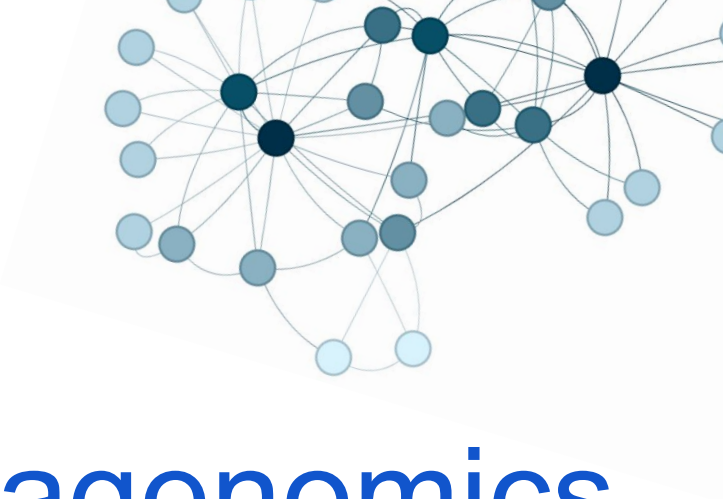


Introduction to Metagenomics

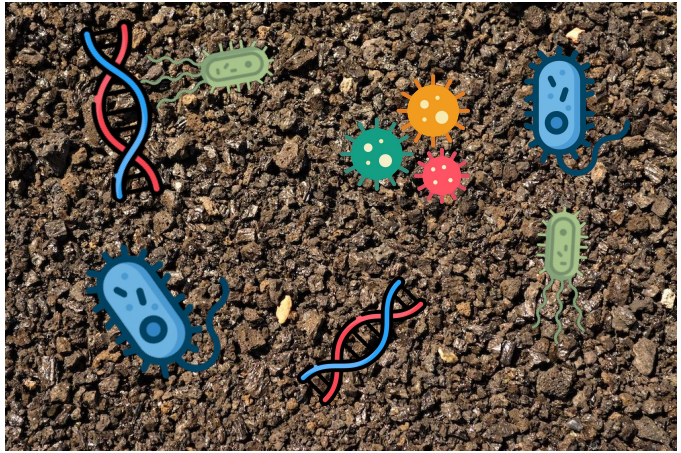
Dr Carla Greco
Basecamp Research

Aims:

- Quick overview of metagenomics
- Introduction to some concepts
  - Read-based analysis
  - Assembly
  - Annotation
- General Bioinformatics tips to help you get started!
- Practical: Kraken2

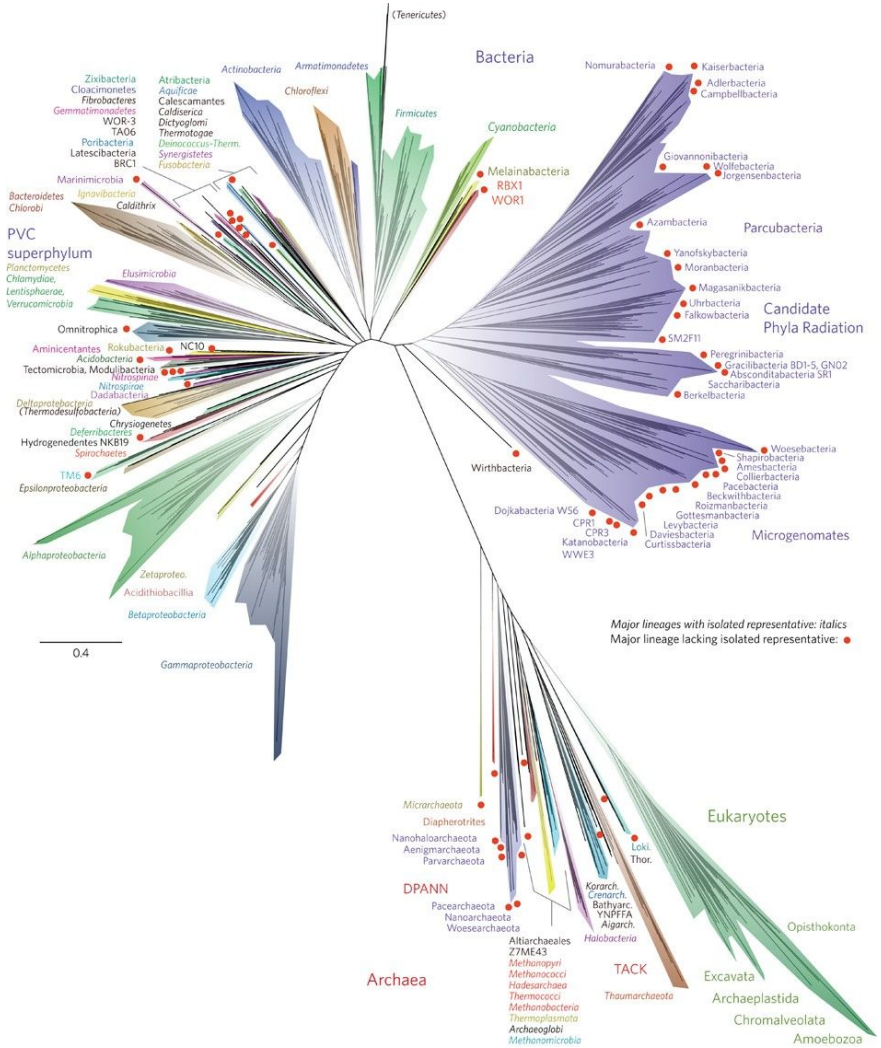
What is metagenomics?

- Coined by Jo Handelsman (2004)
- “Beyond the genome”
- Metagenomics is defined as the **direct genetic analysis** of genomes contained with an environmental sample

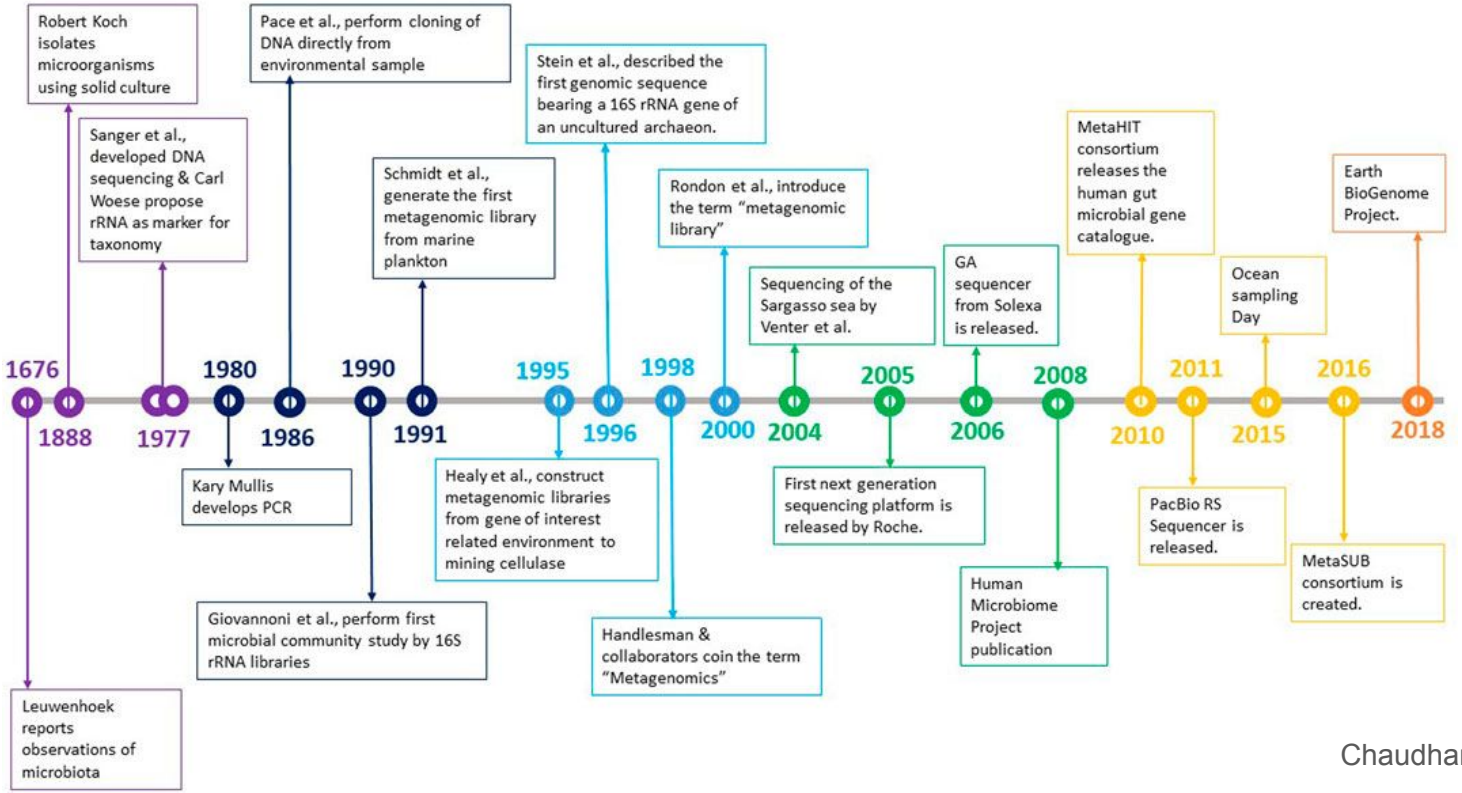


Why metagenomics?

- Historically research focused on cultures
- Less than 1% of microbial diversity
- The ‘great plate count anomaly’
- Hug et al. New tree of life showcasing CPR



Quick History & Landmark papers



Chaudhari et al 2023

Quick History & Landmark papers

Article | Published: 01 February 2004

Community structure and metabolism through reconstruction of microbial genomes from the environment

Gene W. Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar & Jillian F. Banfield

Nature **428**, 37–43 (2004) | Cite this article

22k Accesses | **1628** Citations | **83** Altmetric | Metrics

Article | Open access | Published: 11 September 2017

Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz & Gene W. Tyson

Nature Microbiology **2**, 1533–1542 (2017) | Cite this article

79k Accesses | **1050** Citations | **496** Altmetric | Metrics

Article

Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean

Tom O. Delmont,[1,2,9,*] Morgan Gaia,[1,2] Damien D. Hinsinger,[1,2] Paul Frémont,[1,2] Chiara Vanni,[3] Antonio Fernandez-Guerra,[4] A. Murat Eren,[5] Artem Kourlaiev,[1,2] Leo d'Agata,[1,2] Quentin Clayssen,[1,2] Em Karine Labadie,[1,2] Corinne Cruaud,[1,2] Julie Poulain,[1,2] Corinne Da Silva,[1,2] Marc Wessner,[1,2] Benjamin N Jean-Marc Aury,[1,2] Tara Oceans Coordinators, Colomban de Vargas,[2,6] Chris Bowler,[2,7] Eric Karsenti,[2,6,8] Patrick Wincker,[1,2] and Olivier Jaillon[1,2]

Article | Open access | Published: 11 October 2023

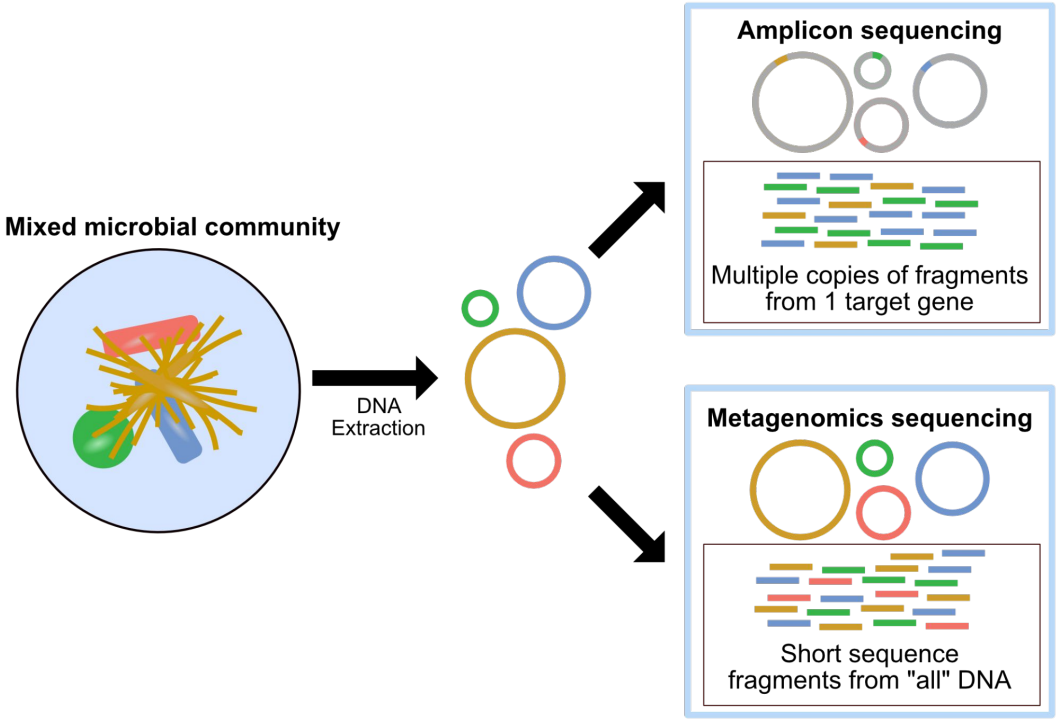
Unraveling the functional dark matter through global metagenomics

Georgios A. Pavlopoulos, Fotis A. Baltoumas, Sirui Liu, Oguz Selvitopi, Antonio Pedro Camargo, Stephen Nayfach, Ariful Azad, Simon Roux, Lee Call, Natalia N. Ivanova, I. Min Chen, David Paez-Espino, Evangelos Karatzas, Novel Metagenome Protein Families Consortium, Ioannis Iliopoulos, Konstantinos Konstantinidis, James M. Tiedje, Jennifer Pett-Ridge, David Baker, Axel Visel, Christos A. Ouzounis, Sergey Ovchinnikov, Aydin Buluç & Nikos C. Kyrpides

Nature **622**, 594–602 (2023) | Cite this article

Differences between WGS metagenomics and Amplicons

- ‘Shotgun metagenomics’ is un-targeted sequencing of DNA fragments
- Amplicons are specific regions (e.g. 16S rRNA gene) amplified through PCR

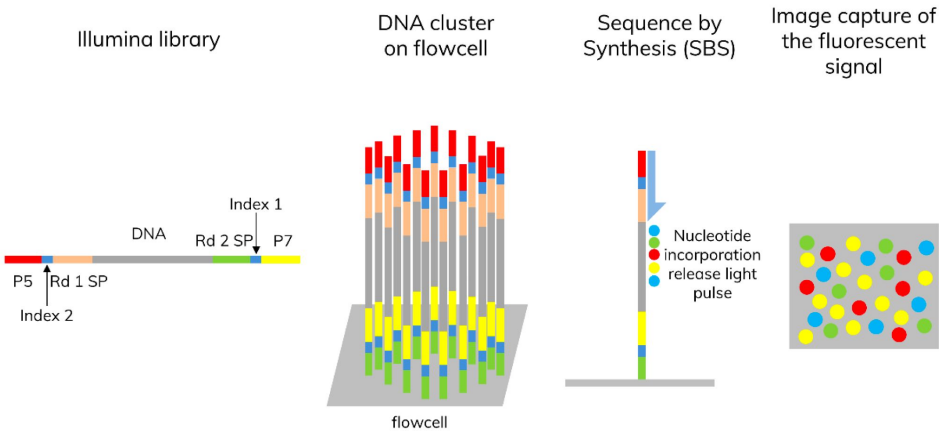


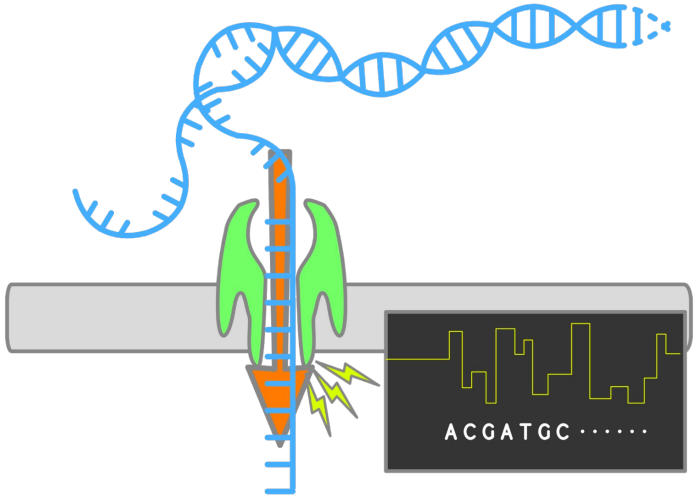
https://astrobiomike.github.io/

Is metagenomics for you?

- ✅ Want to assess functional potential
- ✅ Want to identify entire genes
- ✅ Want to identify species/strains at higher resolution
- ✅ Want to assemble genomes
- ✅ Have some money to throw at it!

Sequencing strategies





Accurate short-read: Illumina

- High accuracy
- Short read length (~300 bp) has limitations

Noisy long-read: ONT

- Low accuracy
- Extremely long read-length (>100 kb)

Read-based analysis

- Using un-assembled reads directly
- Used for quantitative analysis
- Mapping based methods:
  - Aligning reads against references
- Kmer-based methods
- Diamond BLAST

✅ Pros and ❌ Cons

Pro:

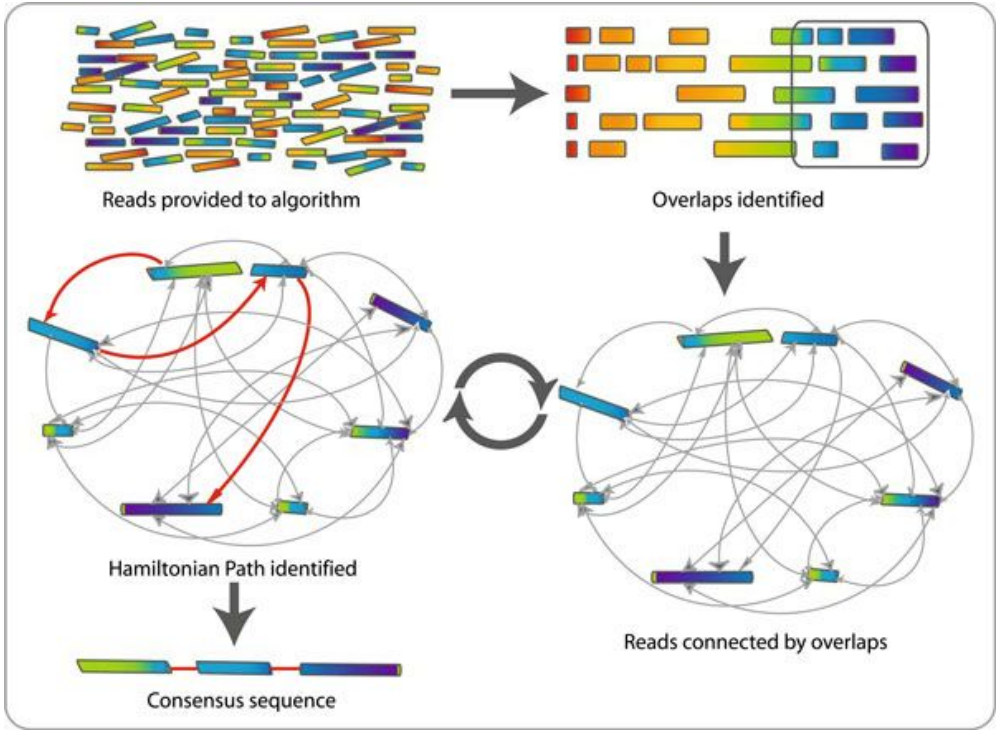
- Computationally quick and easy
- Remove assembly biases
- Allow comparison of metagenomes

Cons:

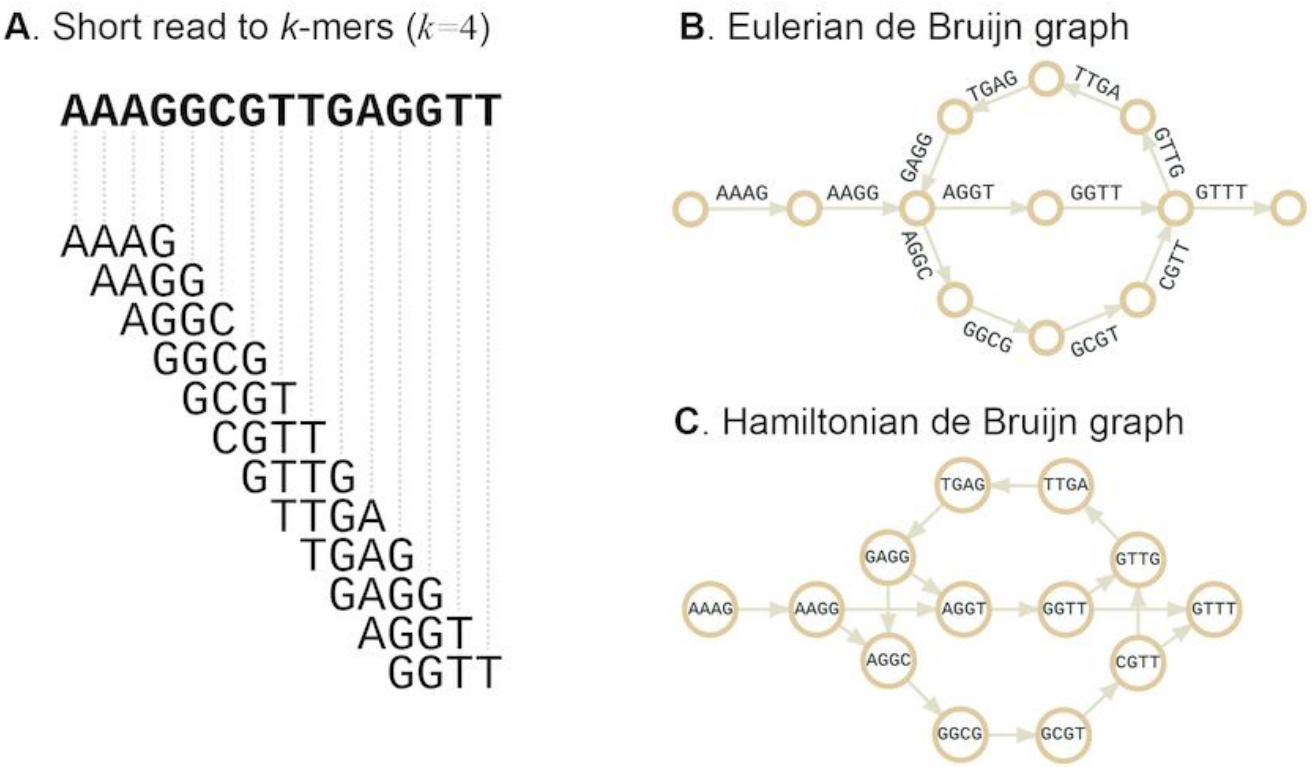
- Database dependant
- Low context - neighbouring genes
- Fragments of genes means often difficult to annotate

De novo Genome assembly: OLC

- From reads to ‘contigs’
- Long reads
- Solving puzzle to move reads to long contiguous sequences
- Different types of assembly algorithms e.g Overlap-Layout-Consensus or de Bruijn
- Used for different cases



De novo Genome assembly: de Bruijn graph

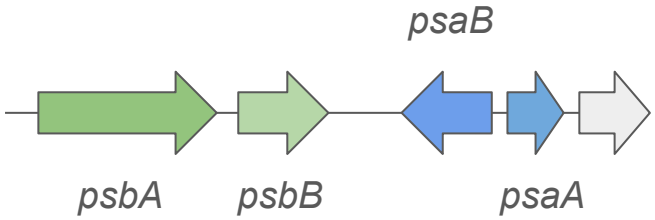


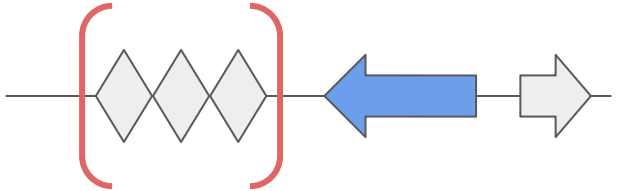
Metagenome assembly

- Multiple similar strains
- Differrening abundance of organisms
- Rare organisms
- Large datasets
- Co-assembly: assembling multiple samples together in a single assembly

Predicting genes and annotating them

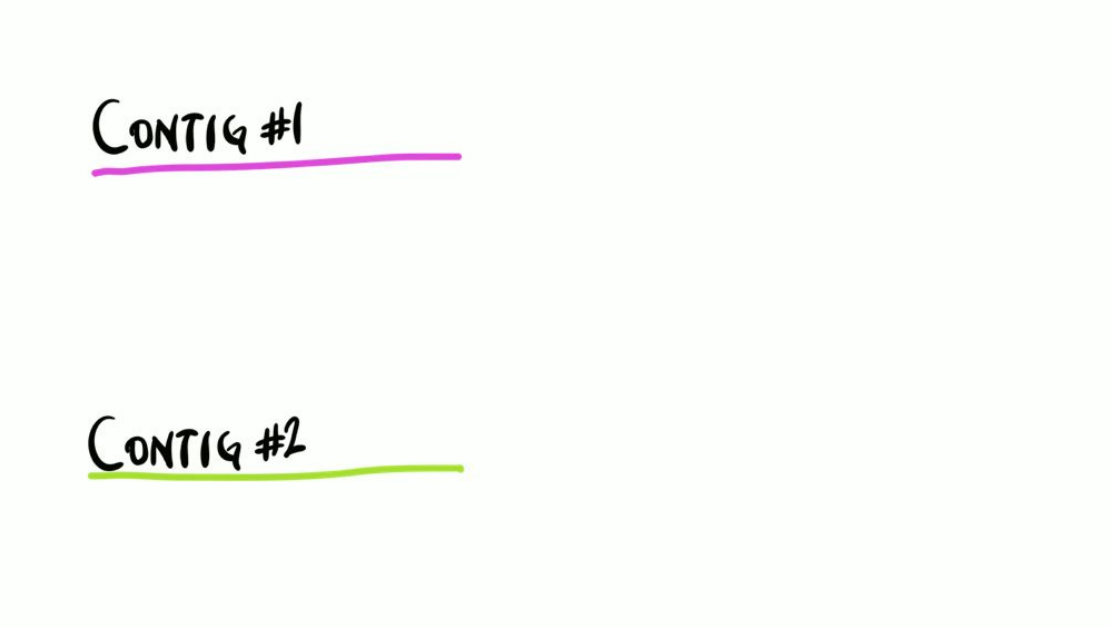
- Identifying genomics of genomic DNA that encode for genes
- Many algorithms to do this
  - Prodigal, Glimmer, MetaGeneMark
- Can also include annotating non-coding genes and other functional elements
- Annotate the genes - assigning function based on homology





CRISPR Cas systems

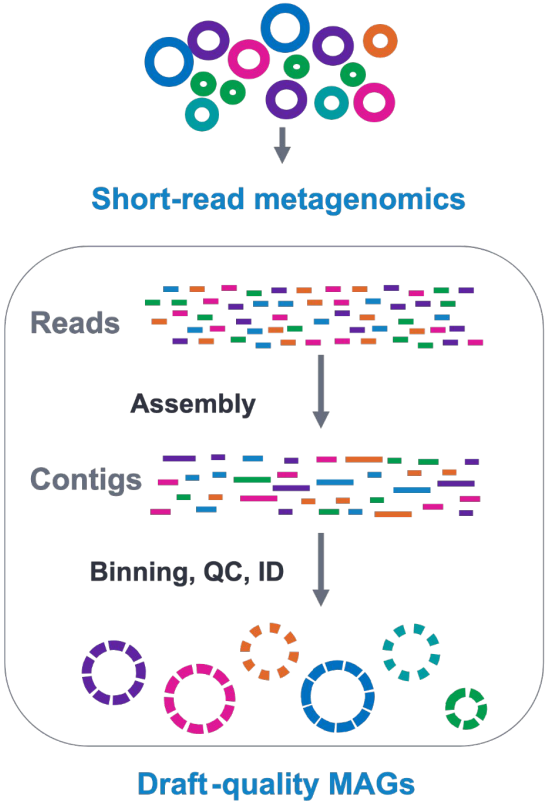
Read recruitment / read mapping



- Estimate abundance in a population
-

anvio.org/vocabulary/

MAGs: Metagenome assembled genomes

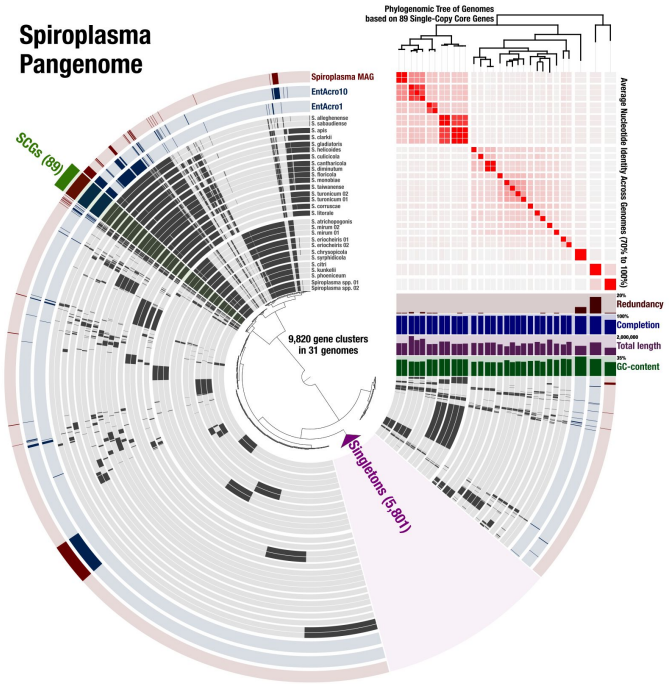


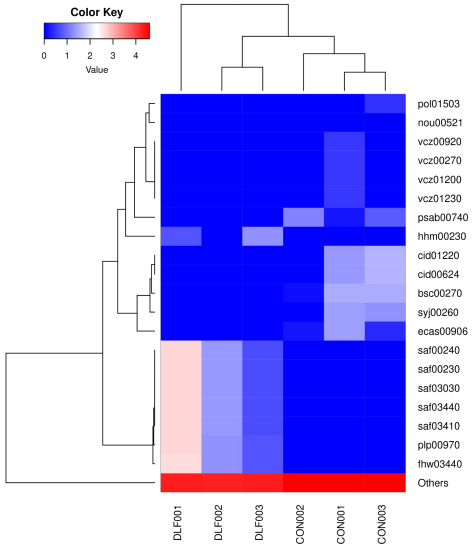
- Grouping contigs together into populations
- Based on characteristics on contigs:
  - Tetranucleotide frequency
  - Abundance/Coverage
  - Marker genes
- Movement towards ‘strain-resolved’ metagenomics

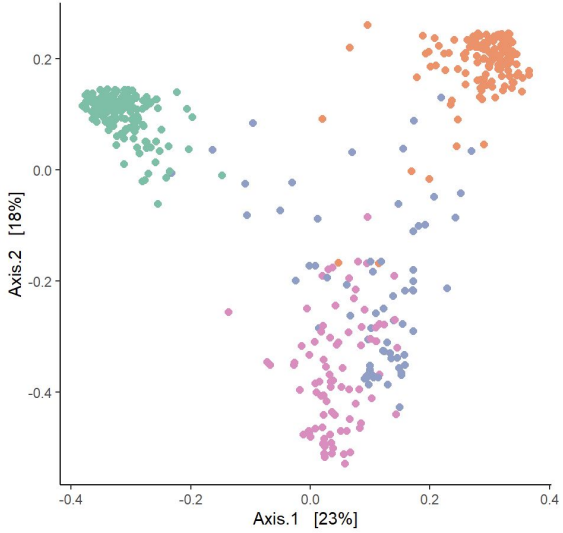
USCGs: Universal Single Copy Genes

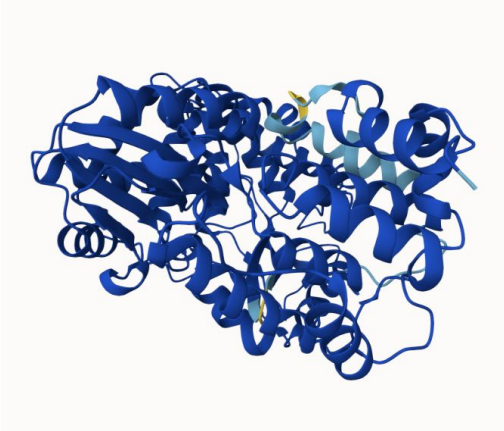
- Genes that occur in majority of genomes
- But occurs only one per genome
- Examples:
  - Ribosomal proteins
- Important in estimating completeness and contamination of an assembly
- Used for phylogenetic inference

What’s next: Analysis metagenomic data



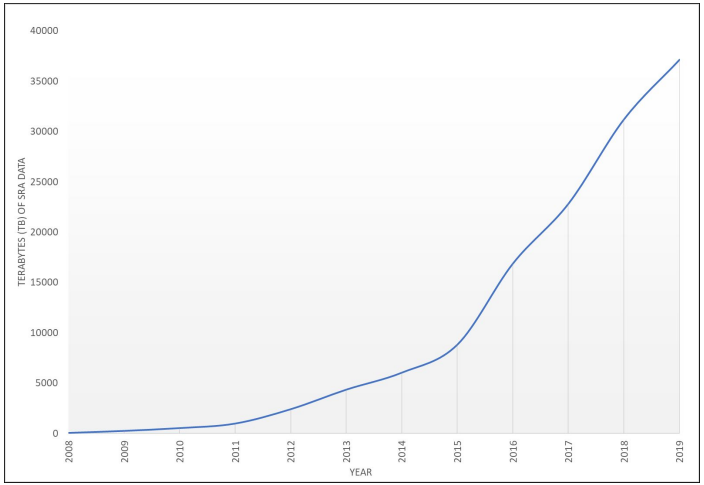






Big data: big problems

- Cheaper sequencing = easier to generate large amounts of data
- Exponential growth of SRA
- Analysis can require large amounts of compute and memory
- Movement towards fast and memory efficient tools
- Tools to use large data



Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity

Rayan Chikhi, Brice Raffestin, Anton Korobeynikov, Robert Edgar, Artem Babaian

doi: https://doi.org/10.1101/2024.07.30.605881

This article is a preprint and has not been certified by peer review [what does this mean?].

sourmash-bio/
branchwater

Searching large collections of sequencing data with genome-scale queries

3 Contributors | 21 Issues | 7 Stars | 2 Forks

General Bioinformatics Resources

Pipeline building

- Automating multi-step analyses
- Automated handling of resources
- Tracking of samples and runs
- Avoids a lot of copy-paste
- “What command did I run again?”

snakemake

nextflow

Installing tools

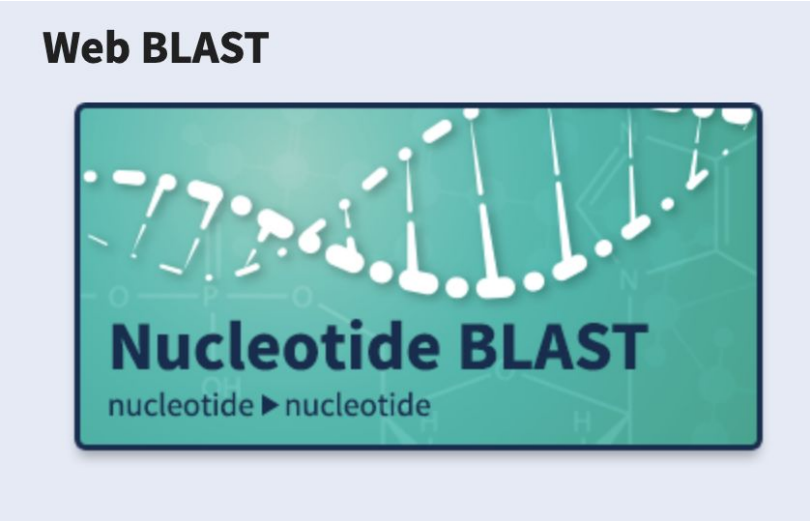
Package management systems: Automate installation of tools and dependencies

Container systems: Isolated environments for applications

Importance in Bioinformatics:

- Reproducibility: Ensures consistent results.
- Efficiency: Simplifies complex tool installation.
- Compatibility: Prevents software conflicts.
- Version Control: Maintains specific software versions.

Don’t want to code?



Useful resources for bioinformatics beginners:

https://astrobiomike.github.io/

https://anvio.org/learn/

https://bioinformaticsworkbook.org/

https://carpentries-lab.github.io/

[quick break - see you in 10!]

Practical session: Classifying reads with kraken2

Short Report | Open access | Published: 28 November 2019

Improved metagenomic analysis with Kraken 2

Derrick E. Wood, Jennifer Lu & Ben Langmead

Genome Biology **20**, Article number: 257 (2019) | Cite this article

A Protocol for this article was published on 28 September 2022

Abstract

Although Kraken's *k*-mer-based approach provides a fast taxonomic classification of metagenomic sequence data, its large memory requirements can be limiting for some applications. Kraken 2 improves upon Kraken 1 by reducing memory usage by 85%, allowing greater amounts of reference genomic data to be used, while maintaining high accuracy and increasing speed fivefold. Kraken 2 also introduces a translated search mode, providing increased sensitivity in viral metagenomics analysis.

Codesandbox:

https://codesandbox.io/p/live/0b5e8628-1622-499d-8395-8eba09f286fa

Other option: Docker file at

https://github.com/carlagreco/metagenomics-tutorial-2024