignación taxonómica

5.1. Kaiju

Pros

- No necesita crear una cuenta
- Maneja bien datos crudos (sin control de calidad)

Contra

- No almacena resultados más de 3 meses
- Solo produce información taxonómica

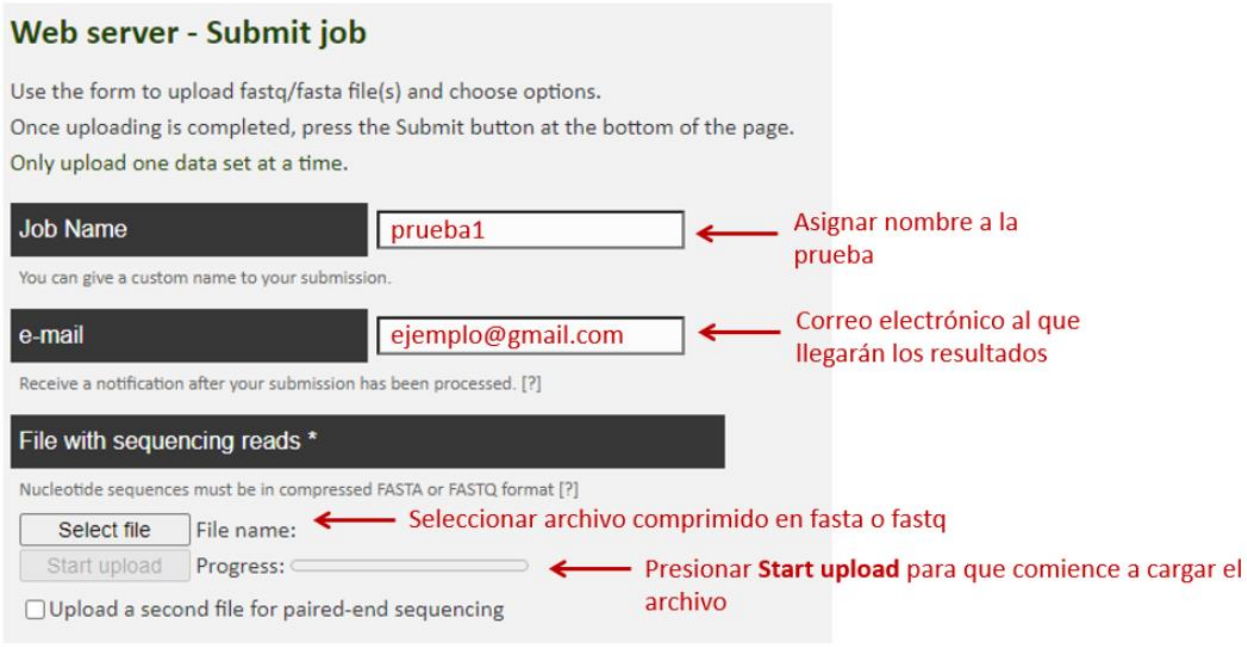
Kaiju realiza la clasificación taxonómica de microorganismos en 3 pasos:
Realiza la conversión de nucleótidos a aminoácidos en sus 6 posibles marcos de lectura, que luego divide en fragmentos por sus codones de terminación.
Los fragmentos son ordenados por su longitud de mayor a menor (MEM) y por su puntuación BLOSUM62 (Greedy).
Inspección en la base de datos mediante una búsqueda hacia atrás modificada en una implementación eficiente en memoria de la transformación de Burrows-Wheeler, que resinge el resultado a coincidencias exactas máximas (MEM) o extiende la búsqueda al permitir sustituciones (Greedy).

Subir datos

En primer lugar, debemos dirigirnos a la página de KAIJU en este link (https://kaiju.binf.ku.dk/), y seleccionamos Web Server. Una vez ahí, diligenciamos la siguiente información.
IMPORTANTE: Para subir archivos a Kaiju deben estar comprimidos.

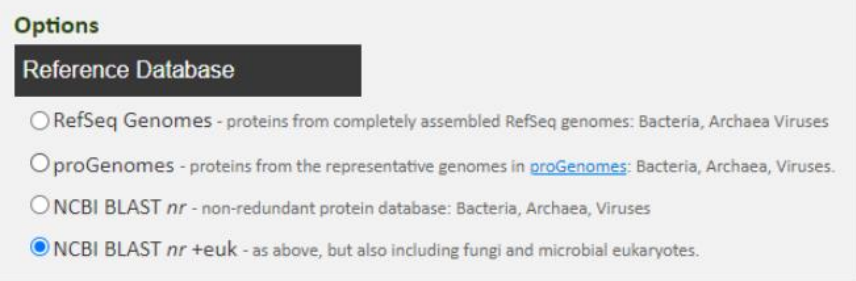
```
!gzip unmp.fastq
```

```
!gzip mp.fastq
```



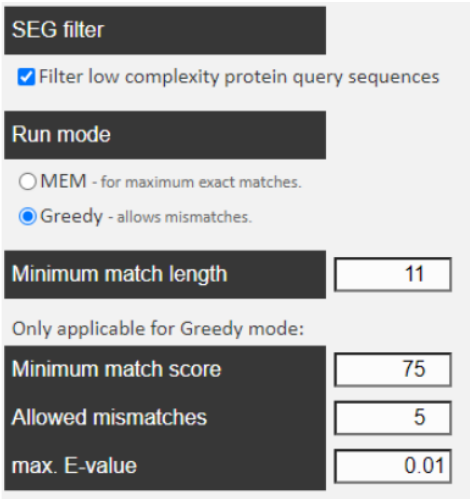
Bases de datos

Elegimos la base de datos de preferencia. Kaiju puede usar el conjunto de genomas completos disponibles de NCBI RefSeq o el subconjunto microbiano de la base de datos de proteínas no redundantes NCBI BLAST nr, que también incluye opcionalmente hongos y eucariotas microbianos.



Otras opciones

Kaiju se puede ejecutar en el modo MEM más rápido (con una longitud mínima de fragmento m = 11), así como en el modo heurístico Greedy (con una puntuación mínima s = 75), permitiendo hasta cinco (Greedy -5) sustituciones de aminoácidos durante la búsqueda.
Mantener valores por defecto.



Resultados

El tiempo de espera es variable e independiente del tamaño del archivo (las solicitudes se van gestionando en el orden en que son sometidas a la plataforma).

Una vez recibe el correo de notificación de Kaiju, el link nos redirige a una primera interfaz de resultados.

Otra forma de visualizar los resultados es mediante el Krona Chart, que le permite explorar las asignaciones por rango taxonómico. Para acceder al krona, damos click en View classification as Krona chart en la interfaz de resultados inicial.

5.2. MG-RAST

Pros

- Conserva los resultados en su cuenta
- Realiza control de calidad, asignación taxonómica y anotación funcional

- Puede emplear la plataforma como repositorio de sus datos crudos. Para esto debe cambiar el estatus de sus muestras de Privado a Público, y compartir la accesión

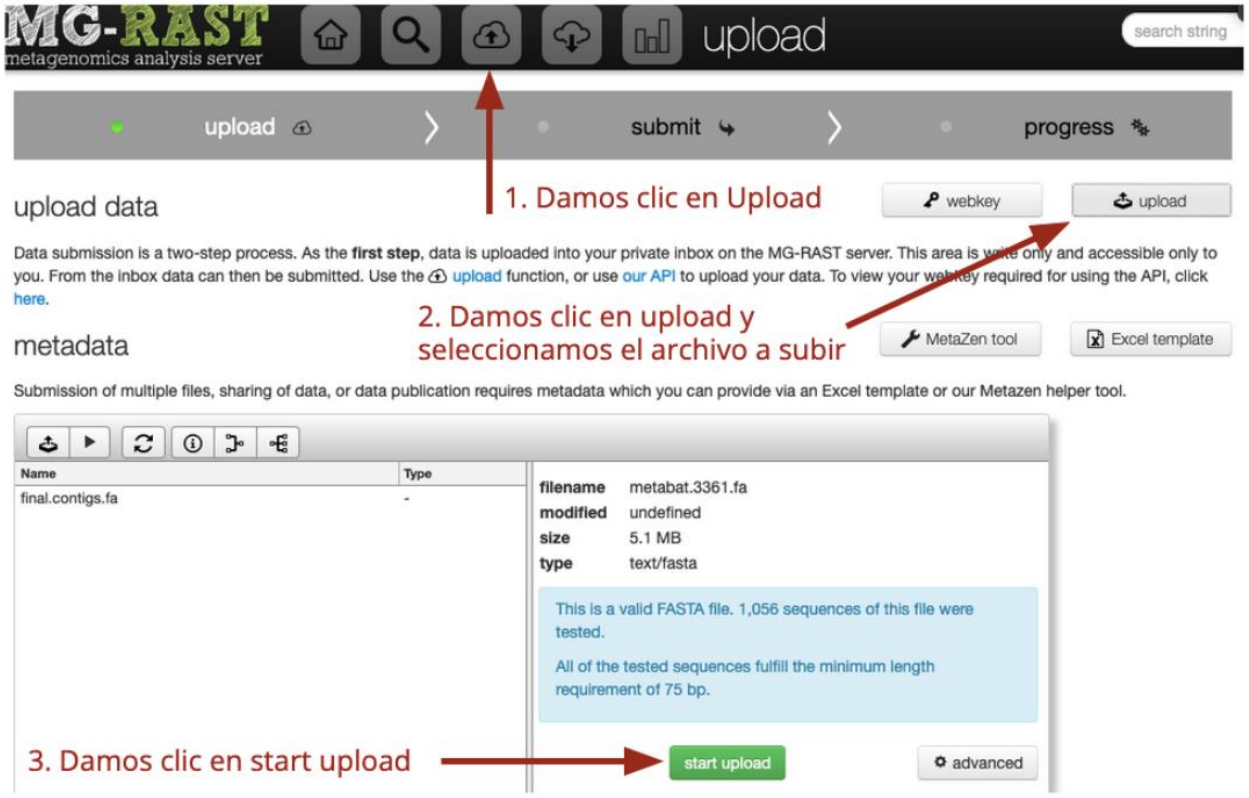
Contra

- Debe realizar una solicitud para abrir una cuenta
- No recibe algunos archivos crudos

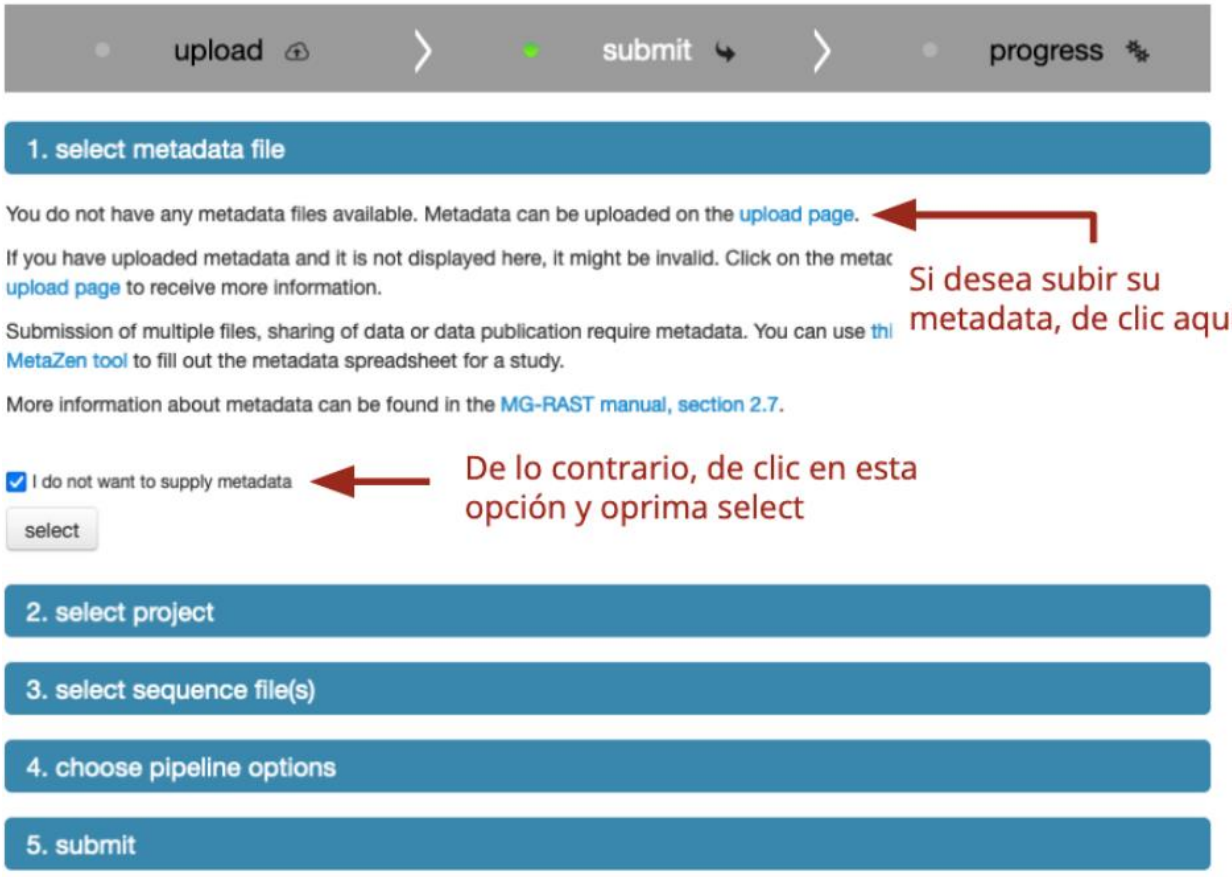
La tubería MG-RAST realiza control de calidad, predicción de proteínas, agrupamiento y anotación basada en similitudes en conjuntos de datos de secuencias de ácidos nucleicos.
Los datos en MG-RAST son privados para el usuario que los envía, a menos que los comparta con otros usuarios o los haga públicos.
MG-RAST presenta las anotaciones a través de las herramientas en la página de análisis que preparan, comparan, muestran y exportan los resultados en el sitio web. La página de descarga ofrece los datos de entrada, datos en etapas intermedias de filtrado, la salida de búsqueda de similitud y tablas resumen de funciones y organismos detectados.

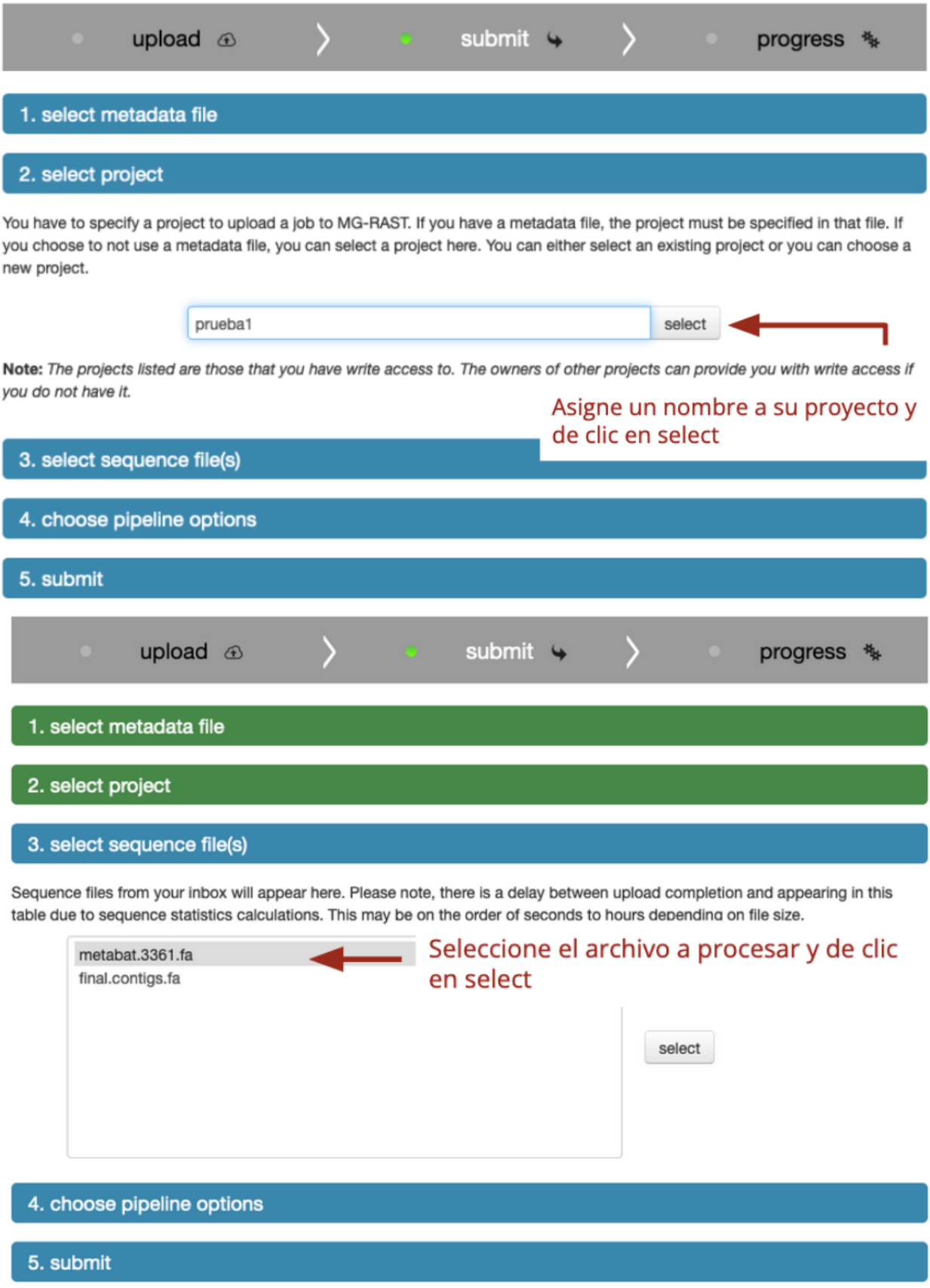
Subir datos

Una vez hayamos abierto la cuenta e ingresado a la página de MG-RAST (https://www.mg-rast.org/), seguimos los siguientes pasos para subir el/los archivo(s) a explorar.
Una vez cargado el archivo, damos click en el botón verde next que aparecerá en la esquina inferior derecha.

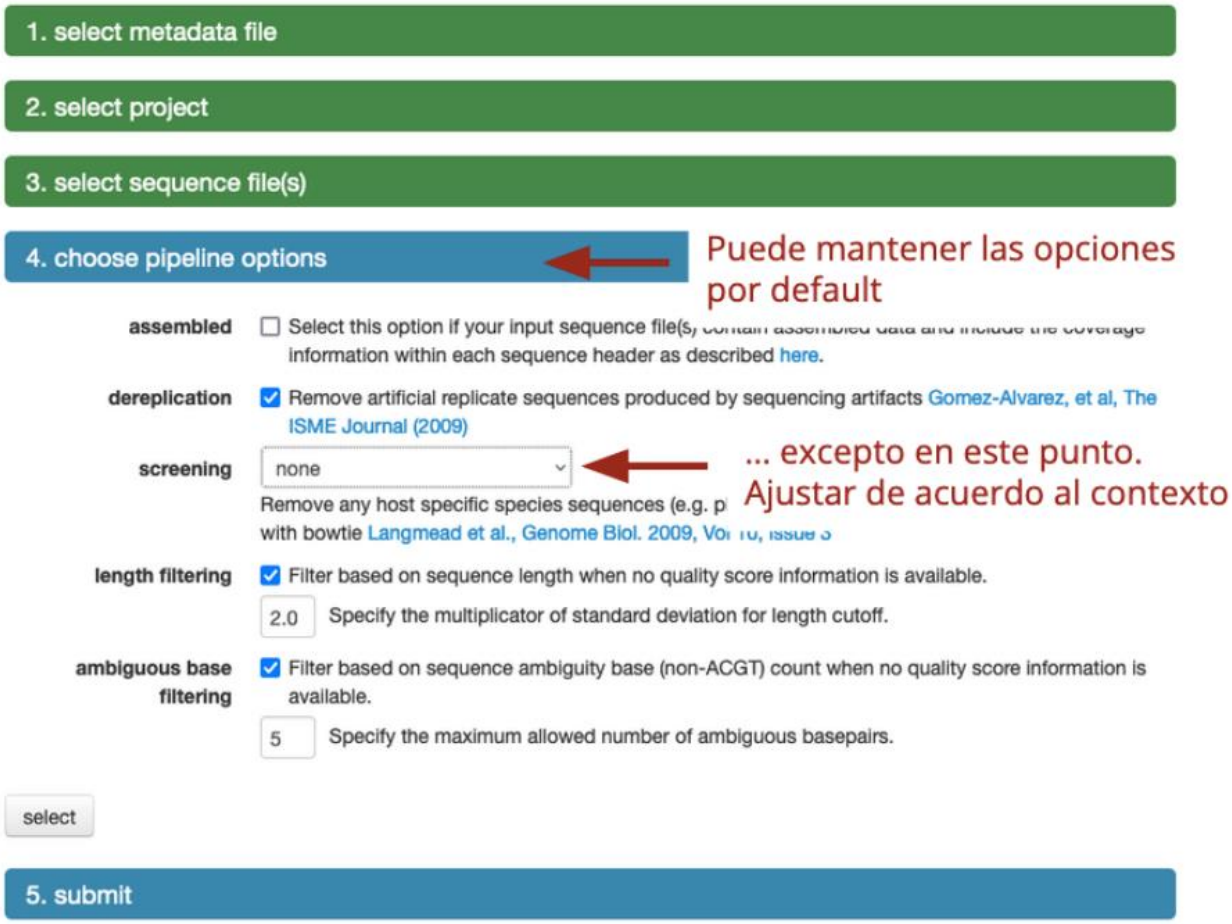


Esto llevará a una nueva ventana, en la que se deben diligenciar los siguientes campos:


upload
submit
progress
1. select metadata file
You do not have any metadata files available. Metadata can be uploaded on the upload page.
If you have uploaded metadata and it is not displayed here, it might be invalid. Click on the metac
upload page to receive more information.
Submission of multiple files, sharing of data or data publication require metadata. You can use thi
MetaZen tool to fill out the metadata spreadsheet for a study.
More information about metadata can be found in the MG-RAST manual, section 2.7.
Si desea subir su metadata, de clic aquí
I do not want to supply metadata
select
De lo contrario, de clic en esta opción y oprima select
2. select project
3. select sequence file(s)
4. choose pipeline options
5. submit

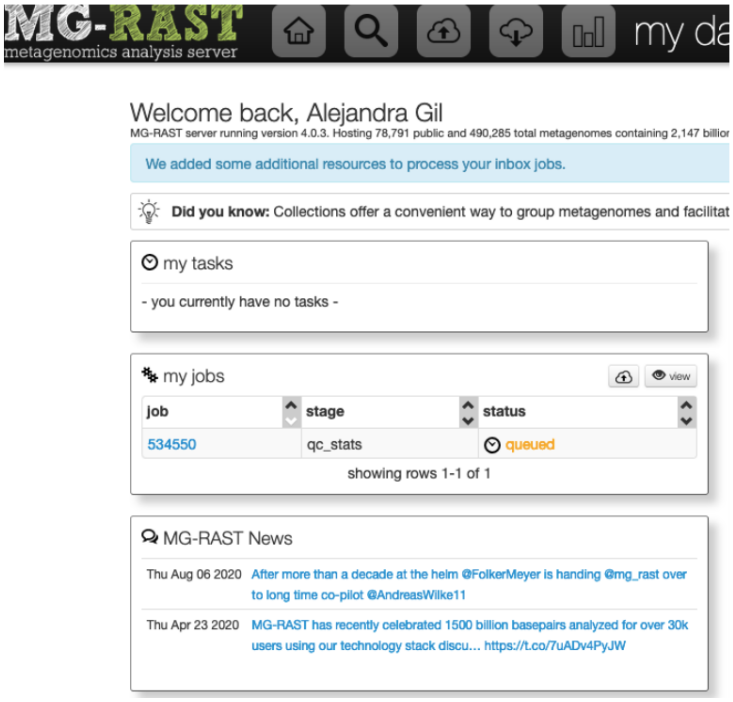

upload
submit
progress
1. select metadata file
2. select project
You have to specify a project to upload a job to MG-RAST. If you have a metadata file, the project must be specified in that file. If you choose to not use a metadata file, you can select a project here. You can either select an existing project or you can choose a new project.
prueba1
select
Note: The projects listed are those that you have write access to. The owners of other projects can provide you with write access if you do not have it.
Asigne un nombre a su proyecto y de clic en select
3. select sequence file(s)
4. choose pipeline options
5. submit
upload
submit
progress
1. select metadata file
2. select project
3. select sequence file(s)
Sequence files from your inbox will appear here. Please note, there is a delay between upload completion and appearing in this table due to sequence statistics calculations. This may be on the order of seconds to hours depending on file size.
metabat.3361.fa
final.contigs.fa
Seleccione el archivo a procesar y de clic en select
select
4. choose pipeline options
5. submit

En el ítem screening puede seleccionar entre algunos organismos modelo (*Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, etc.) para realizar el filtrado de lecturas asociadas a estos organismos. Si las lecturas provienen de una muestra ambiental (suelo, agua, etc.) o del tejido de un organismo no presente en esta lista, se selecciona none.


1. select metadata file
2. select project
3. select sequence file(s)
4. choose pipeline options
Puede mantener las opciones por default
assembled
Select this option if your input sequence file(s)
information within each sequence header as described here.
dereplication
Remove artificial replicate sequences produced by sequencing artifacts Gomez-Alvarez, et al, The ISME Journal (2009)
screening
none
... excepto en este punto. Ajustar de acuerdo al contexto
Remove any host specific species sequences (e.g. p
with bowtie Langmead et al., Genome Biol. 2009,
length filtering
Filter based on sequence length when no quality score information is available.
2.0
Specify the multiplicator of standard deviation for length cutoff.
ambiguous base filtering
Filter based on sequence ambiguity base (non-ACGT) count when no quality score information is available.
5
Specify the maximum allowed number of ambiguous basepairs.
select
5. submit

Resultados

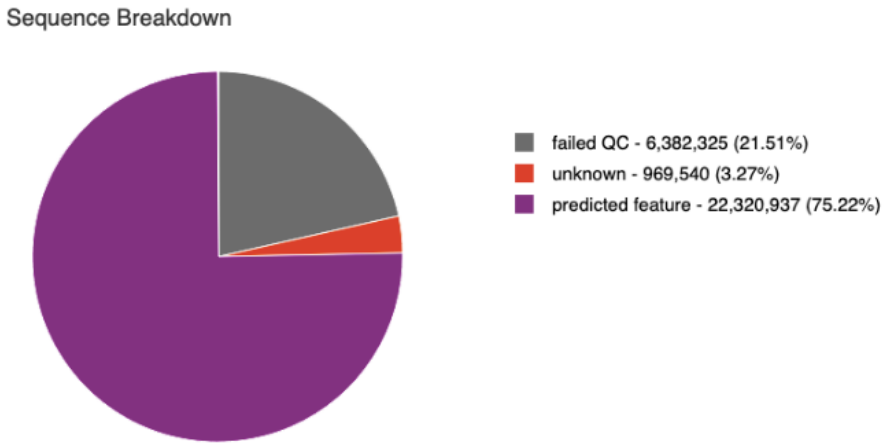
El tiempo de espera es variable, pero puede dar seguimiento al avance de su proyecto en el panel my jobs de la página principal. Inmediatamente después de subir el archivo, posiblemente aparezca en fila (queued); una vez comience a ser analizado, podrá monitorear el avance en este panel.



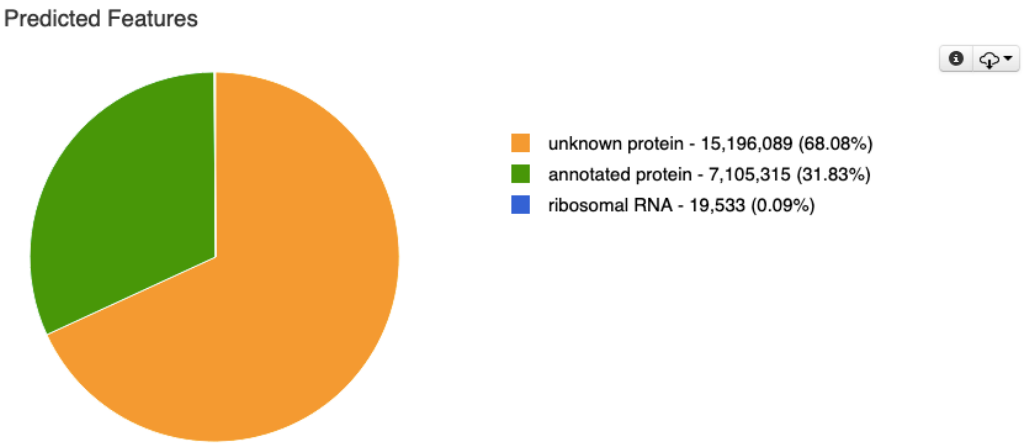
Al igual que en Kaiju, al finalizar el análisis MG-RAST le enviará una notificación vía correo electrónico.

Esto nos redirigirá a una ventana con un reporte extendido sobre la muestra. Incluye gráficos y estadísticos del control de calidad, la asignación taxonómica y la anotación funcional. Para mayor información sobre cada uno de estos aspectos, puede consultar el manual de MG-RAST (https://help.mg-rast.org/user_manual.html).

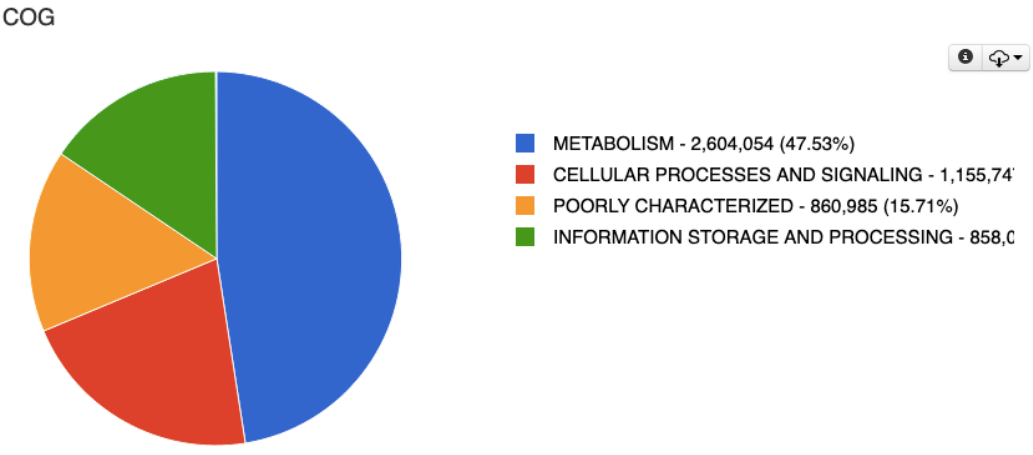
Gráfica control de calidad



Gráfica sobre funciones predichas



Gráfica con funciones predichas por COG



Gráfica con Dominios y Phylum identificados