nmdc

National Microbiome Data Collaborative

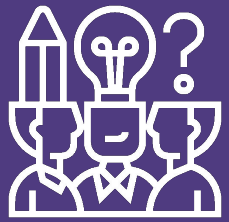
Workflows metagenómicos estandarizados utilizando NMDC EDGE

April 13, 2023

Pacific Northwest NATIONAL LABORATORY

Qué es el NMDC?

nmdc
National Microbiome Data Collaborative



Visión

Conectar data, **personas**, e **ideas** para avanzar la innovación y el descubrimiento de microbiomas

Misión

Apoyar una red de compartimiento de datos a base de principios FAIR para enfrentar retos en las ciencias ambientales mediante **infraestructura**, **estándares de datos**, y **formación comunidades**

FAIR Data

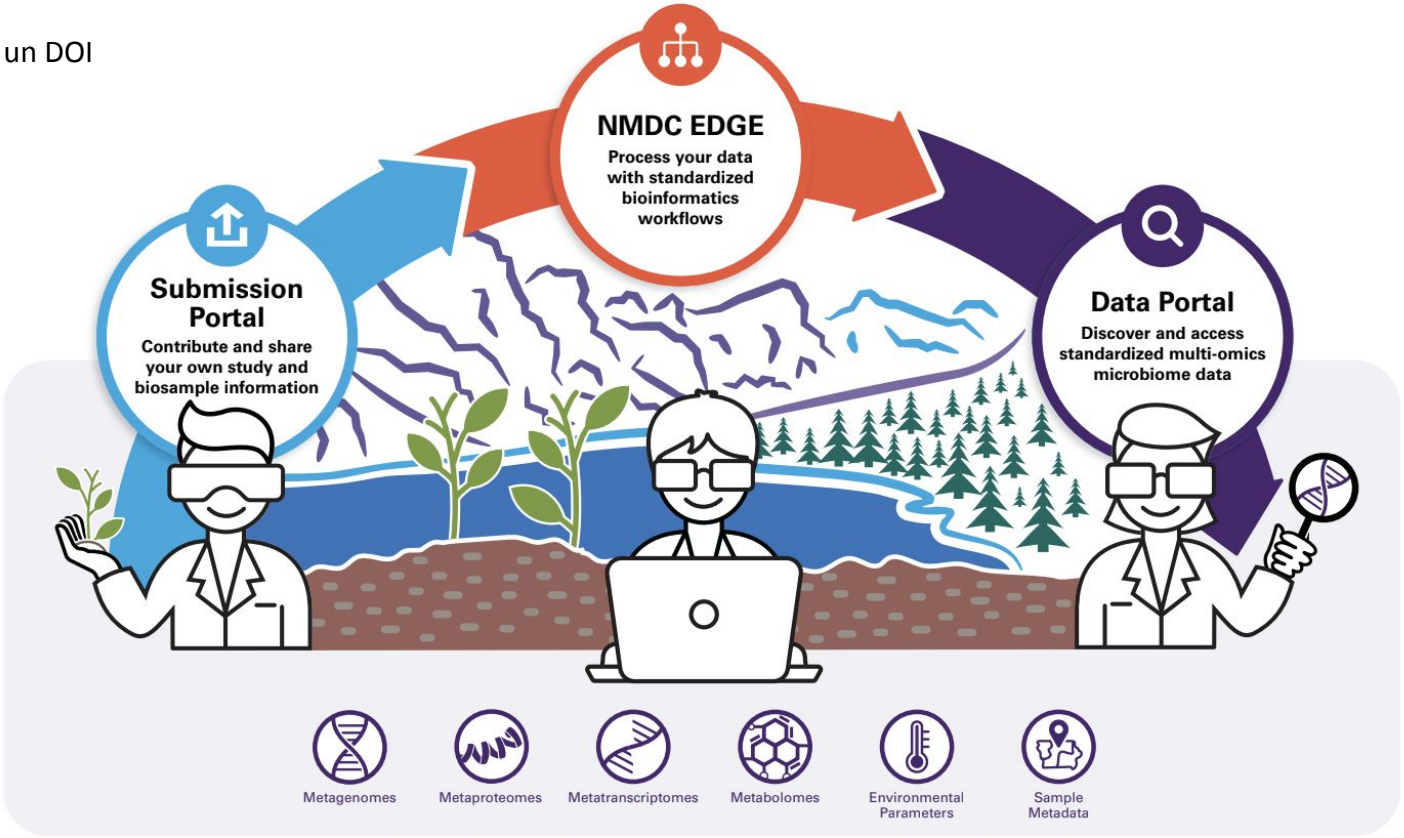
nmdc
National Microbiome Data Collaborative

- NMDC está comprometido a los principios FAIR para asegurar que todo dato sea fácil de encontrar, accesible, interoperable, y reusable.
- La data “cruda” al igual que la procesada deberia ser FAIR!
- Procesar data que está estandarizada ayuda muchísimo en crear data interoperable y reusable.

Los 3 componentes del NMDC

nmdc
National Microbiome Data Collaborative

1. Portal de someter datos:
Comparte tu investigación y adquiere un DOI

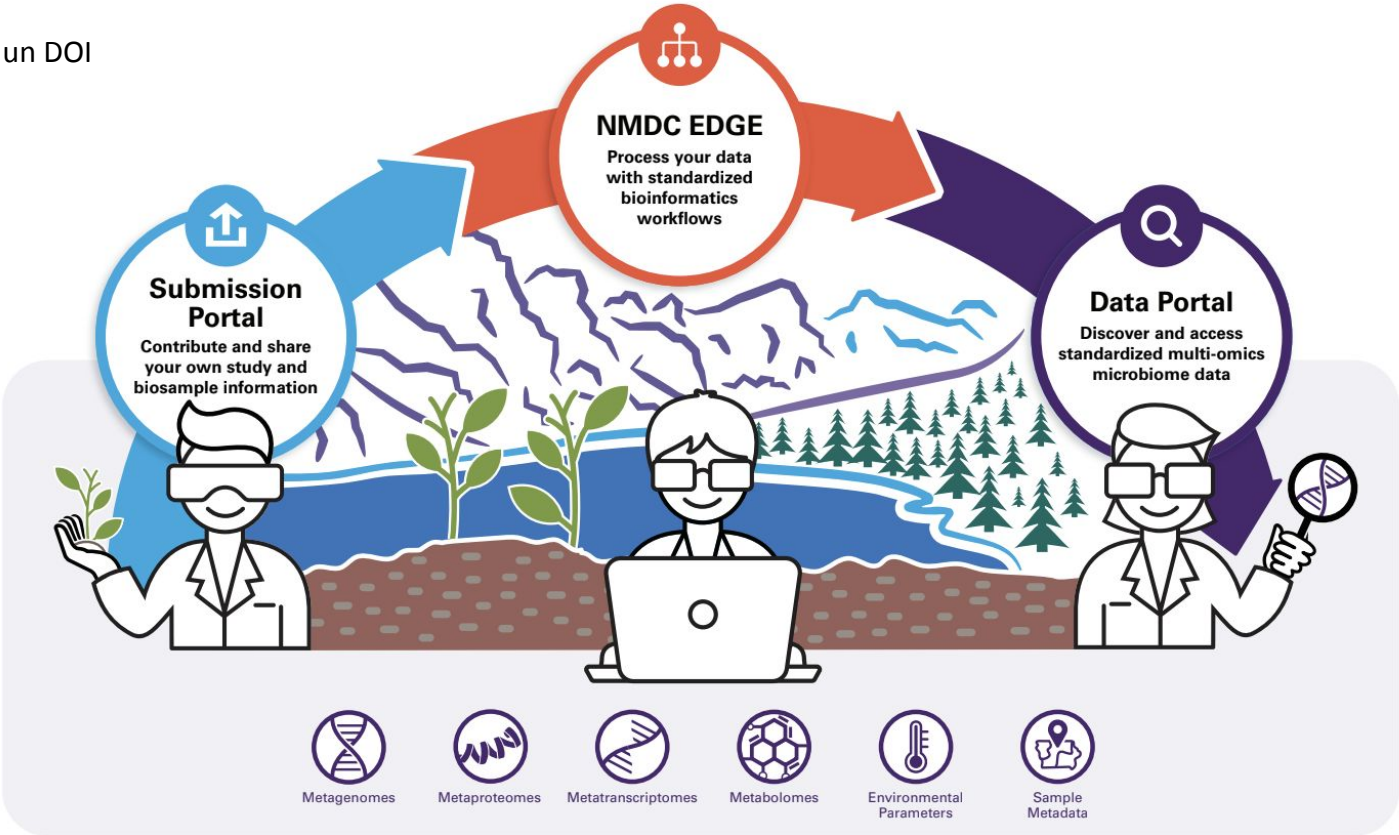


Los 3 componentes del NMDC

nmdc
National Microbiome Data Collaborative

1. Portal de someter datos:
Comparte tu investigación y adquiere un DOI

2. NMDC EDGE:
Procesamiento de datos y workflows bioinformaticos a base web.



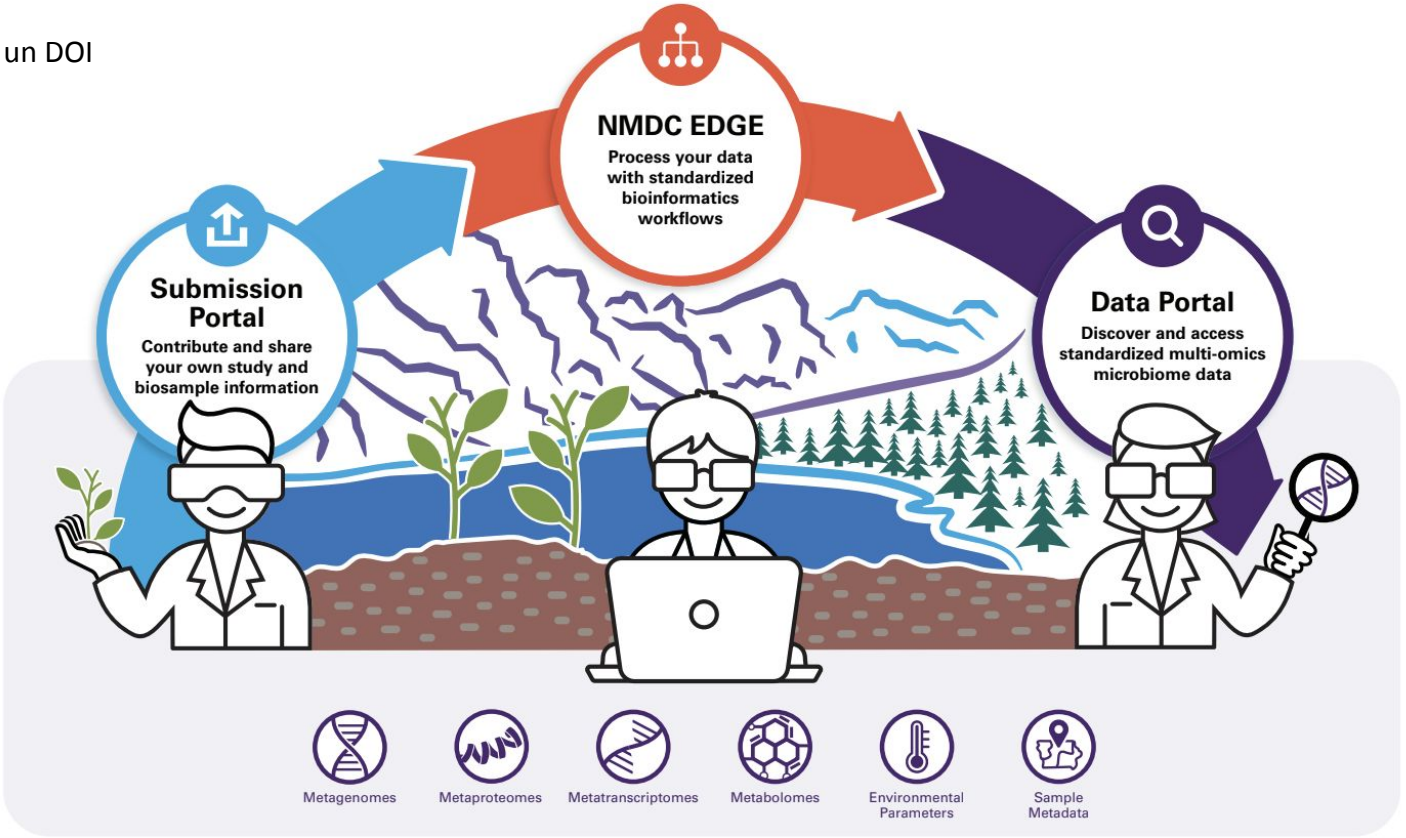
Los 3 componentes del NMDC

nmdc
National Microbiome Data Collaborative

1. Portal de someter datos:
Comparte tu investigación y adquiere un DOI

2. NMDC EDGE:
Procesamiento de datos y workflows bioinformaticos a base web.

3. Portal de datos:
Descubre, analiza, y accede datos de microbiomas estandarizados



Los 3 componentes del NMDC

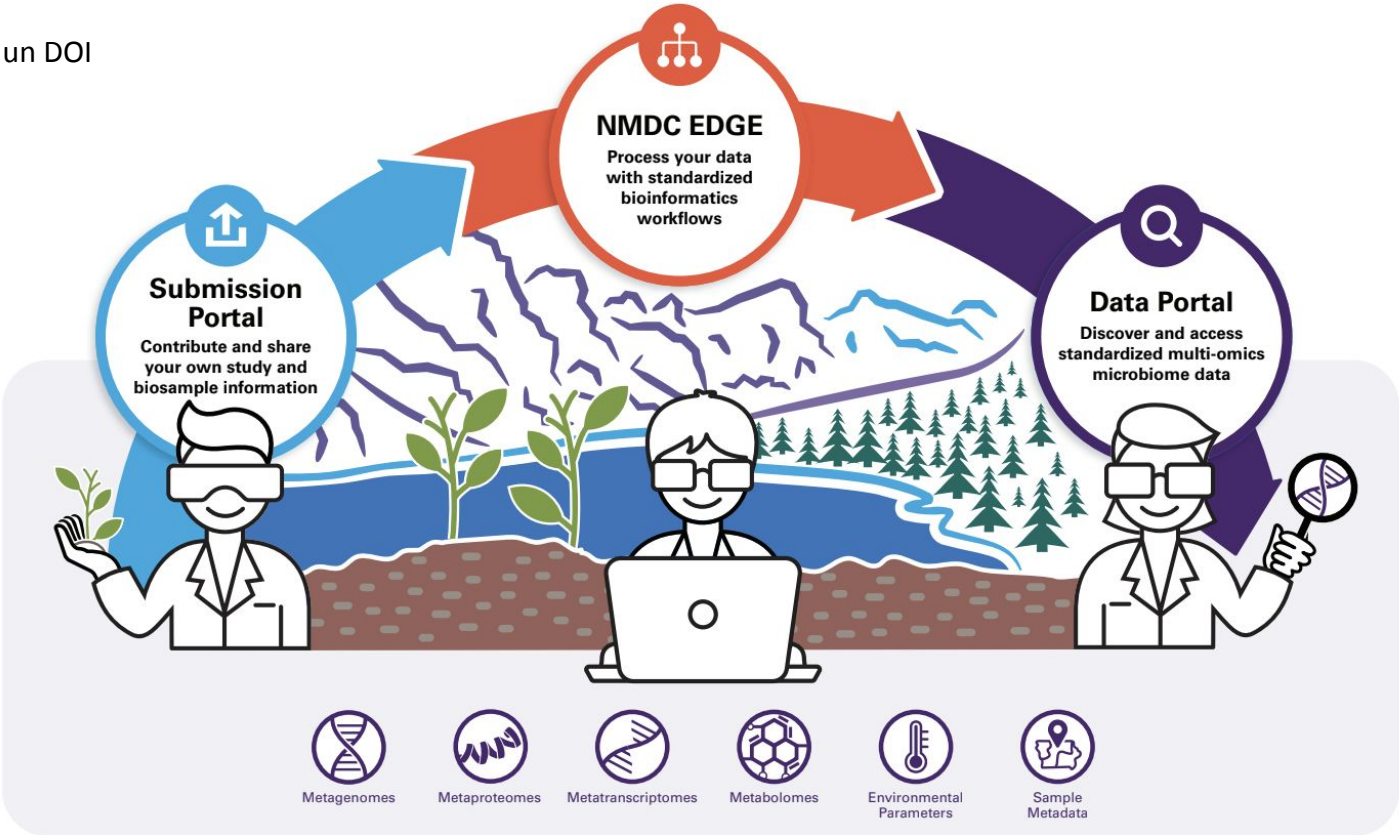
nmdc
National Microbiome Data Collaborative

1. Portal de someter datos:
Comparte tu investigación y adquiere un DOI

2. NMDC EDGE:
Procesamiento de datos y workflows bioinformaticos a base web.

3. Portal de datos:
Descubre, analiza, y accede datos de microbiomas estandarizados

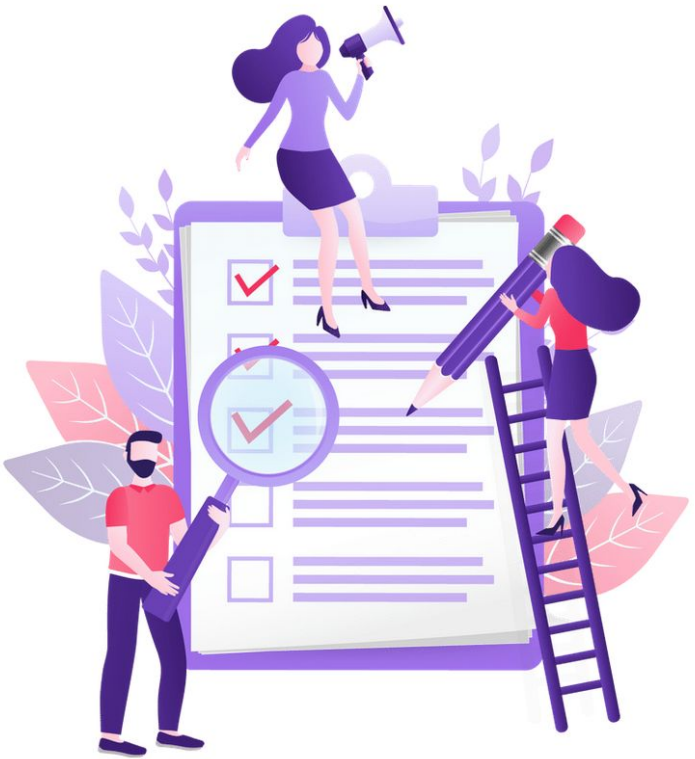
Tipos de datos que se pueden procesar / someter a NMDC:
-Metagenomas
-Metaproteomas
-Metatranscriptomas
-Metabolomas
-Medidas de medio ambiente
-Metadata asociada a muestras



Agenda para hoy

nmdc
National Microbiome Data Collaborative

- Portal de someter datos NMDC
  - Metadata estandarizada
  - Proposito de metadata
- Workflows estandarizados
  - Workflow metagenómico
  - Workflows adicionales
- NMDC EDGE
  - Walk-through
  - Activity
  - Beta Testing
- Preguntas
- Recursos



nmdc

National Microbiome Data Collaborative

Estandarización de Metadata y el Portal para Someter Datos

Qué es Metadata?

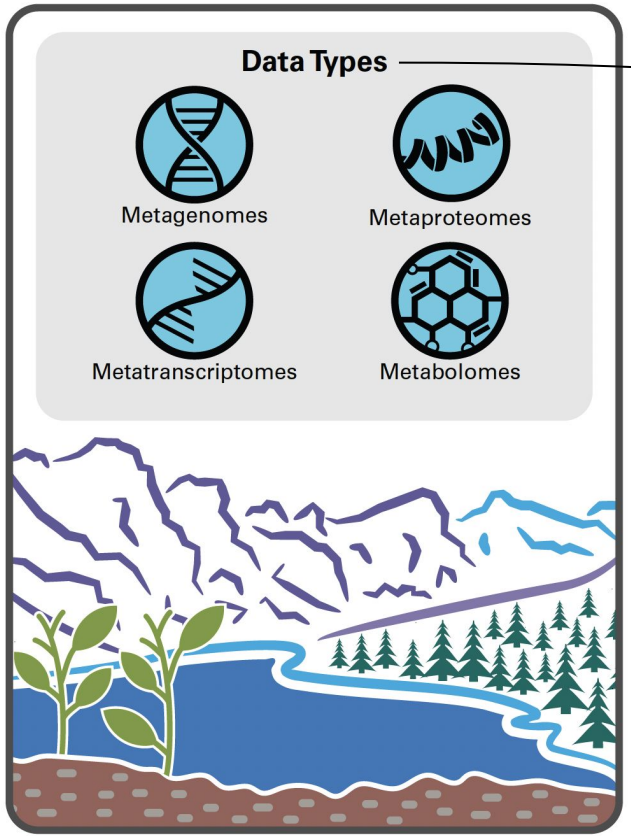
nmdc
National Microbiome Data Collaborative

Metadata es ...

- Data contextual sobre tu data
- Esencial para la data:
  - Publicación y depósito de datos
  - Preservación
  - Descubrimiento
  - Acceso
  - Reusabilidad

Hay una gran cantidad de información en estudios de microbiomas

nmdc
National Microbiome Data Collaborative



Información sobre el estudio

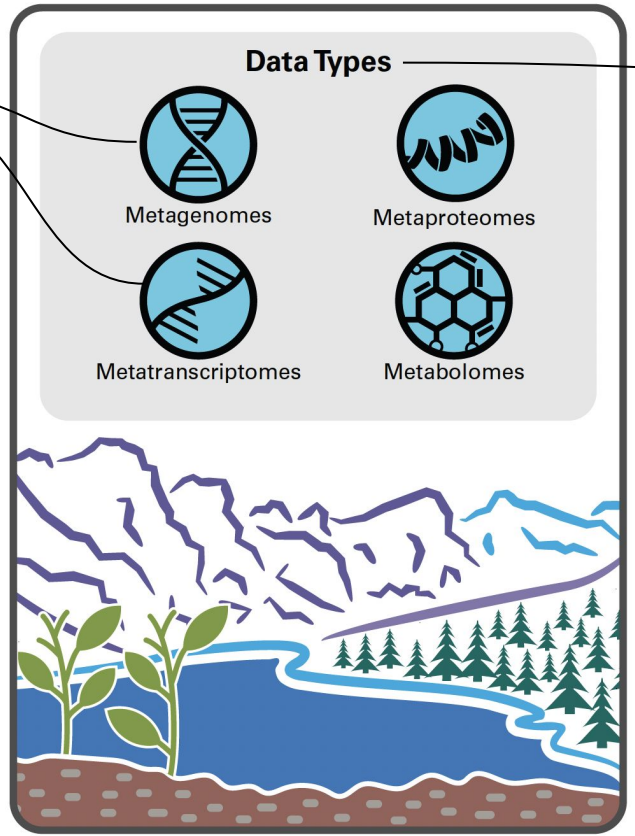
PI & Contribuidores?
Diseño experimental?

Hay una gran cantidad de información en estudios de microbiomas

nmdc
National Microbiome Data Collaborative

Método de secuenciación

Kits de extracción?
Tipo de secuenciación?



Información sobre el estudio

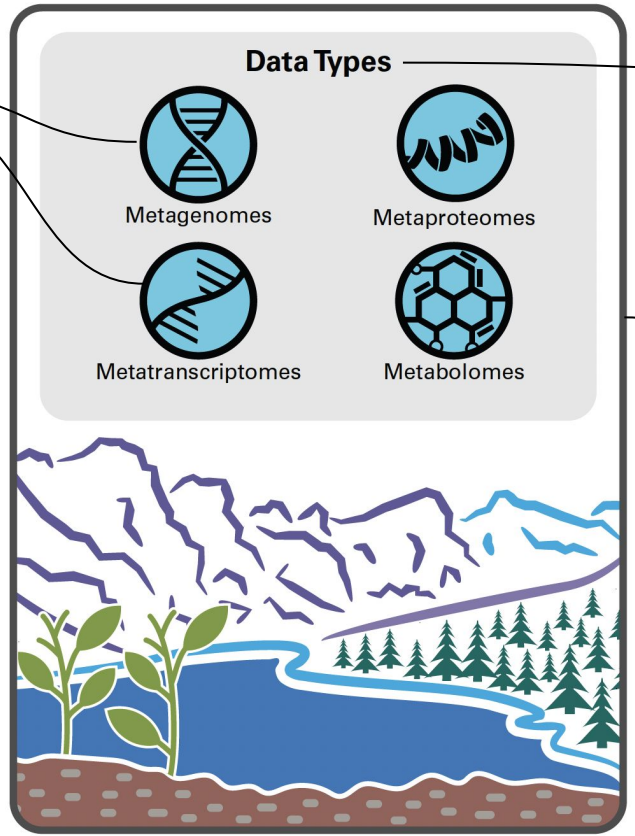
PI & Contribuidores?
Diseño experimental?

Hay una gran cantidad de información en estudios de microbiomas

nmdc
National Microbiome Data Collaborative

Método de secuenciación

Kits de extracción?
Tipo de secuenciación?



Información sobre el estudio

PI & Contribuidores?
Diseño experimental?

Muestras con tratamientos y preparaciones distintos

Dispositivo de muestreo?
Cómo se procesaron?

Hay una gran cantidad de información en estudios de microbiomas

nmdc
National Microbiome Data Collaborative

Método de secuenciación

Kits de extracción?
Tipo de secuenciación?

Múltiples parámetros ambientales y biogeoquímicos

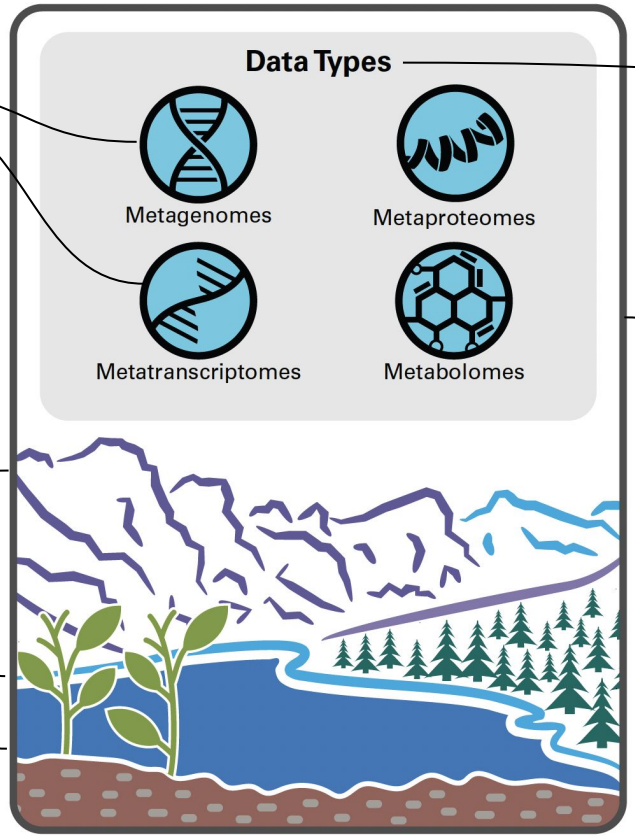
Clima
Elevación

lat/lon
bioma
material
profundidad

Tipo veg.
Parte de planta

pH
temperatura
carbono
nitrogeno
...

Tipo de suelo



Información sobre el estudio

PI & Contribuidores?
Diseño experimental?

Muestras con tratamientos y preparaciones distintos

Dispositivo de muestreo?
Cómo se procesaron?

Hay una gran cantidad de información en estudios de microbiomas

nmdc
National Microbiome Data Collaborative

Método de secuenciación

Kits de extracción?
Tipo de secuenciación?

Múltiples parámetros ambientales y biogeoquímicos

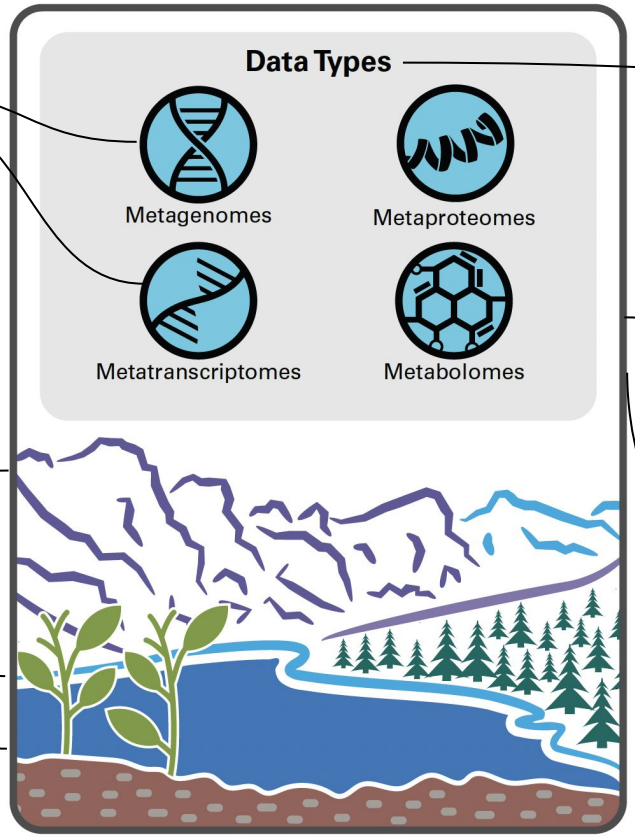
Clima
Elevación

lat/lon
bioma
material
profundidad

Tipo veg.
Parte de planta

pH
temperatura
carbono
nitrogeno
...

Tipo de suelo



Información sobre el estudio

PI & Contribuidores?
Diseño experimental?

Muestras con tratamientos y preparaciones distintos

Dispositivo de muestreo?
Cómo se procesaron?

Resultados de análisis

Estadisticas de ensamblaje
Función de genes
Cantidad de metabolitos y péptidos
Abundancias taxonómicas

Metadata estandarizada es esencial

nmdc
National Microbiome Data Collaborative

Data similares son difíciles de utilizar si carecen de consistencia en lenguaje y formato

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Metadata estandarizada es esencial

nmdc
National Microbiome Data Collaborative

Data similares son difíciles de utilizar si carecen de consistencia en lenguaje y formato

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Metadata estandarizada es esencial

nmdc
National Microbiome Data Collaborative

Data similares son difíciles de utilizar si carecen de consistencia en lenguaje y formato

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Metadata estandarizada es esencial

nmdc
National Microbiome Data Collaborative

Data similares son difíciles de utilizar si carecen de consistencia en lenguaje y formato

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

sampleNum	substance	sample depth	temp
8725	dirt	45 cm	21.1
2312	ground liquid	105 cm	7

Metadata estandarizada es esencial

nmdc
National Microbiome Data Collaborative

Data similares son difíciles de utilizar si carecen de consistencia en lenguaje y formato

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

idNumber	material	sample depth	temperature
8725	soil	.45 m	21.1 °C
2312	groundwater	1.05 m	7 °C

Metadata estandarizada es esencial

nmdc
National Microbiome Data Collaborative

Data similares son difíciles de utilizar si carecen de consistencia en lenguaje y formato

idNumber	material	sample depth	temperature
3928	soil	0.03 m	23.2 °C
3234	groundwater	1 m	9.02 °C

idNumber	material	sample depth	temperature
8725	soil	.45 m	21.1 °C
2312	groundwater	1.05 m	7 °C

nmdc

National Microbiome
Data Collaborative

Estándares de Metadata NMDC

Estándares de la comunidad

nmdc
National Microbiome Data Collaborative

NMDC utiliza requisitos de metadata basados en estándares de la comunidad

1. MIxS: Minimum Information about any (x) Sequence
   Genomic Standards Consortium (GSC)
2. GOLD: Genomes OnLine Database
   Joint Genome Institute (JGI)
4. EnvO: Environment Ontology
   Open Biological and Biomedical Ontology (OBO) Foundry

Estándares de la comunidad

nmdc
National Microbiome Data Collaborative

MIxS

EnvO

GOLD

MIxS / EnvO		
Ambiente-Escala Amplia	Ambiente-Escala Local	Ambiente-Medio
Bioma de lago de agua dulce	Costado de lago	Sedimento
Bioma de lago de agua dulce	Lago	Floración de alga

Términos de GSC, EnvO y GOLD nos dan un contexto ambiental mejorado sobre muestras de microbiomas!

Clasificación de ecosistemas GOLD				
Ecosistema	Ecosystem Category	Ecosystem Type	Specific Ecosystem	Ecosystem Tree
Ambiente	Acuático	Agua dulce	Lago	Sedimento
Ambiente	Acuático	Agua dulce	Lago	Floración de alga

GOLD - Ecosistemas en 5 niveles

nmdc
National Microbiome Data Collaborative

Clasificación de ecosistema GOLD	*Ejemplo: Sedimento de un lago*
Ecosistema	Ambiental
Categoría de ecosistema	Acuático
Tipo de ecosistema	Agua dulce
Sub-tipo de ecosistema	Lago
Ecosistema específico	Sedimento

https://gold.jgi.doe.gov/ecosystemtree

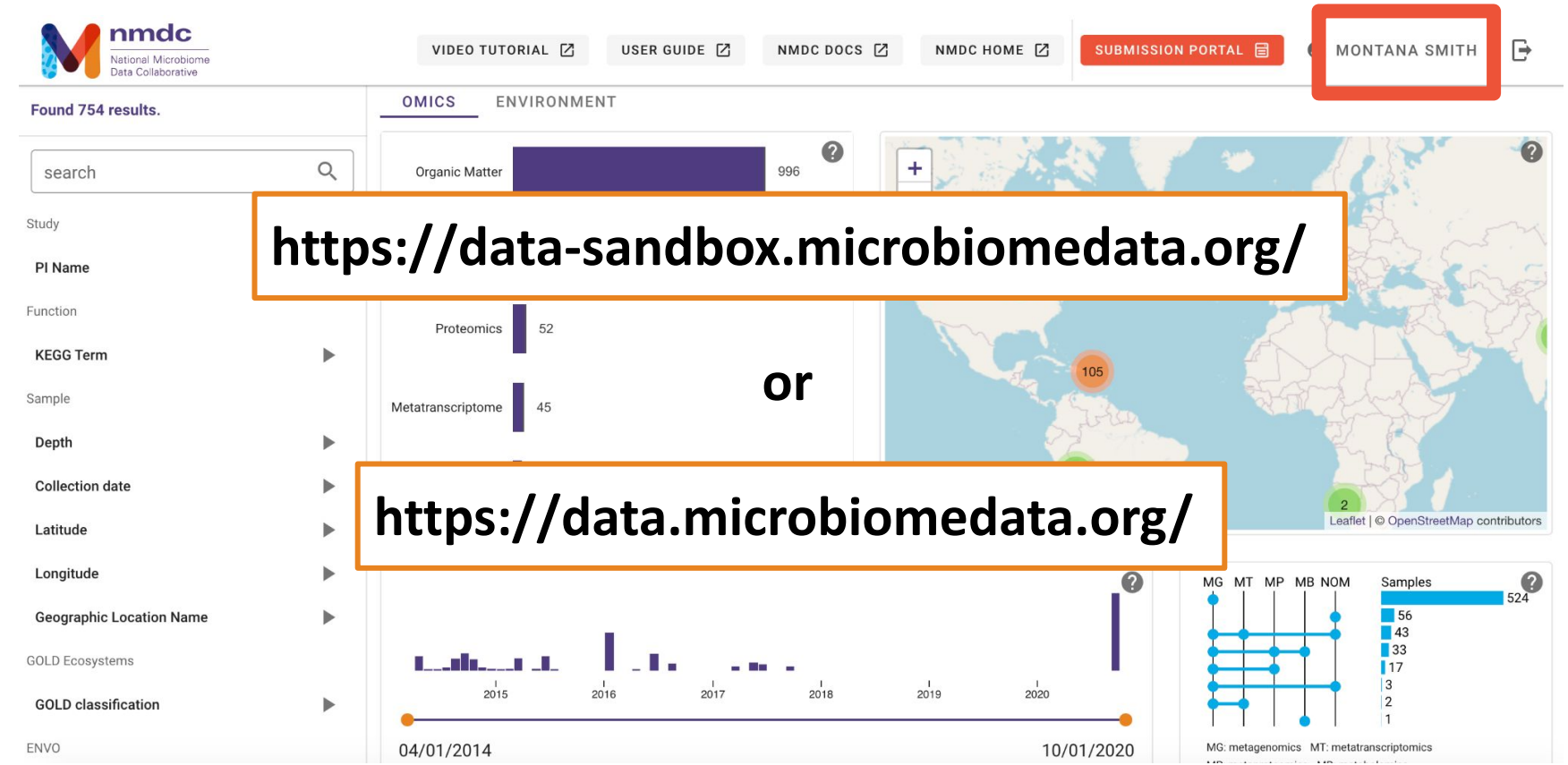
nmdc

National Microbiome
Data Collaborative

Portal de someter datos NMDC

Portal para someter datos NMDC

nmdc
National Microbiome Data Collaborative



https://data-sandbox.microbiomedata.org/

or

https://data.microbiomedata.org/

nmdc

National Microbiome
Data Collaborative

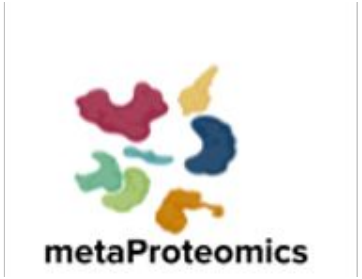
Workflows metagenómicos

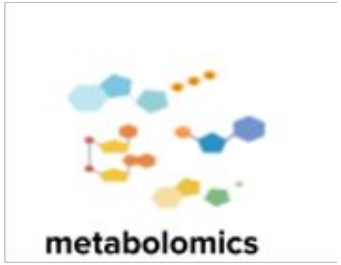
Procesamiento de data multi-ómica

nmdc
National Microbiome Data Collaborative

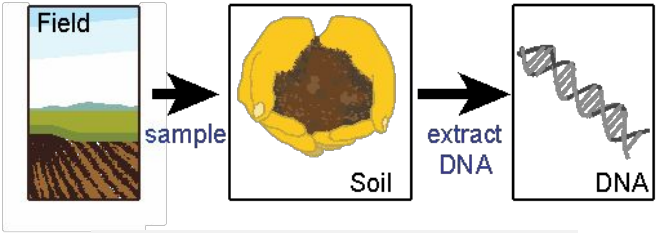
- Colección de datos multi-ómicos se está convirtiendo en una de las maneras más eficaces de interrogar los microbiomas
  - La infraestructura para este tipo de data no da abasto
    - Es necesario una gran capacidad computacional para almacenar y procesar estos datos
    - Datos usualmente no adhieren a los principios FAIR
    - Distintos tipos de datos “ómicos” no están conectados y no se pueden comparar

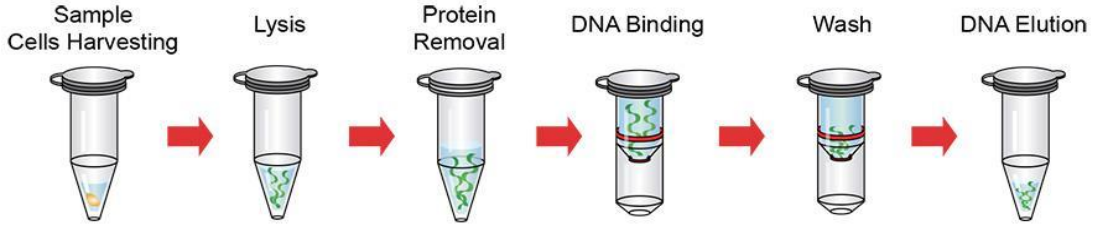
- Es difícil procesar estos datos ómicos de manera efectiva para comparar entre estudios distintos y para reusabilidad de datos.





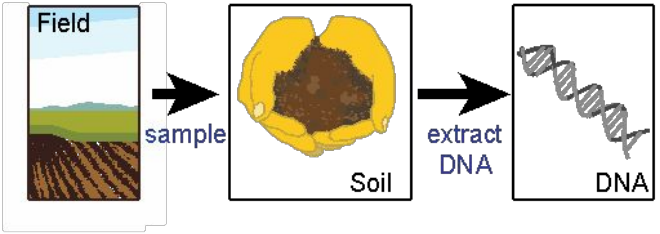
Parte 1: Muestreo y extracción de ADN

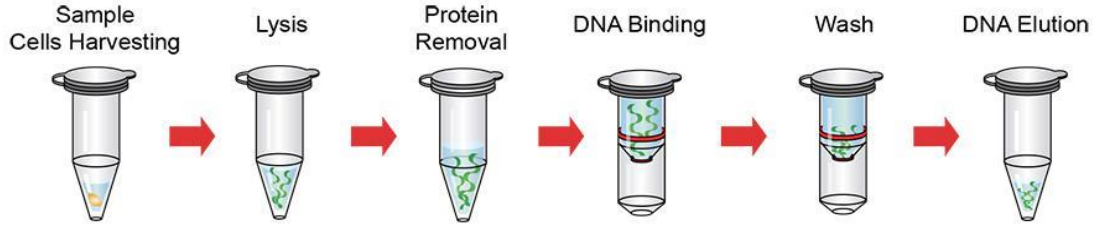


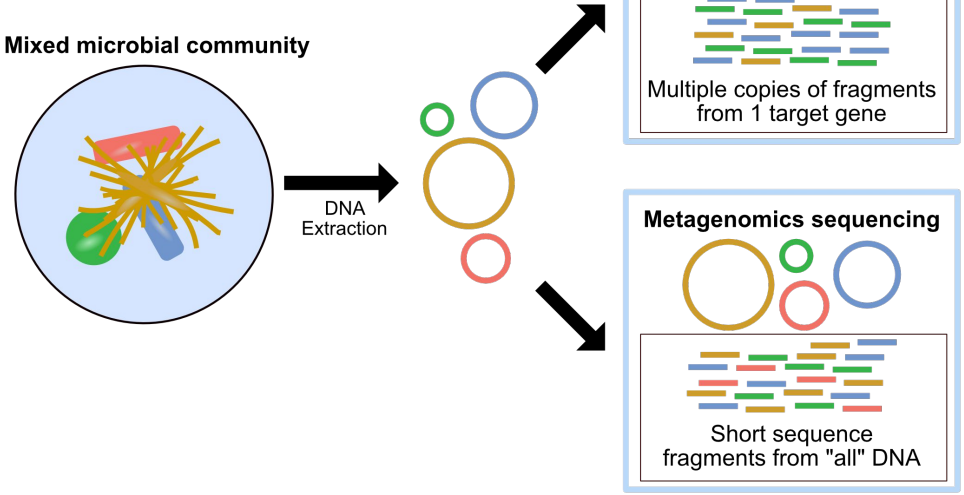


Se extrae el ADN directamente desde una muestra ambiental utilizando un kit de extracción

Parte 1: Metagenómica **NO** es 16S rRNA





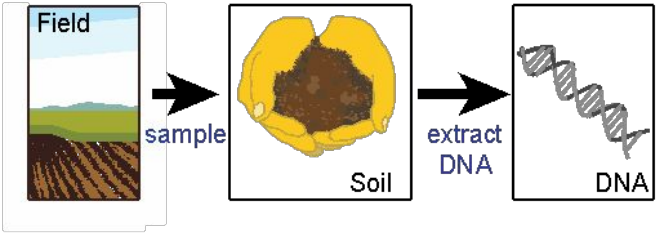


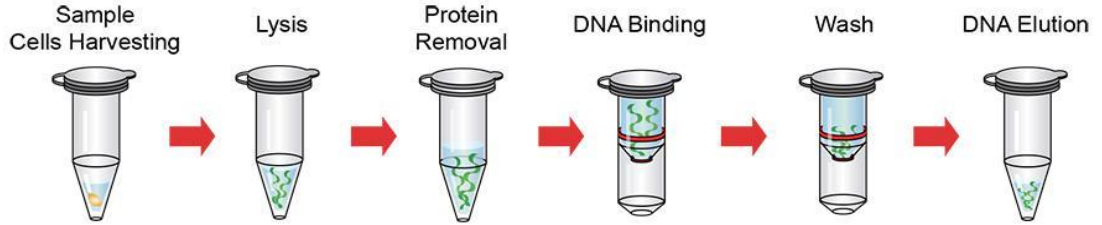
Amplificas y secuencias 1 solo gen

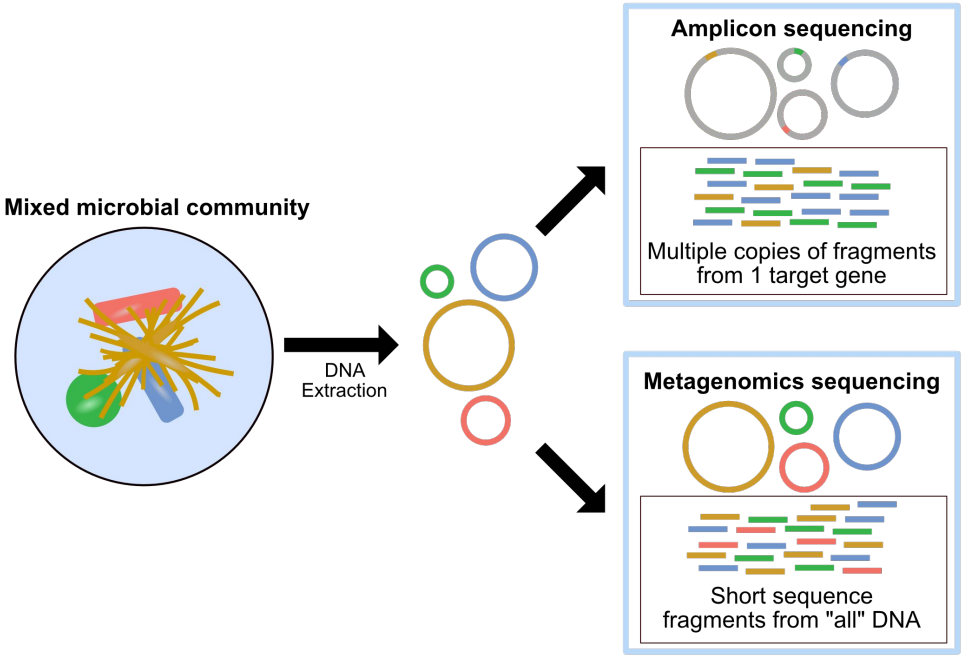
Secuenciación de todos los genes de una muestra

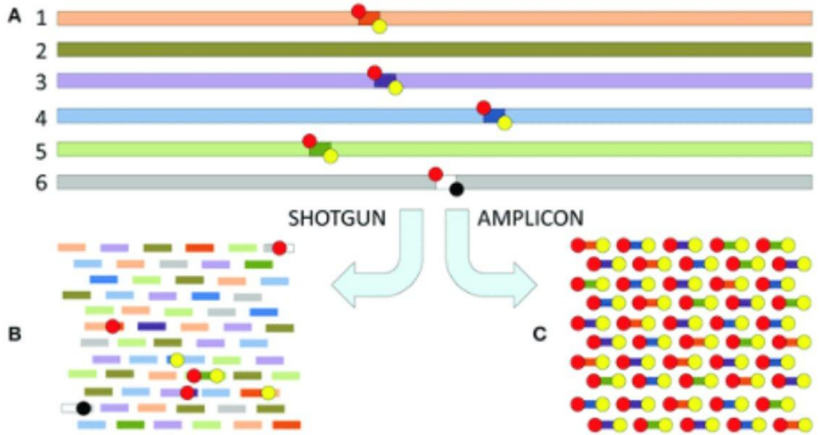
https://astrobiomike.github.io/misc/amplicon_and_metagen

Parte 1: Metagenómica **NO** es 16S rRNA

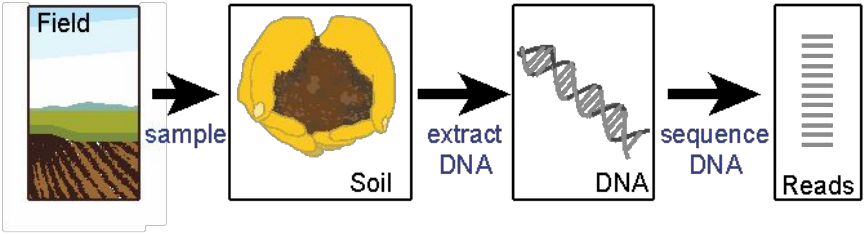








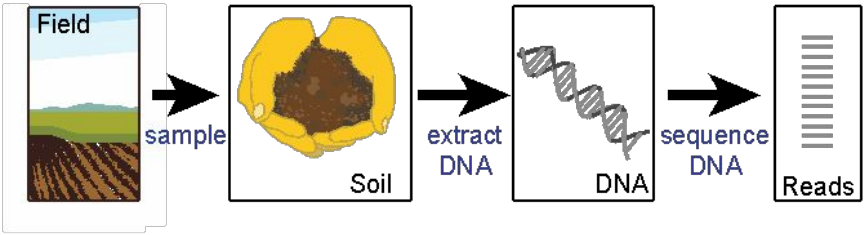
Parte 2: Enviar a secuenciar y recibir “reads”



Se envía la ADN a un centro de secuenciación como el Joint Genome Institute (JGI) o University of Colorado Denver.

Ellos nos devuelven “reads”

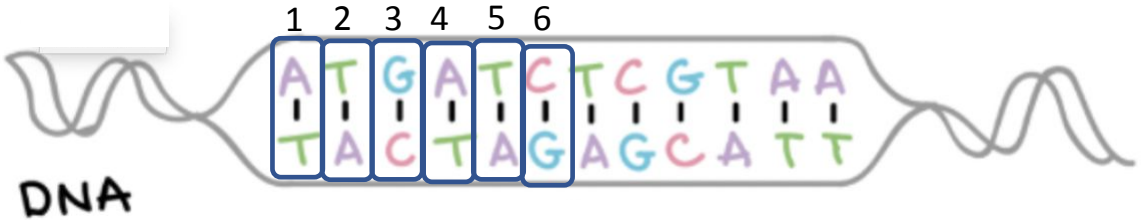
Parte 2: Enviar a secuenciar y recibir “reads”



Se envía la ADN a un centro de secuenciación como el Joint Genome Institute (JGI) o University of Colorado Denver.

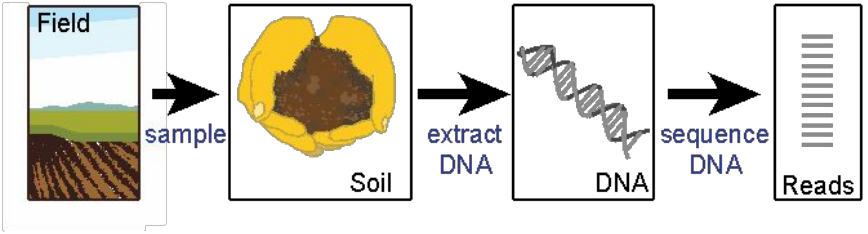
Ellos nos devuelven “reads”

Reads se miden en “base pair length” (bp)



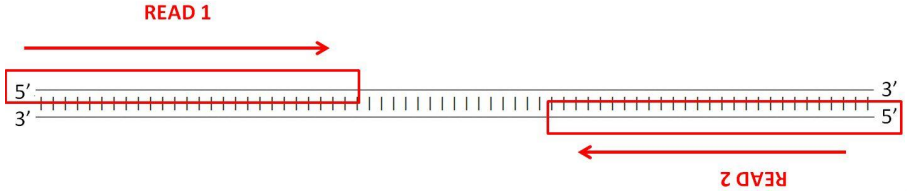
Este es el tipo de secuencia de ADN que nos devuelven

Parte 2: Enviar a secuenciar y recibir “reads”

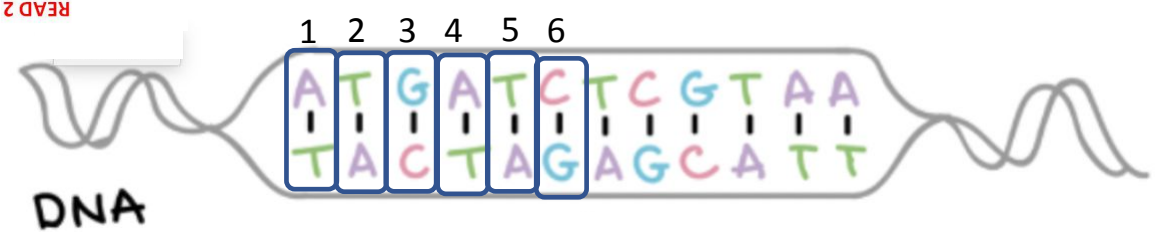


Se envía la ADN a un centro de secuenciación como el Joint Genome Institute (JGI) o University of Colorado Denver.

Ellos nos devuelven “reads”



Reads se miden en “base pair length” (bp)



Este es el tipo de secuencia de ADN que nos devuelven

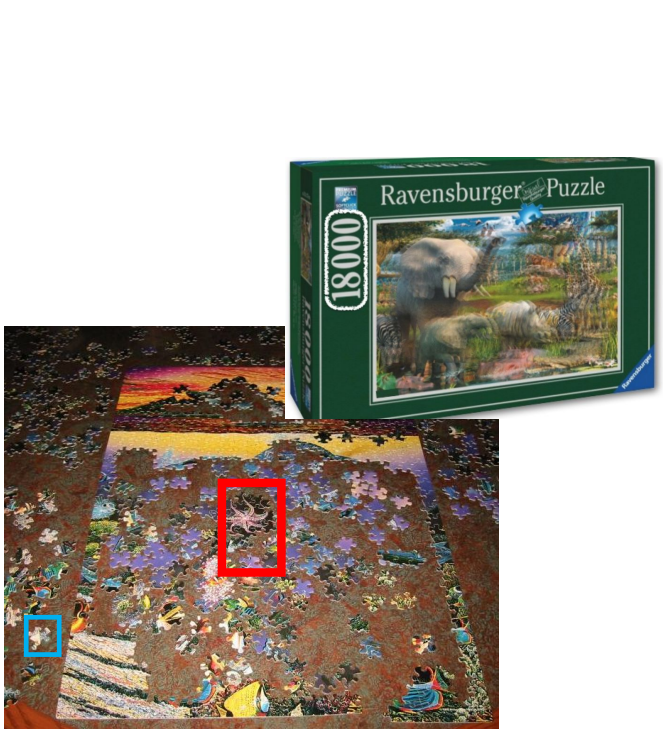
Parte 3: Ensamblaje de reads

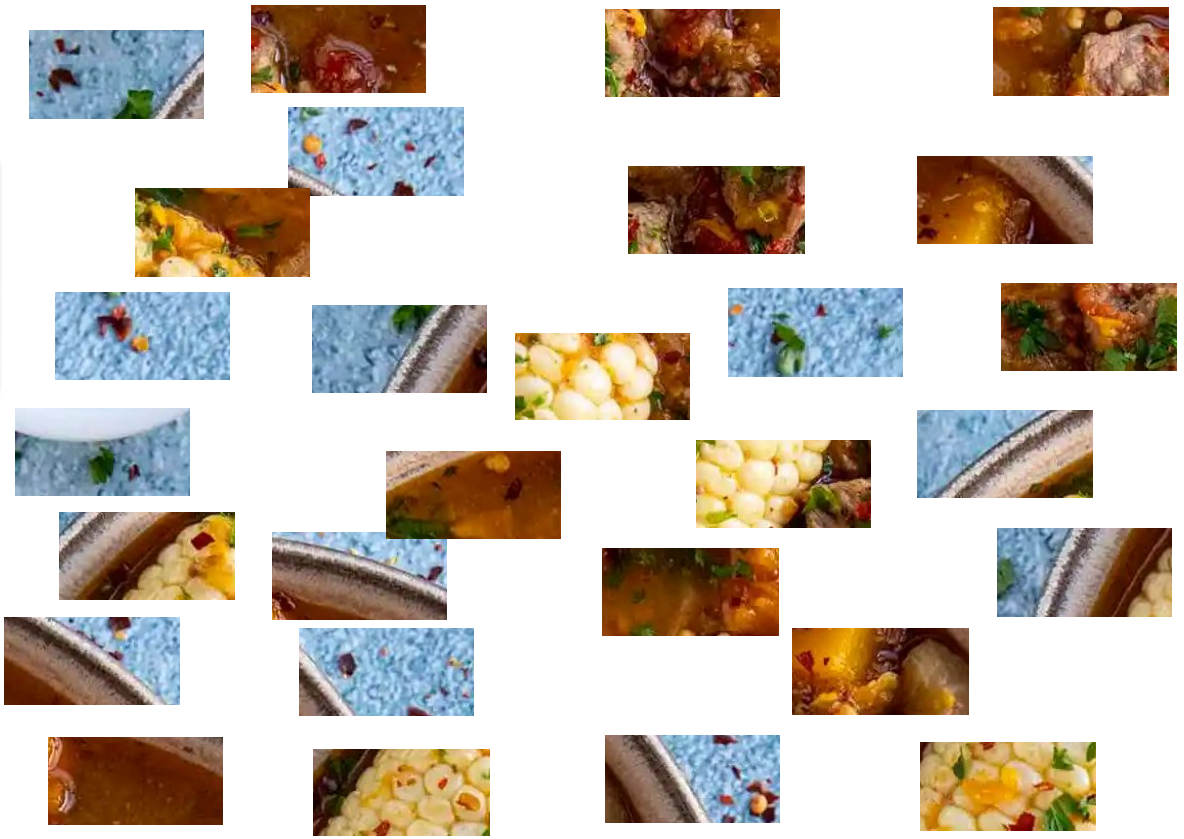
Tienes millones de reads que proveen de una muestra, sin mucha idea de cómo se orientan, y cuales van juntos

MidJourney AI: “Millions of metagenomic reads in a bowl”

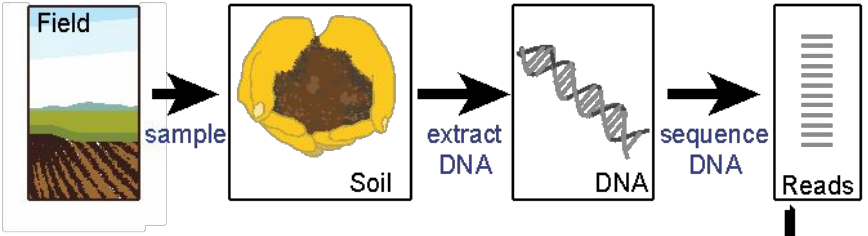
Parte 3: Ensamblaje de reads

Parte 3: Ensamblaje de reads





Part 3: Ensamblaje de reads

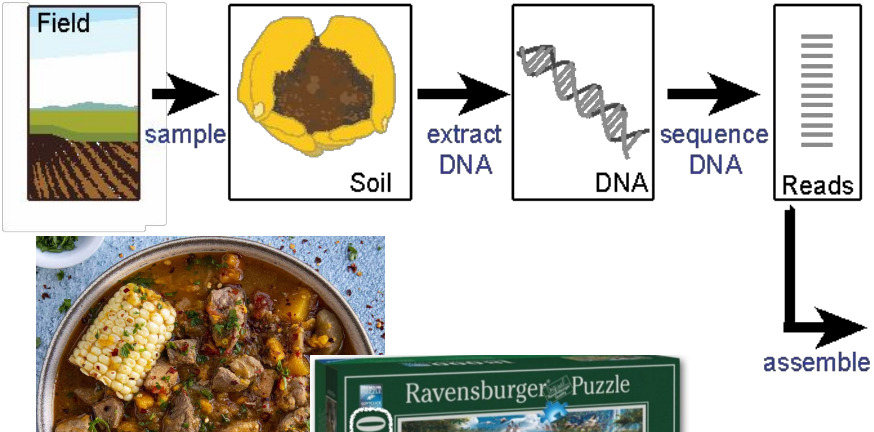


Un “assembly” es el ensamblaje computacional de tus reads de ~150 bp.

La meta es ir ensamblando reads poco a poco para crear secuencias contiguas o “contigs”. También se les dice “scaffolds”.

Mega-Hit
IDBA-UD
MetaSPAdes
(…) etc. etc. etc.

Part 3: Ensamblaje de reads



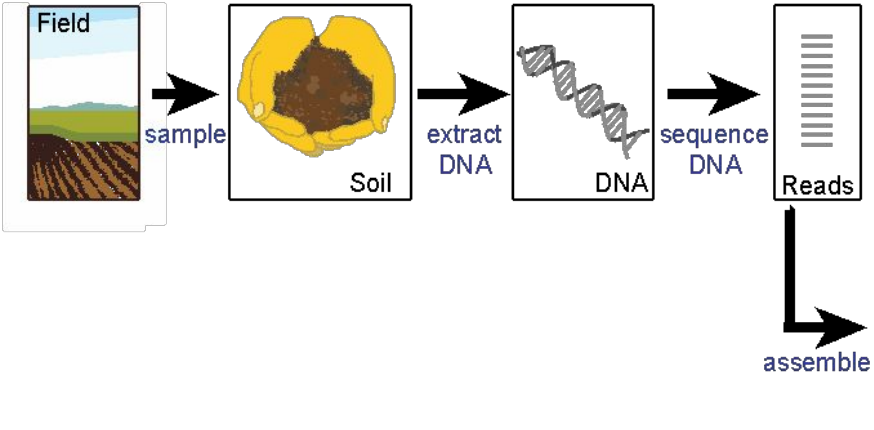
Un “assembly” es el ensamblaje computacional de tus reads de ~150 bp.

La meta es ir ensamblando reads poco a poco para crear secuencias contiguas o “contigs”. También se les dice “scaffolds”.

Muchas veces, se ensambla “de-novo” sin ningún tipo de guía. Es muy intensivo en cuestión de memoria (RAM + procesadores).

~400GB de memoria con 20 procesadores en serie

Part 3: Ensamblaje de reads



Un “assembly” es el ensamblaje computacional de tus reads de ~150 bp.

La meta es ir ensamblando reads poco a poco para crear secuencias contiguas o “contigs”. También se les dice “scaffolds”.

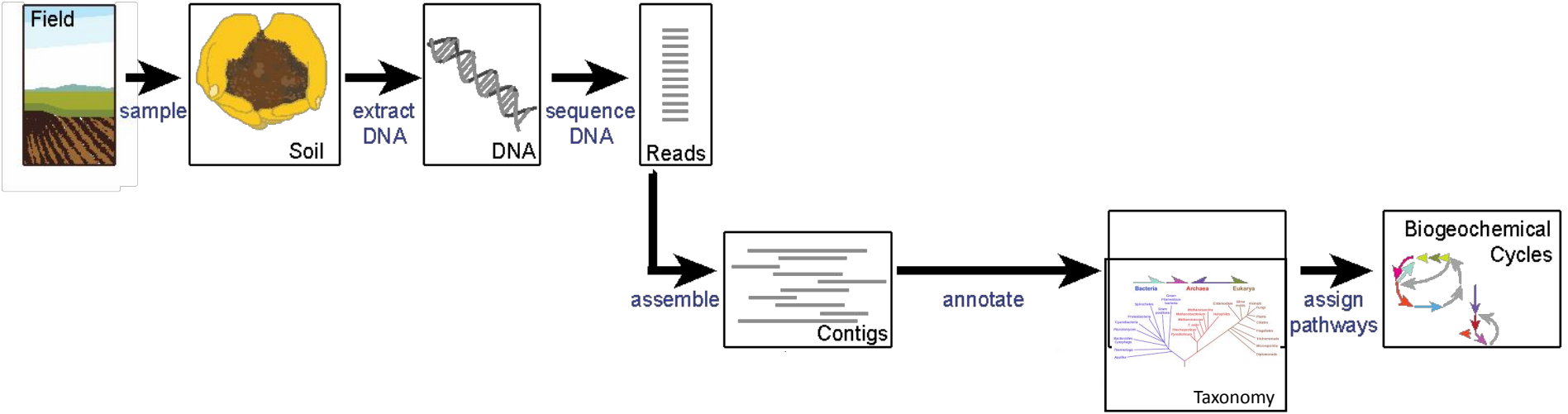
Mega-Hit
IDBA-UD
MetaSPAdes
(…) etc. etc. etc.

- Cada “read” es un pedazo del rompecabeza (~150bp)
- Un “contig” / ”scaffold” es un fragmento ensamblado de reads
- Un “assembly” es el rompecabezas completado

Tenemos partes más grandes de un rompecabezas relativamente ensamblado (contigs).

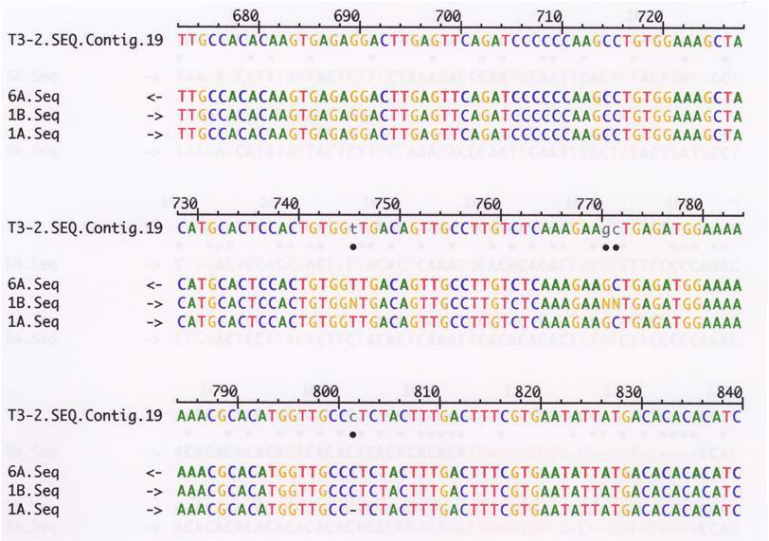
Y ahora qué?

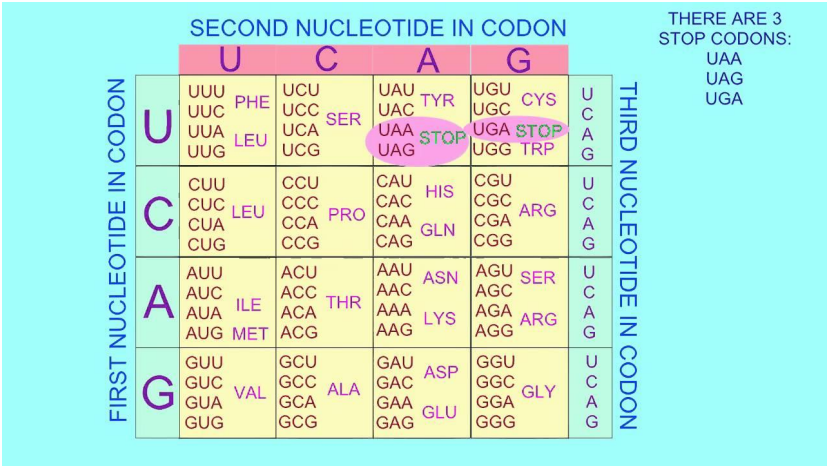
Parte 4a: Método “un-binned”

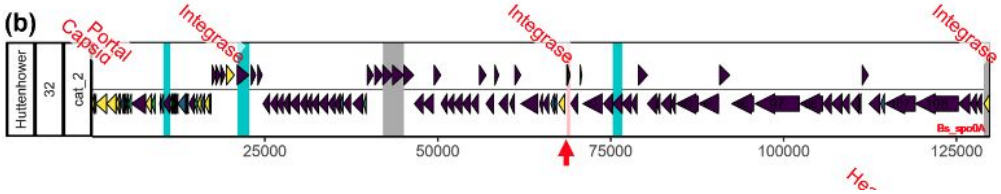


Parte 4a: Método de metagenómica “un-binned”

- Llamamos “genes” dentro de nuestros contigs utilizando programas como Prodigal que usan la tabla de código genético.



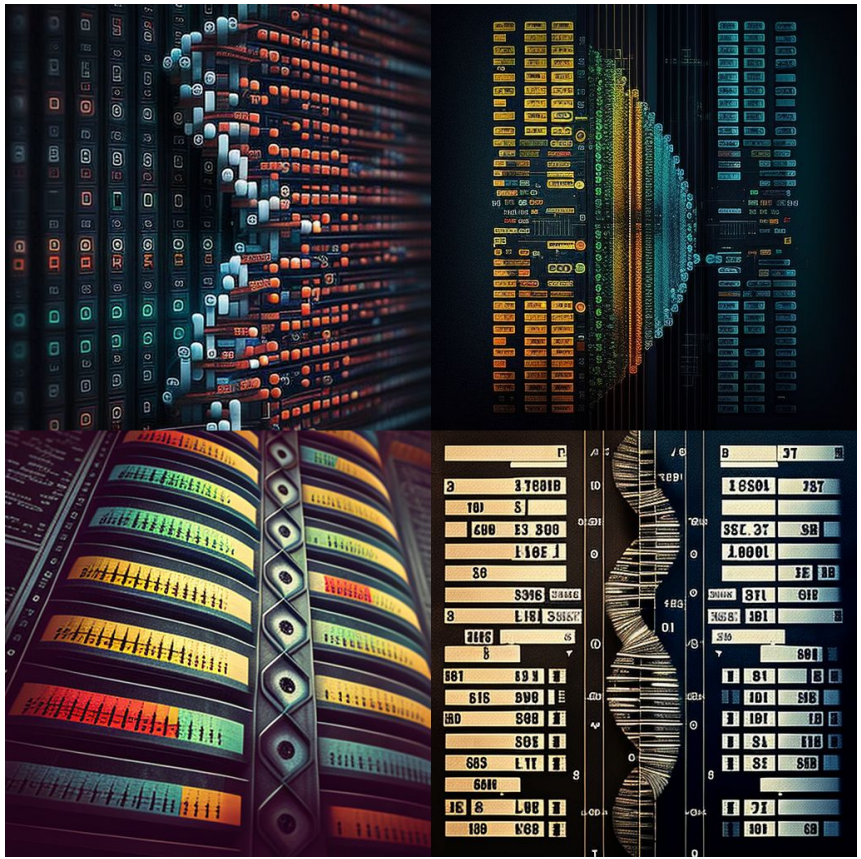




Parte 4a: Método de metagenómica “un-binned”

- Llamamos “genes” dentro de nuestros contigs utilizando programas como Prodigal que usan la tabla de código genético.

- Luego alineamos esos “genes” a databases de genes que ya conocemos, y adivinamos a base de homología / similitud qué son. (KRAKEN2, MG-Rast, SEED, BLASTn/BLASTp, KEGG)



MidJourney AI: “Matching a genetic sequence to a

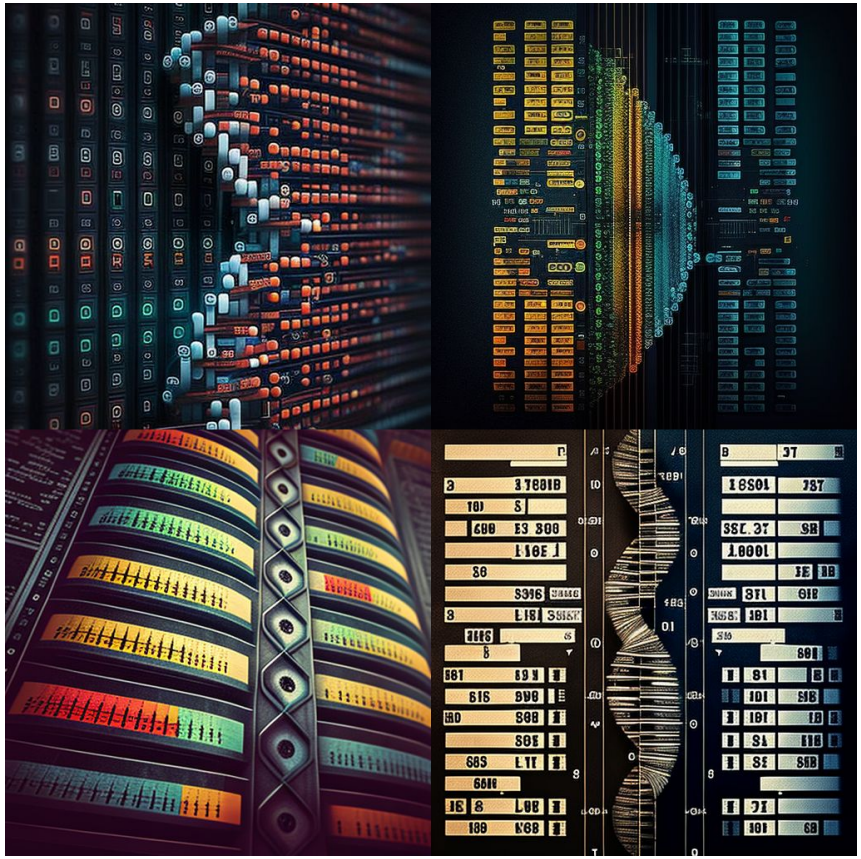
Parte 4a: Método de metagenómica “un-binned”

- Llamamos “genes” dentro de nuestros contigs utilizando programas como Prodigal que usan la tabla de código genético.

- Luego alineamos esos “genes” a databases de genes que ya conocemos, y adivinamos a base de homología / similitud qué son. (KRAKEN2, MG-Rast, SEED, BLASTn/BLASTp, KEGG)

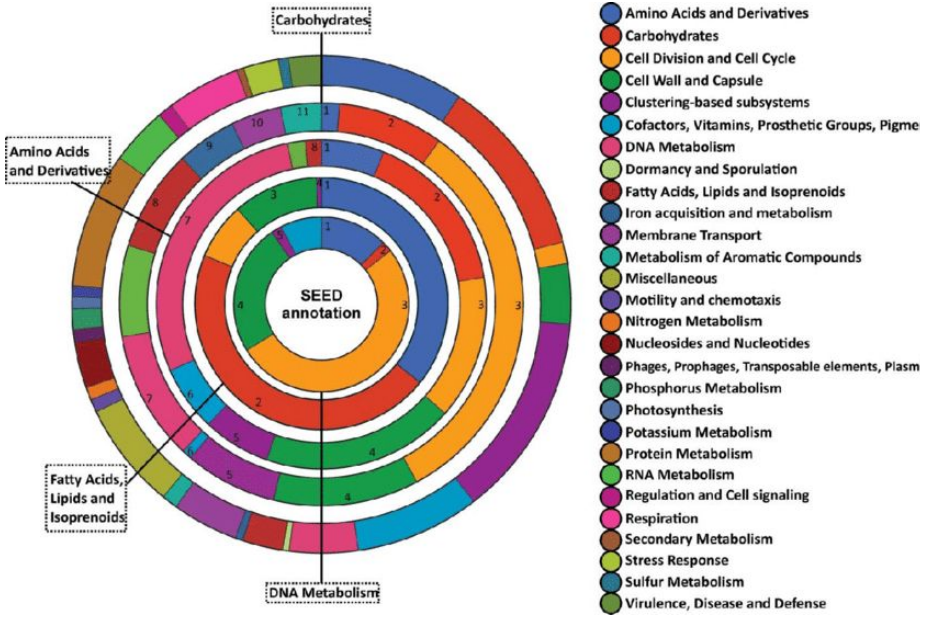
97% similar a maíz

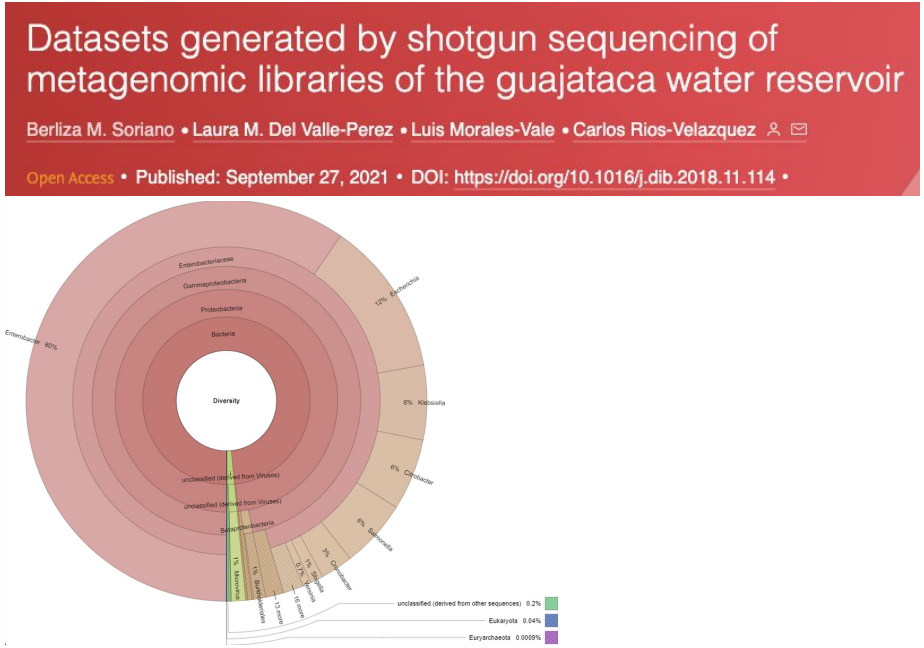
93% similar a carne guisada



MidJourney AI: “Matching a genetic sequence to a

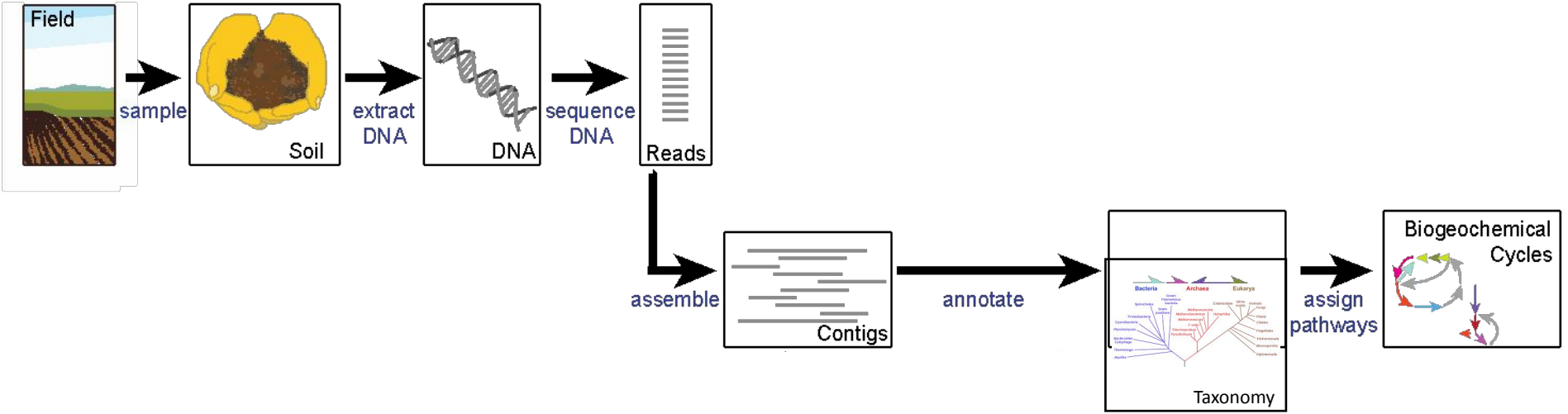
Resultados de **método “un-binned”**



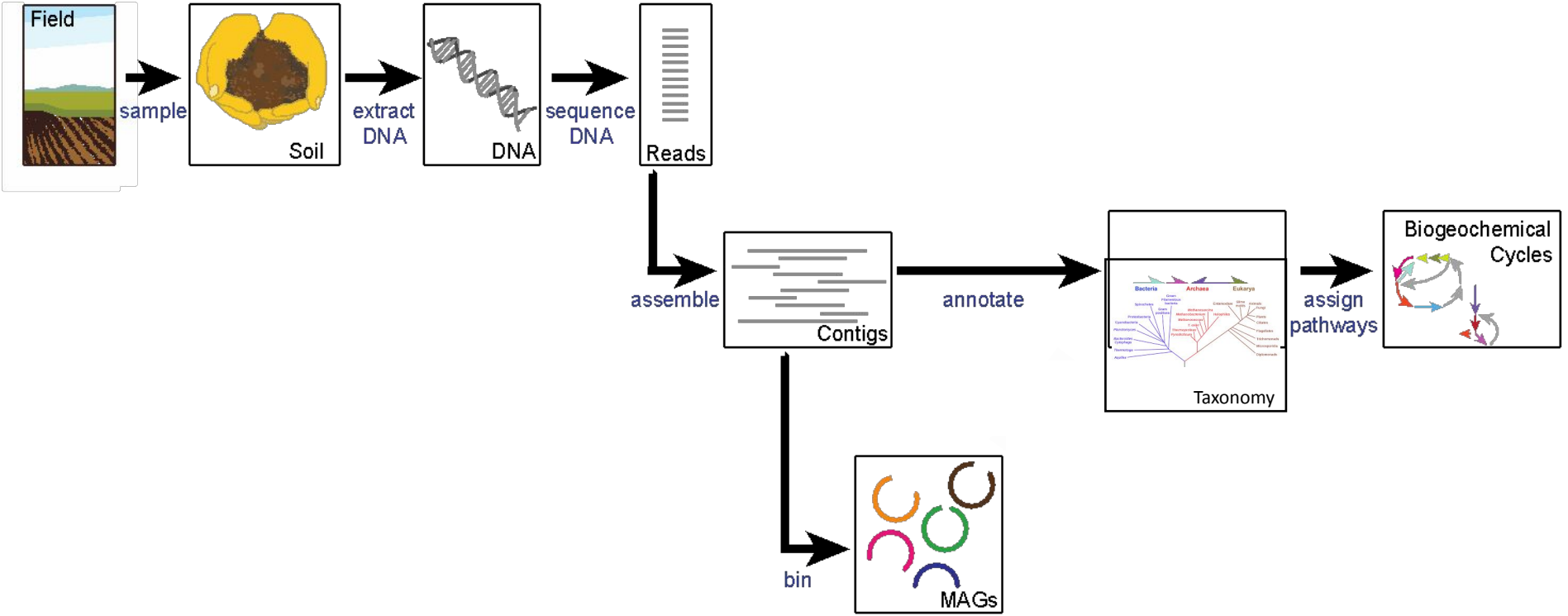


- Desglose de genes y procesos que se pueden encontrar en tu muestra.
  - Por ejemplo: desnitrificación, genes de fotosíntesis, genes para esporulación, etc.
- Desglose de qué organismos están presentes en tu muestra.

Parte 4a: Método “un-binned”



Parte 4b: Método “binned”

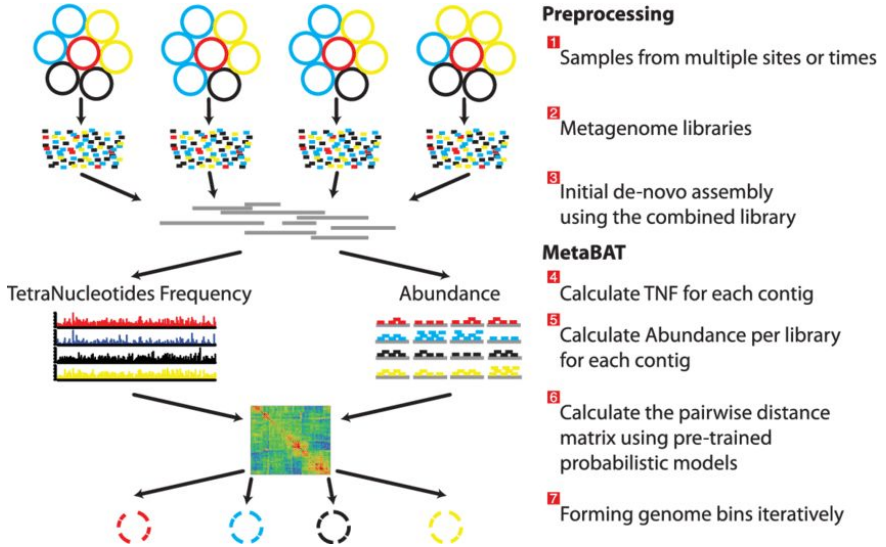


MAG: Metagenome Assembled Genome

Los “bins” o MAGs se forman cuando juntamos “contigs” que pertenecen al mismo organismo

- Contigs (>2,500 bp) son las partes del pato (el genoma)

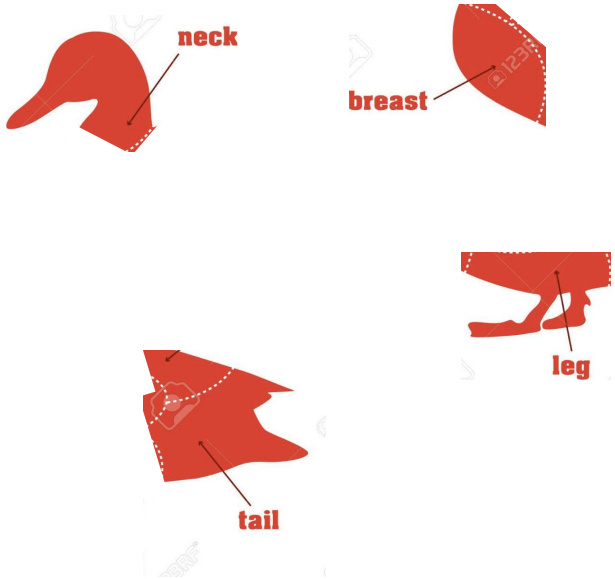
Haces un “mapeo” de reads metagenómicos a tus contigs o scaffolds y a base de cobertura y otras métricas.



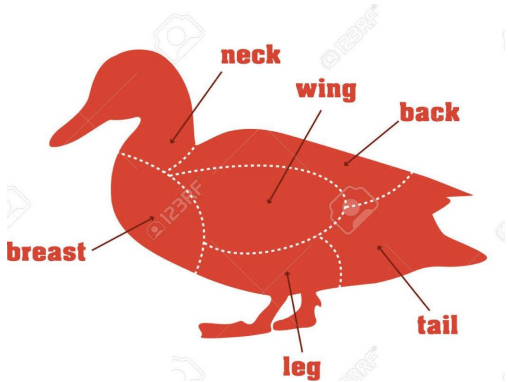
Pipeline MetaBAT

Los “bins” o MAGs se forman cuando juntamos “contigs” que pertenecen al mismo organismo

- Contigs (>2,000 bp) son las partes del pato (el genoma)

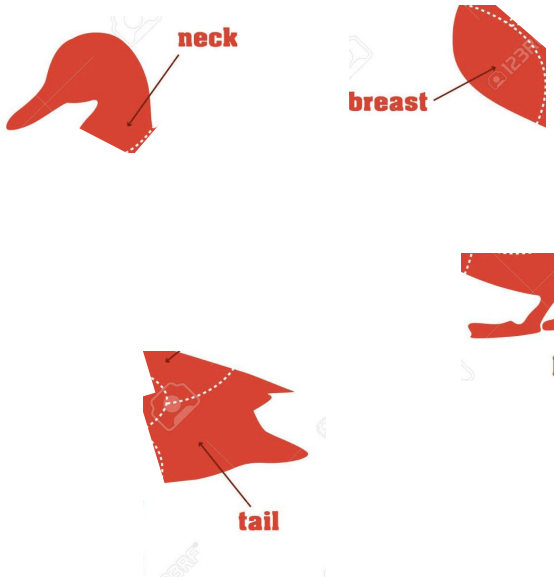


- Los contigs se agrupan y forman un genoma (el genoma representa el conjunto de ADN de un microbio)
- El programa más utilizado es “MetaBat2”



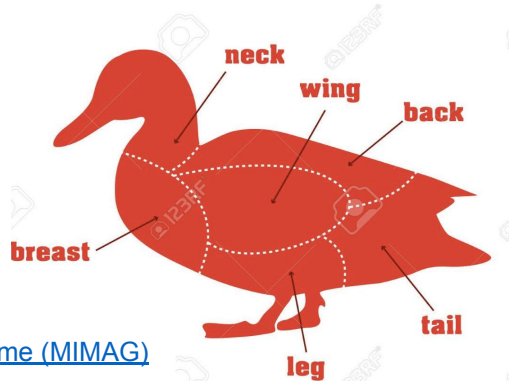
Los “bins” o MAGs se forman cuando juntamos “contigs” que pertenecen al mismo organismo

- Contigs (>2,000 bp) son las partes del pato (el genoma)



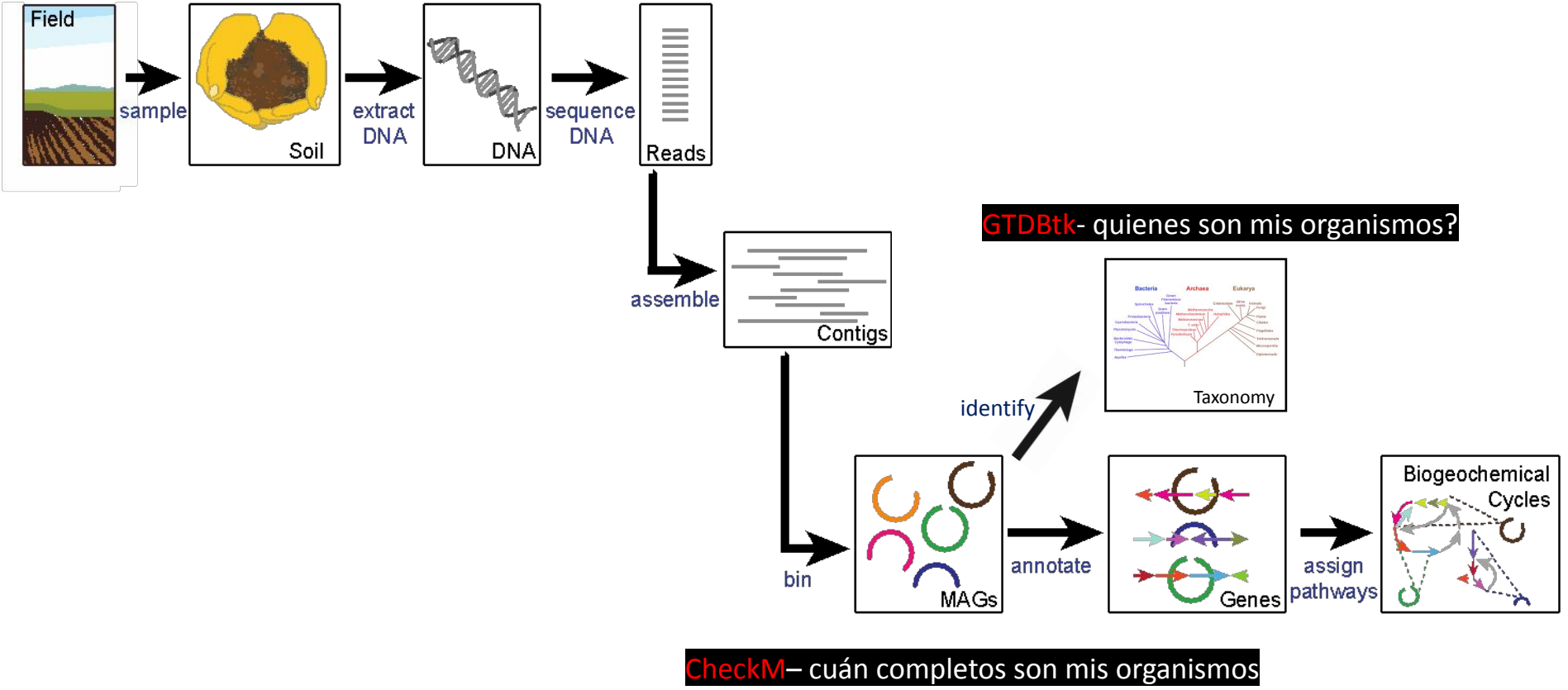
- Los contigs se agrupan y forman un genoma (el genoma representa el conjunto de ADN de un microbio)
- El programa más utilizado es “MetaBat2”

Importante: el ”binning” no es perfecto. Hay diferentes niveles de contaminación y totalidad.



Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG)

Parte 4b: Método “binned”



nmdc

National Microbiome
Data Collaborative

Workflows metagenómicos

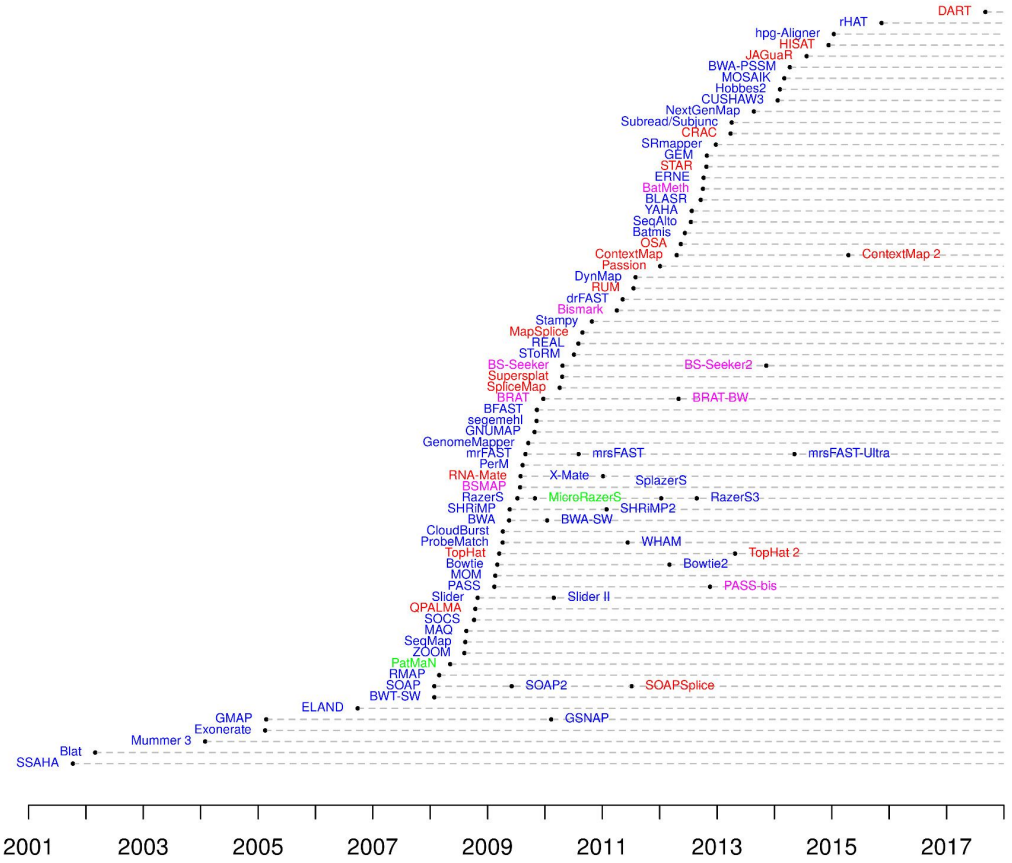
Explosión bioinformática

nmdc
National Microbiome Data Collaborative

La explosión masiva de herramientas bioinformáticas y sus workflows ha causado que la data sea procesada de diferentes maneras.

Esto limita nuestra habilidad de enlazar los diferentes estudios.

Esta gráfica incluye **SOLAMENTE** herramientas para “mapear” reads metagenómicos entre 2001-2018.



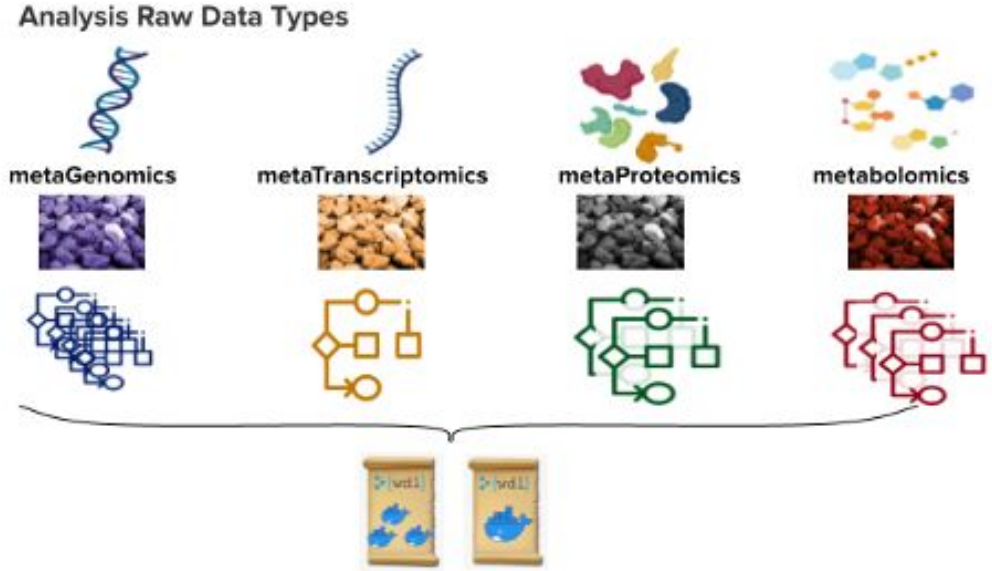
Workflows bioinformáticos NMDC

nmdc
National Microbiome Data Collaborative

El NMDC ha integrado herramientas bioinformáticas open-source en workflows estandarizados. Estos workflows pueden procesar muestras multi-omics “crudas” para producir resultados interoperables y funcionalmente anotados. Todo online, todo gratis!

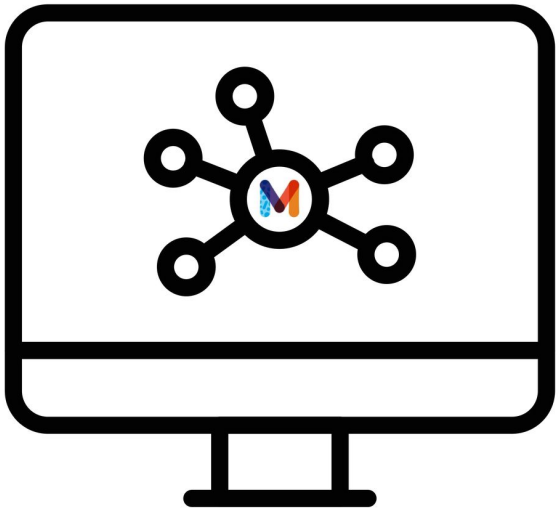
Workflows NMDC:

- Data metagenómica
  - QC de reads
  - Taxonomía a base de reads
  - Ensamblaje de reads
  - Anotación de metagenomas
  - Metagenome assembled genomes (MAGs)
- Data metatranscriptómica
- Data de materia natural orgánica
- Data metabolómica
- Data metaproteómica
- Virus y Plásmidos



Por qué usar workflows NMDC?

nmdc
National Microbiome Data Collaborative

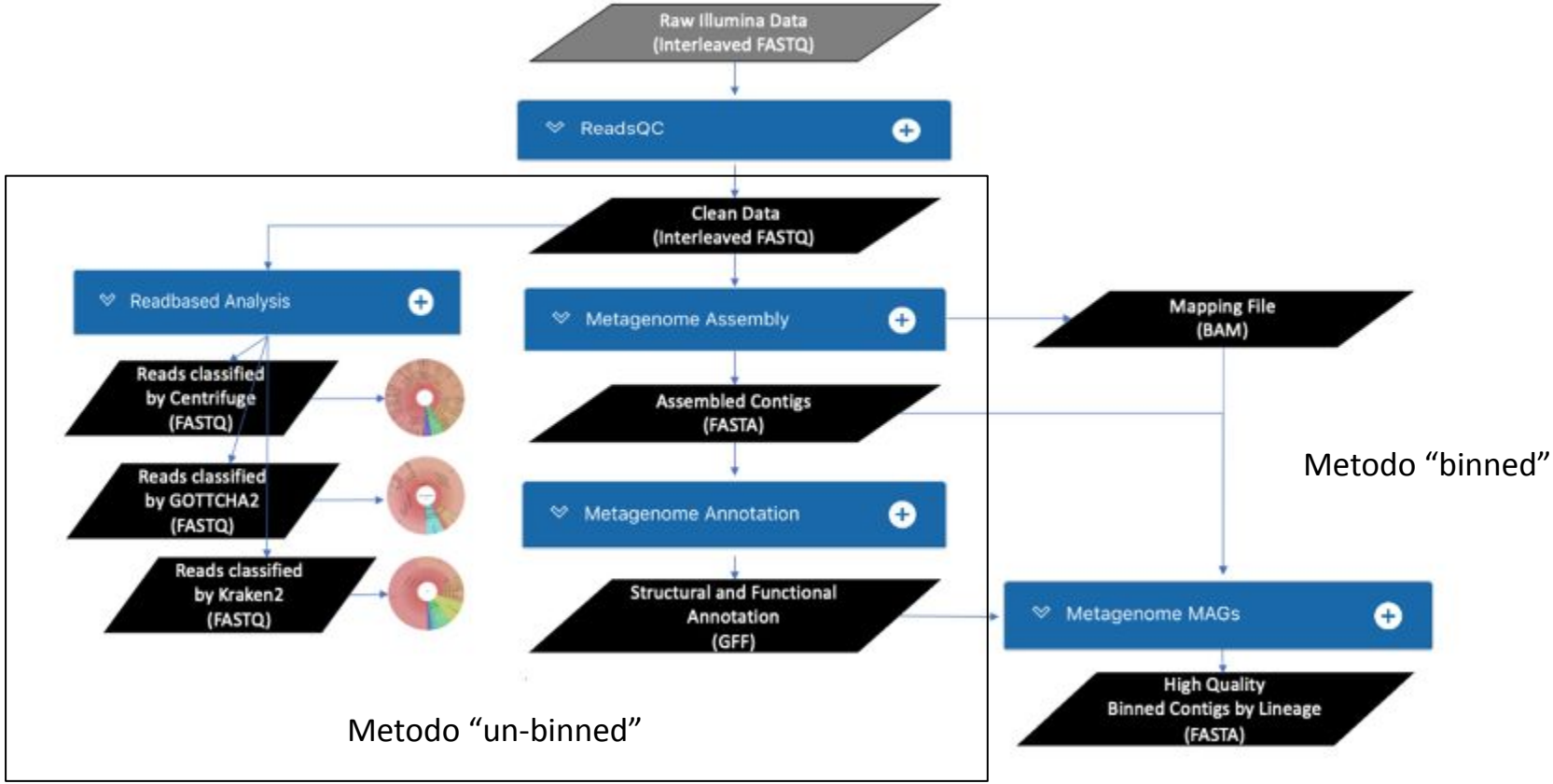


Beneficios de utilizar workflows NMDC:

- Las herramientas fueron extensamente estudiadas, seleccionadas y modificadas para desempeño óptimo
- Los workflows fueron probados en data recibida de docenas de instituciones y tipos de muestras
- Usuarios pueden correr todos estos workflows a través de recursos computacionales compartidos
  - Los usuarios NO necesitan descargar herramientas ni databases, ni tienen que tener acceso a recursos computacionales masivos.
- Los workflows están disponibles en un GUI sencillo, apto para usuarios de todo tipo de experiencia bioinformática.
- Open source

Workflow metagenómico NMDC

nmdc
National Microbiome Data Collaborative



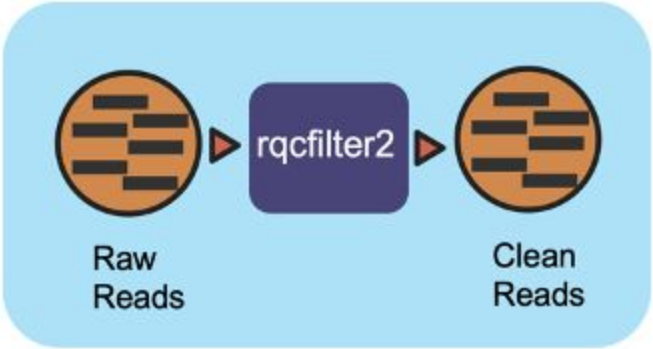
Reads QC

nmdc
National Microbiome Data Collaborative

rqcfilter2 aplica control de calidad (QC) a reads metagenómicos crudos para remover data de baja calidad y artefactos de secuenciación (e.g., reads asociados a contaminación, adapters de secuenciación).

Input: Data de reads cruda

Output: Archivo de reads “limpios” y estadísticas de QC.



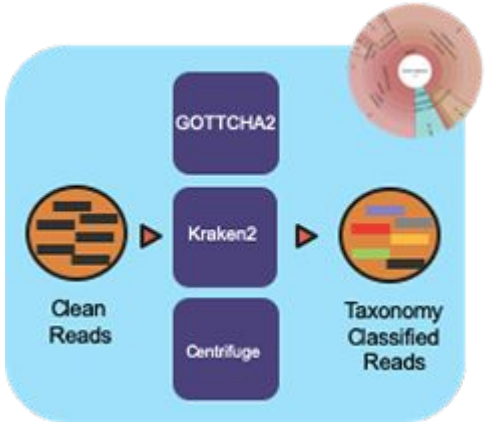
Clasificación taxonómica basada en reads

nmdc
National Microbiome Data Collaborative

Utiliza reads “limpios” metagenómicos y analiza el contenido taxonómico usando 3 herramientas diferentes que tienen distintos rangos de sensitividad y especificidad (GOTTCHA2, Kraken2, Centrifuge).

Input: Reads “limpios” del ReadsQC workflow

Output: Resultados para cada herramienta a distintos niveles taxonómicos (Epecie, Genero, y Familia). Genera plots Krona interactivos.



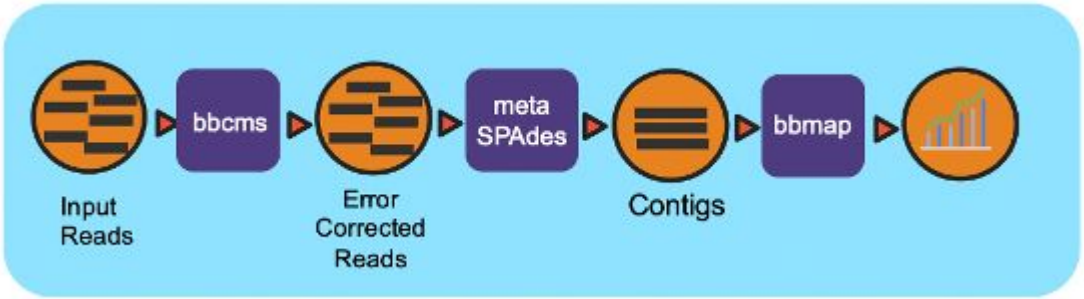
Ensamblaje de metagenoma

nmdc
National Microbiome Data Collaborative

Utiliza los reads metagenómicos limpios, hace una corrección de errores, los ensambla, y corre validación de ensamblaje.

Input: Reads “limpios” del ReadsQC workflow

Output: Contigs / Scaffolds ensamblados; estadísticas de ensamblaje



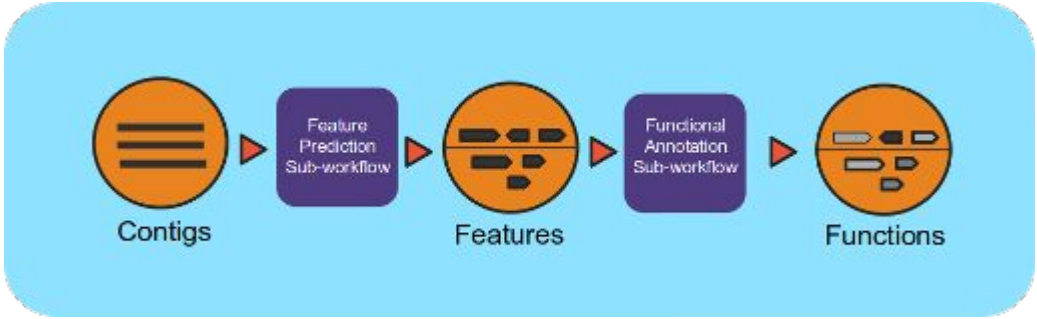
Anotación de metagenoma

nmdc
National Microbiome Data Collaborative

Utiliza los metagenomas generados por el último paso, y genera anotaciones funcionales y estructurales de genes presentes.

Input: Ensamblajes: recomendados del metagenome assembly workflow

Output: Anotación estructural, functional, y varios archivos resumiendo



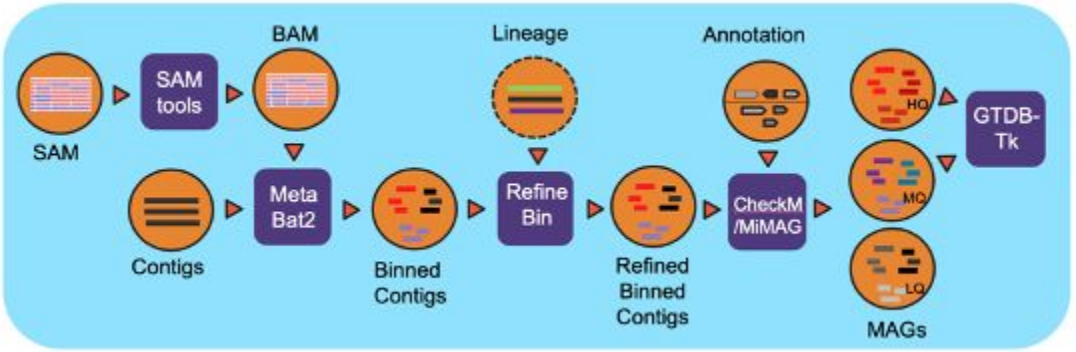
Metagenome Assembled Genomes

nmdc
National Microbiome Data Collaborative

Clasifica los contigs / scaffolds ensamblados en bins (MAGs). Luego son refinados utilizando el archivo de anotación funcional y evaluados para completación y contaminación. Da la calidad de bins y sus respectivos linajes.

Input: Contigs ensamblados, archivo de mapeo de reads proveniente del paso anterior, anotación funcional de los ensamblajes de la muestra anterior

Output: Summary statistics, file of high quality (HQ) and medium quality (MQ) bins



Workflow de virus y plásmidos

nmdc
National Microbiome Data Collaborative

geNomad es una herramienta que identifica genomas de virus y plásmidos a base de secuencias de nucleótidos. Provee clasificación rápida que se puede utilizar para encontrar elementos móviles genéticos en genomas, metagenomas o metatranscriptomas.

nmdc

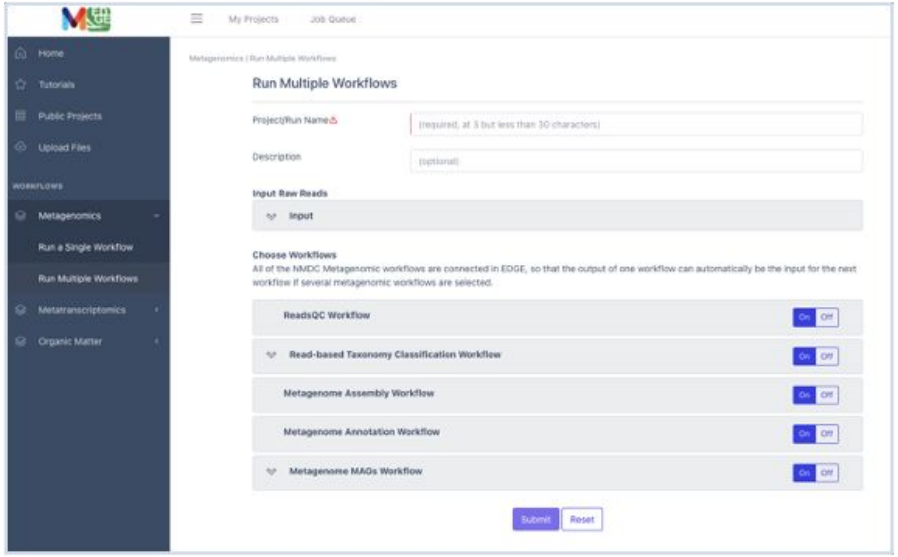
National Microbiome
Data Collaborative

NMDC EDGE

NMDC EDGE

Los workflows bioinformáticos se pueden correr en una interfase de red fácil de utilizar y gratuita llamada **NMDC EDGE**

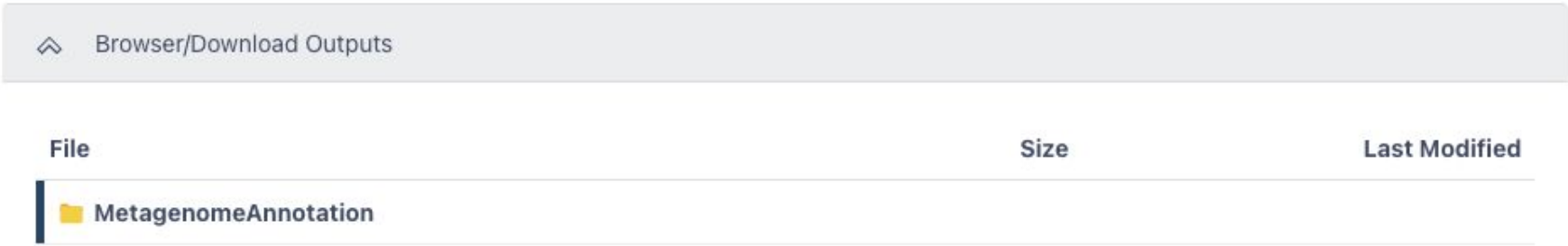
Diseñada para acomodar cualquier tipo de experiencia bioinformáticas (incluyendo novatos) https://nmdc-edge.org/home



Resumiendo grupos funcionales

nmdc
National Microbiome Data Collaborative

1. ProjectID.fna = nucleótidos de todos los genes encontrados
2. ProjectID.faa = amino ácidos de todos los genes encontrados
3. ProjectID_gene_phylogeny.tsv = taxonomía de genes encontrados
4. ProjectID_structural_annotation_stats.tsv = estadísticas de anotación
5. ProjectID_product_names.tsv = anotación de genes



Resumiendo grupos funcionales

nmdc
National Microbiome Data Collaborative

JOURNAL ARTICLE

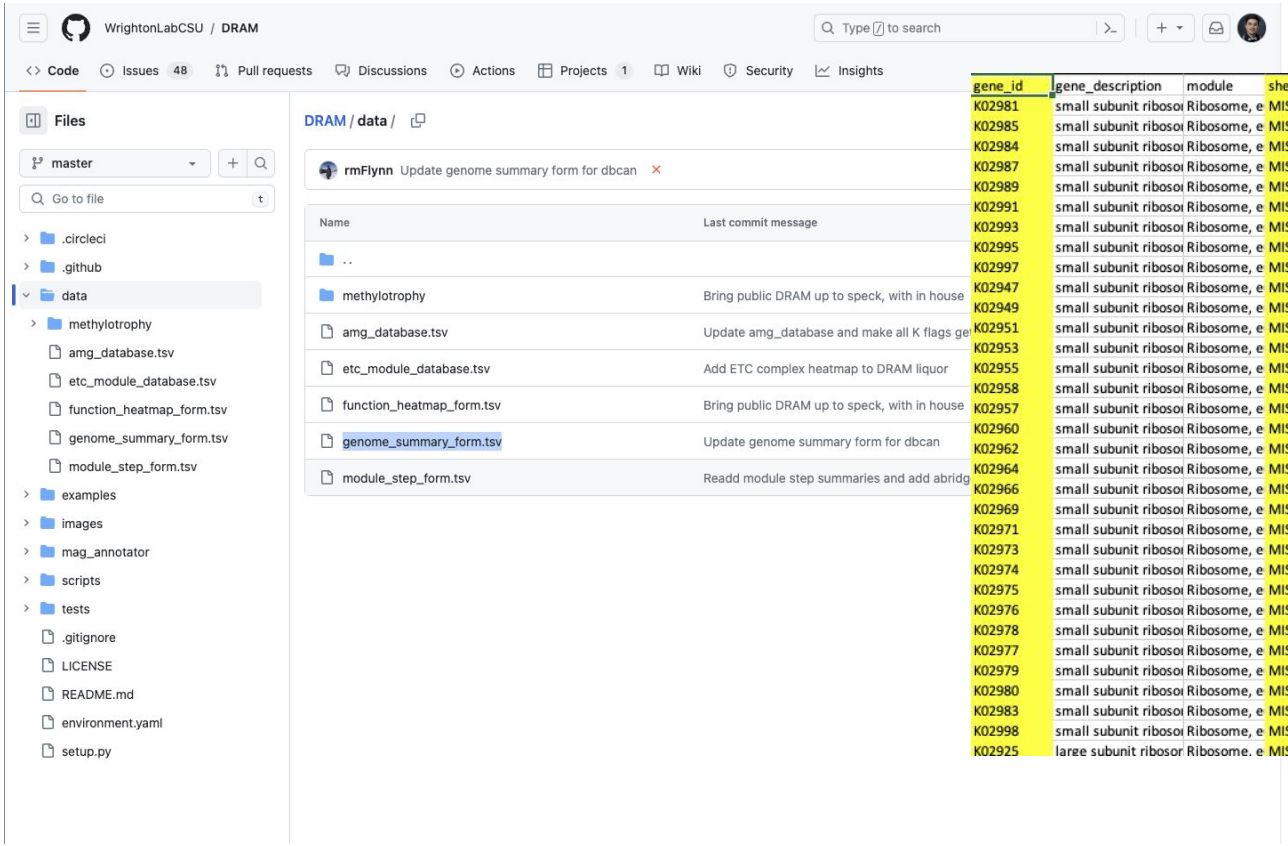
DRAM for distilling microbial metabolism to automate the curation of microbiome function

Michael Shaffer, Mikayla A Borton, Bridget B McGivern, Ahmed A Zayed, Sabina Leanti La Rosa, Lindsey M Solden, Pengfei Liu, Adrienne B Narrowe, Josué Rodríguez-Ramos, Benjamin Bolduc, M Consuelo Gazitúa, Rebecca A Daly, Garrett J Smith, Dean R Vik, Phil B Pope, Matthew B Sullivan, Simon Roux, Kelly C Wrighton

WrightonLabCSU/DRAM: Distilled and Refined Annotation of Metabolism: A tool for the annotation and curation of function for microbial and viral genomes (github.com)

Resumiendo grupos funcionales

nmdc
National Microbiome Data Collaborative



gene_id	gene_description	module	sheet	header	subheader	potential_amg
K02981	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02985	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02984	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02987	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02989	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02991	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02993	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02995	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02997	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02947	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02949	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02951	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02953	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02955	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02958	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02957	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02960	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02962	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02964	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02966	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02969	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02971	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02973	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02974	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02975	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02976	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02978	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02977	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02979	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02980	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02983	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02998	small subunit ribosor	Ribosome, e	MISC	Information systems		TRUE
K02925	large subunit ribosor	Ribosome, e	MISC	Information systems		TRUE

Resumiendo grupos funcionales

nmdc
National Microbiome Data Collaborative

module: “=VLOOKUP(C2,genome_summary_form.tsv!$A:$E,3,FALSE)”

sheet: “=VLOOKUP(C2,genome_summary_form.tsv!$A:$E,4,FALSE)”

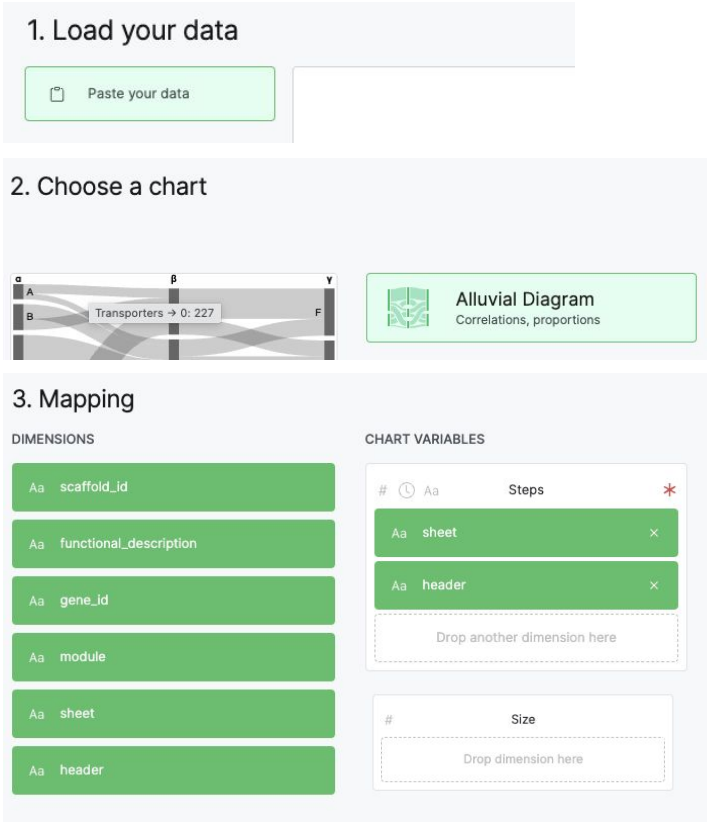
header: “=VLOOKUP(C2,genome_summary_form.tsv!$A:$E,5,FALSE)”

scaffold_id	functional_description	gene_id	module	sheet	header
Test_Run_GSP_scf_25_c1_130053_130616	2-aminoethylphosphonate transport system substrate-binding protein	K11081	2-Aminoethylphosphonate transport system	Transporters	0
Test_Run_GSP_scf_25_c1_130707_131066	2-aminoethylphosphonate transport system substrate-binding protein	K11081	2-Aminoethylphosphonate transport system	Transporters	0
Test_Run_GSP_scf_25_c1_131072_132181	2-aminoethylphosphonate transport system ATP-binding protein	K11084	2-Aminoethylphosphonate transport system	Transporters	0
Test_Run_GSP_scf_25_c1_132184_133044	2-aminoethylphosphonate transport system permease protein	K11083	2-Aminoethylphosphonate transport system	Transporters	0
Test_Run_GSP_scf_25_c1_133047_133844	2-aminoethylphosphonate transport system permease protein	K11082	2-Aminoethylphosphonate transport system	Transporters	0
Test_Run_GSP_scf_40_c1_93060_93275	3-methylfumaryl-CoA hydratase	K09709	3-Hydroxypropionate bi-cycle	Energy	C1
Test_Run_GSP_scf_26_c1_89139_90263	ABC-2 type transport system permease protein	K01992	ABC-2 type transport system	Transporters	0
Test_Run_GSP_scf_40_c1_46401_47507	ABC-2 type transport system permease protein	K01992	ABC-2 type transport system	Transporters	0
Test_Run_GSP_scf_40_c1_47623_48753	ABC-2 type transport system permease protein	K01992	ABC-2 type transport system	Transporters	0
Test_Run_GSP_scf_40_c1_48746_50482	ABC-2 type transport system ATP-binding protein	K01990	ABC-2 type transport system	Transporters	0
Test_Run_GSP_scf_50_c1_47833_48759	ABC-2 type transport system ATP-binding protein	K01990	ABC-2 type transport system	Transporters	0
Test_Run_GSP_scf_50_c1_48756_49526	ABC-2 type transport system permease protein	K01992	ABC-2 type transport system	Transporters	0
Test_Run_GSP_scf_178_c1_19_192	adenylosuccinate synthase	K01939	Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	MISC	Information systems
Test_Run_GSP_scf_193_c1_1_270	adenylosuccinate synthase	K01939	Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	MISC	Information systems
Test_Run_GSP_scf_25_c1_66712_67356	adenylate kinase	K00939	Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	MISC	Information systems
Test_Run_GSP_scf_37_c1_37374_38783	adhesin transport system outer membrane protein	K12543	Adhesin protein transport system	Transporters	0
Test_Run_GSP_scf_37_c1_38780_40960	ATP-binding cassette subfamily C protein LapB	K12541	Adhesin protein transport system	Transporters	0
Test_Run_GSP_scf_37_c1_40941_42131	adhesin transport system membrane fusion protein	K12542	Adhesin protein transport system	Transporters	0
Test_Run_GSP_scf_60_c1_81188_81994	AI-2 transport system ATP-binding protein	K10558	AI-2 transport system	Transporters	0
Test_Run_GSP_scf_60_c1_81937_82722	AI-2 transport system ATP-binding protein	K10558	AI-2 transport system	Transporters	0
Test_Run_GSP_scf_60_c1_82716_83759	AI-2 transport system permease protein	K10556	AI-2 transport system	Transporters	0
Test_Run_GSP_scf_60_c1_83756_84757	AI-2 transport system permease protein	K10557	AI-2 transport system	Transporters	0
Test_Run_GSP_scf_60_c1_84741_85808	AI-2 transport system substrate-binding protein	K10555	AI-2 transport system	Transporters	0
Test_Run_GSP_scf_117_c1_27780_29255	outer membrane protein	K12340	alpha-Hemolysin/cyclolysin transport system	Transporters	0
Test_Run_GSP_scf_127_c1_23657_25057	asparaginyl-tRNA synthetase	K01893	Aminoacyl-tRNA biosynthesis, eukaryotes	MISC	Information systems
Test_Run_GSP_scf_130_c1_25893_28280	phenylalanyl-tRNA synthetase beta chain	K01890	Aminoacyl-tRNA biosynthesis, eukaryotes	MISC	Information systems
Test_Run_GSP_scf_130_c1_28296_29291	phenylalanyl-tRNA synthetase alpha chain	K01889	Aminoacyl-tRNA biosynthesis, eukaryotes	MISC	Information systems
Test_Run_GSP_scf_202_c1_6980_9850	isoleucyl-tRNA synthetase	K01870	Aminoacyl-tRNA biosynthesis, eukaryotes	MISC	Information systems
Test_Run_GSP_scf_25_c1_12893_14278	cysteinyl-tRNA synthetase	K01883	Aminoacyl-tRNA biosynthesis, eukaryotes	MISC	Information systems

Añade un “row” y ponle esos ^ headers a las columnas. En columna “gene_id”, elimina la porción “KO:” para que sea más fácil.

Visualización: RAWGraphs 2.0

National Microbiome Data Collaborative



Añadir grupos funcionales:

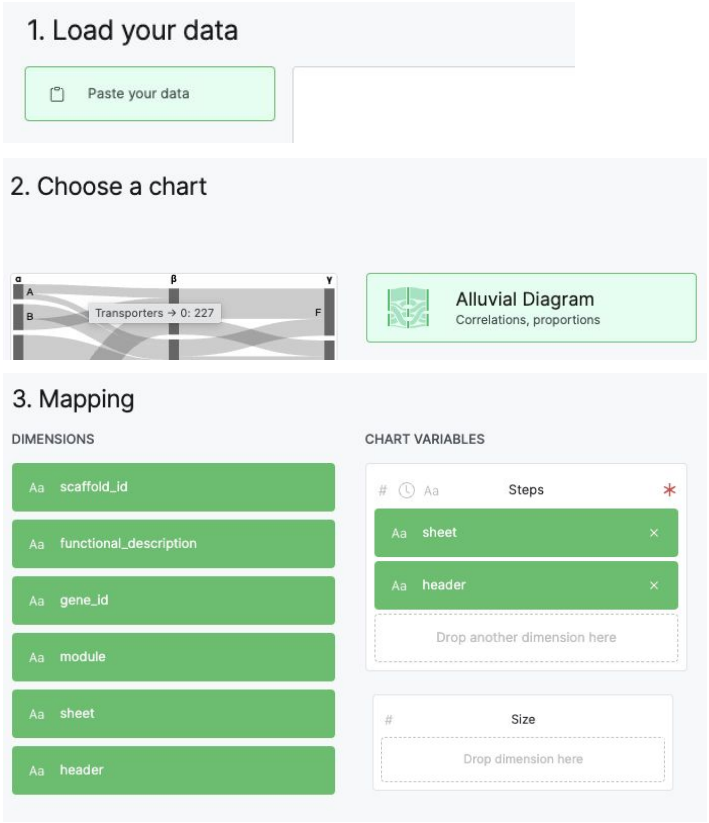
Grupos adicionales que se pueden añadir manualmente usando función de filtro en la columna “gene_description” de excel a todo lo que está en la categoría “otro”.

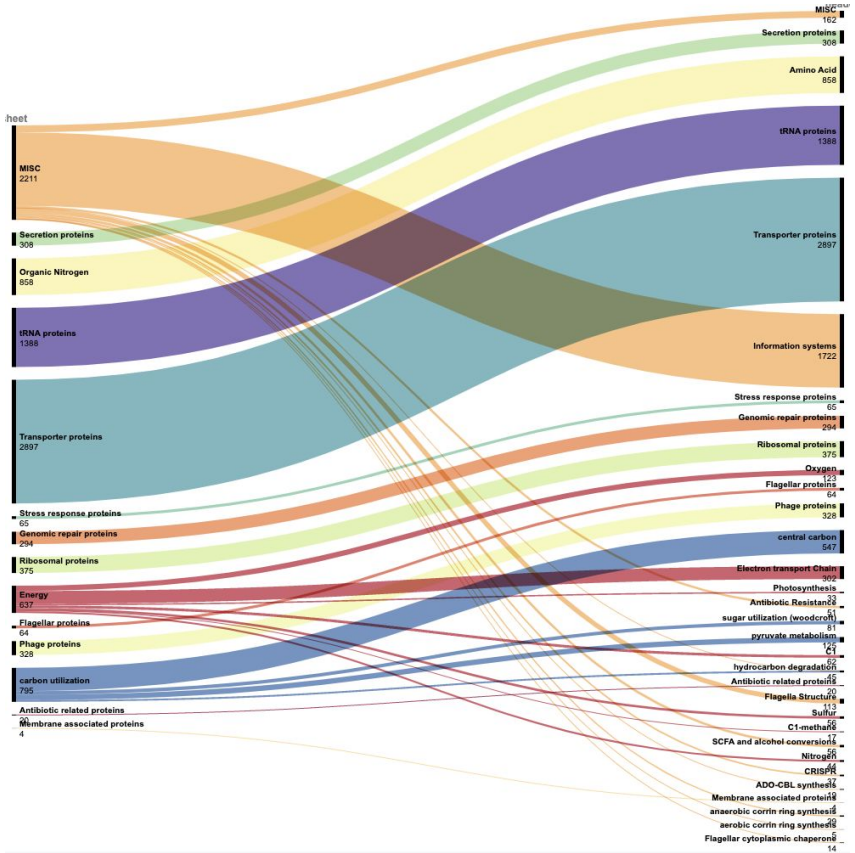
Para que la visualización se vea mejor, yo siempre quito los “hypothetical”, “uncharacterized”, y lo que sobre que no está en ninguna categoría (otro)

Nombre de grupo	**Término de búsqueda**
Hypothetical proteins	“Hypothetical” - remover
Uncharacterized proteins	“Uncharacterized” - remover
Other transport proteins	“transport”
Phage proteins	“phage” y “virus”
tRNA proteins	“tRNA”
Membrane associated proteins	“membrane”
Stress response proteins	“Stress”, y “heat shock”
Ribosomal proteins	“ribosomal”
Genomic repair proteins	“repair”
Antibiotic related proteins	“antibiotic”
Secretion proteins	“secretion”
Flagellar proteins	“flagell”
Pillus related proteins	“pill”

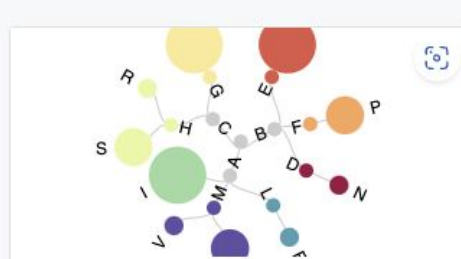
Visualización: RAWGraphs 2.0

National Microbiome Data Collaborative





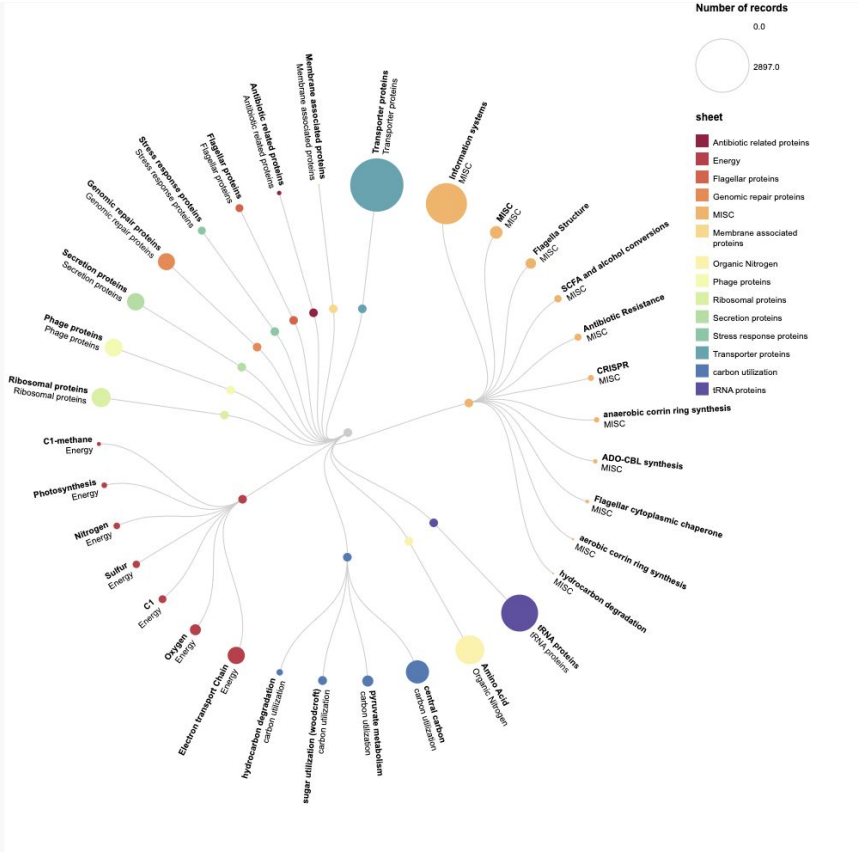
Visualización: RAWGraphs 2.0



Circular dendrogram

It displays hierarchically structured data with a radial tree structure, where the root node is in the center with the hierarchies moving outward. The area of nodes can be used to encode a further quantitative dimension and a quantitative or categorical dimension with color.

Code Tutorial



NMDC EDGE Beta Testing

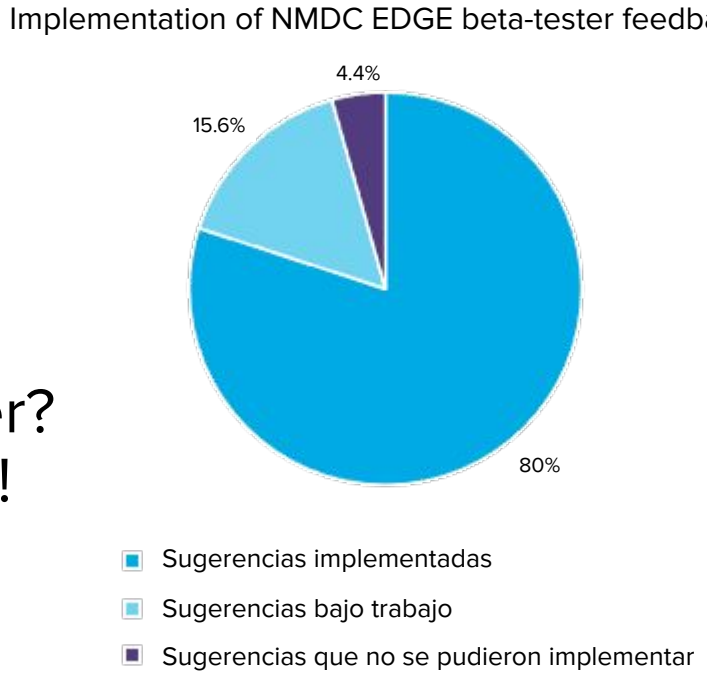
nmdc
National Microbiome Data Collaborative

Áreas para retroalimentación de usuarios para NMDC EDGE:

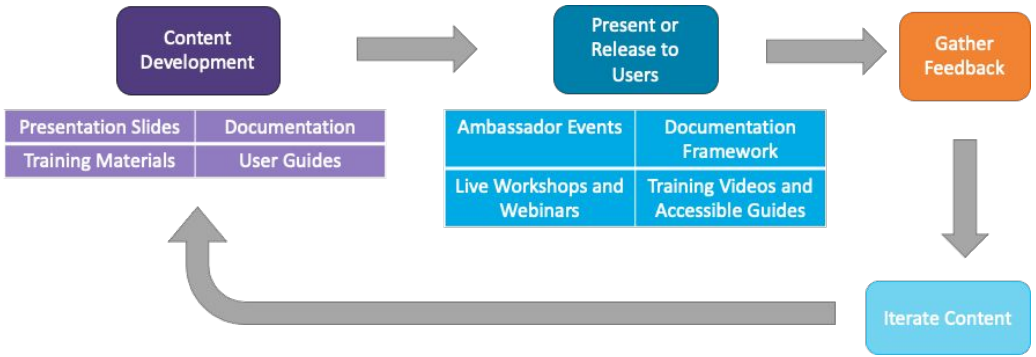
- Cómo correr workflows en NMDC EDGE
- Interfase de NMDC EDGE
- Materiales de entrenamiento para los NMDC workflow

Estás interesado en ser un beta-tester? Te damos crédito en nuestro website!

- Email nmdc-edge@lanl.gov



Diseño a base de la comunidad



Sus comentarios son una parte crítica del proceso de diseño

- Identificar las necesidades de la comunidad y evolucionar constantemente
- Trabajar con miembros de la comunidad a través de nuestro programa de “Champions” y “Ambassadors” que llevan a cabo workshops como este
- Investigación de los usuarios y cuán fácil es usar nuestras herramientas

Mecanismos de reportaje

- Formulario de “beta-tester”
  - Qué tipos de workflows corriste?
  - Información sobre tamaños de archivos, errores que surgieron, etc.
- Email: nmdc-edge@lanl.gov
  - Pueden recibir ayuda de troubleshooting con sus workflows de NMDC EDGE.
- Formulario de reportes general:
  - Tuviste algún error utilizando un producto NMDC?
  - Cualquier feedback de de herramientas
  - https://forms.gle/yxu9gkbufPigtbrB8

Recursos del NMDC

nmdc
National Microbiome Data Collaborative

Website: https://microbiomedata.org/
Data Portal: https://data.microbiomedata.org/
Submission Portal: https://data.microbiomedata.org/submission/home
NMDC EDGE: https://nmdc-edge.org/home
Github: https://github.com/microbiomedata
Docker Hub: https://hub.docker.com/u/microbiomedata
Documentation:
https://nmdc-documentation.readthedocs.io/en/latest/overview/nmdc_overview.html
YouTube: https://www.youtube.com/channel/UCyBqKc46NQZ_YgZlKGYeglw/featured

Sign up for our newsletter
microbiomedata.org

Find us on Twitter
@microbiomedata

Become a NMDC Champion
Champions Program - National Microbiome Data Collaborative

Find us on LinkedIn
(1) National Microbiome Data Collaborative: Overview | LinkedIn

Read more about the NMDC

frontiers

Hu B, Canon S, Eloe-Fadrosh EA, et al.. Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale. *Front Bioinform*. 1:826370. (2022) doi: 10.3389/fbinf.2021.826370.

Nucleic Acids Research

Eloe-Fadrosh EA *et al*. The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res*. 7;60(D1):D828–D836. (2022) doi: 10.1093/nar/gkab990.

nature REVIEWS MICROBIOLOGY

Wood-Charlson, E.M., Anubhav, Auberry, D. *et al.* The National Microbiome Data Collaborative: enabling microbiome science. *Nat Rev Microbiol* **18,** 313–314 (2020). doi.org/10.1038/s41579-020-0377-0

mSystems

Vangay, P *et al.* Microbiome metadata standards: Report of the National Microbiome Data Collaborative's workshop and follow-on activities. *mSystems* **6**, e01194-20 (2021). doi.org/10.1128/mSystems.01194-20