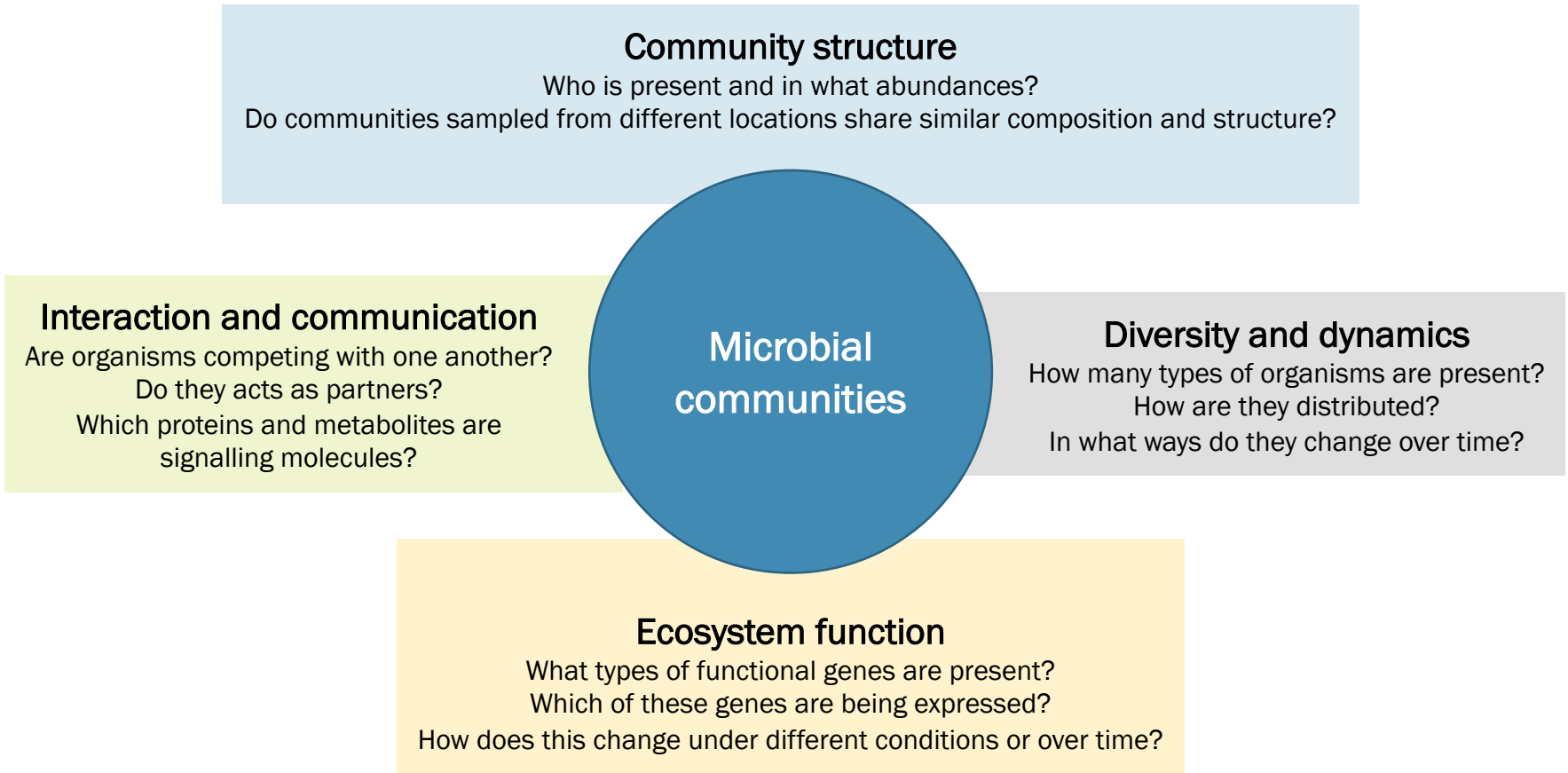
Environmental metagenomics

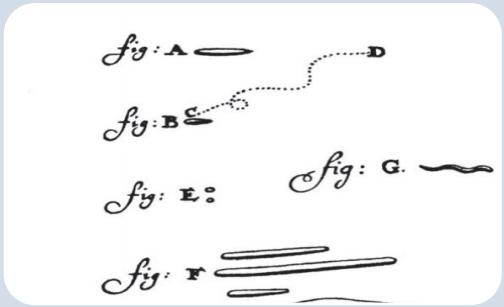
Introduction to metagenomics

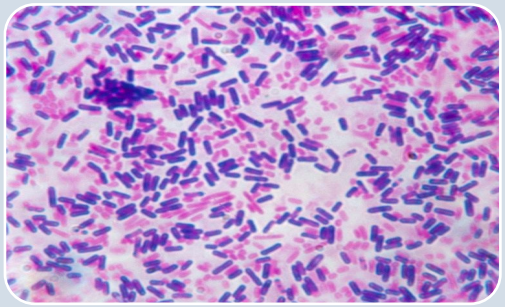
Igor S. Pessi & Antti Karkman, University of Helsinki

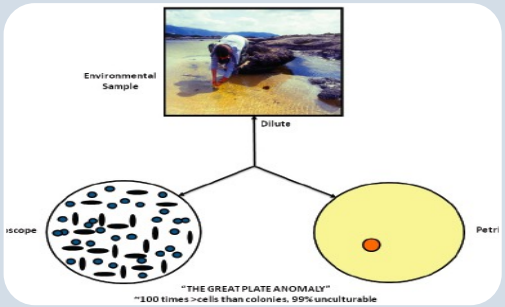
Metagenomics is the ultimate way to study microbial communities



Microbiology and technology go hand in hand







1670's

First observation of microbes under the microscope

1880's

Development of the Gram staining method

First isolation of a bacterium in solid media

1980's

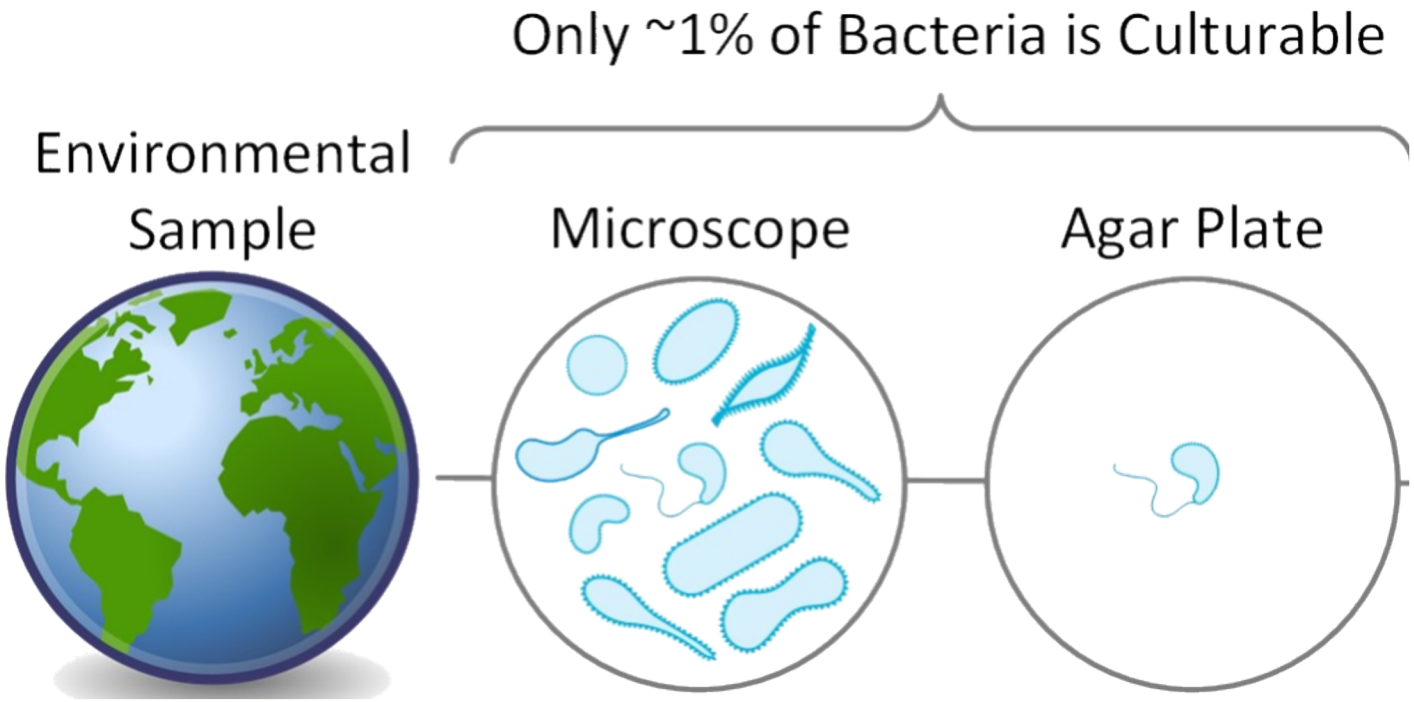
The Great Plate Count Anomaly

First culture-independent studies

2000's

Advent of high-throughput sequencing

The Great Plate Count Anomaly



Ann. Rev. Microbiol. 1985. 39:321–46

MEASUREMENT OF IN SITU ACTIVITIES OF NONPHOTOSYNTHETIC MICROORGANISMS IN AQUATIC AND TERRESTRIAL HABITATS

James T. Staley

Department of Microbiology and Immunology, University of Washington, Seattle, Washington 98195

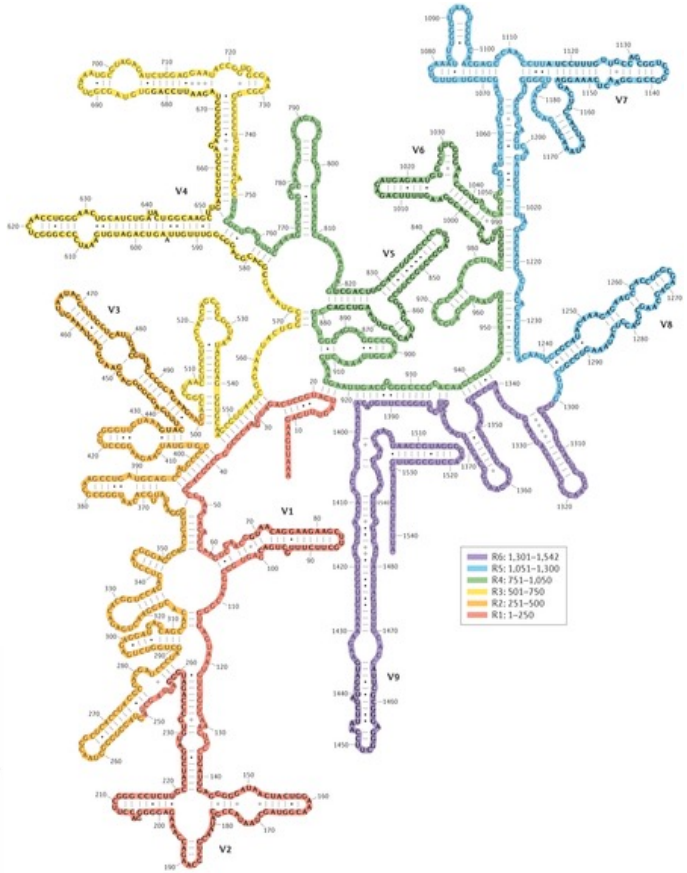
Allan Konopka

Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907

The 80's boom of culture-independent studies

The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences

NORMAN R. PACE, DAVID A. STAHL,
DAVID J. LANE, and GARY J. OLSEN



Nature Reviews | Microbiology

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

What is metagenomics?

JOURNAL OF BACTERIOLOGY, Feb. 1996, p. 591–599
0021-9193/96/$04.00+0

Vol. 178, No. 3

Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-Kilobase-Pair Genome Fragment from a Planktonic Marine Archaeon

JEFFEREY L. STEIN,[1]* TERENCE L. MARSH,[2] KE YING WU,[3] HIROAKI SHIZUYA,[4] AND EDWARD F. DELONG[3]*

Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products

Jo Handelsman[1], Michelle R Rondon[1], Sean F Brady[2], Jon Clardy[2] and Robert M Goodman[1]

meta | genome

"beyond the genome"

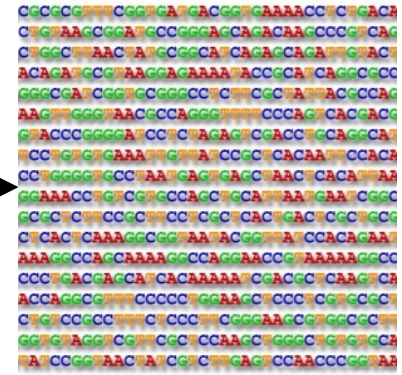
"[cloning of environmental DNA into *E. coli* for phenotype screening] has been made possible by advances in molecular biology and Eukaryotic genomics, which have laid the groundwork for cloning and functional analysis of the collective genomes of soil microflora, which we term the metagenome of the soil"

What is metagenomics?

Environmental sample

DNA extraction

Sequencing

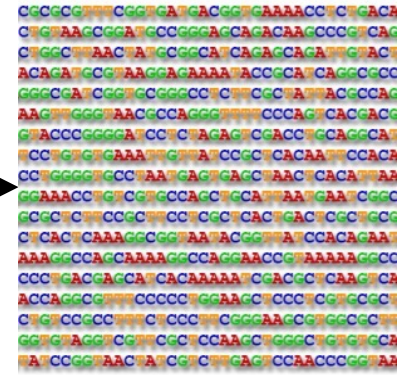


What is NOT metagenomics?

Environmental sample

DNA extraction

Sequencing



PCR
(e.g. 16S rRNA gene)

Sequencing

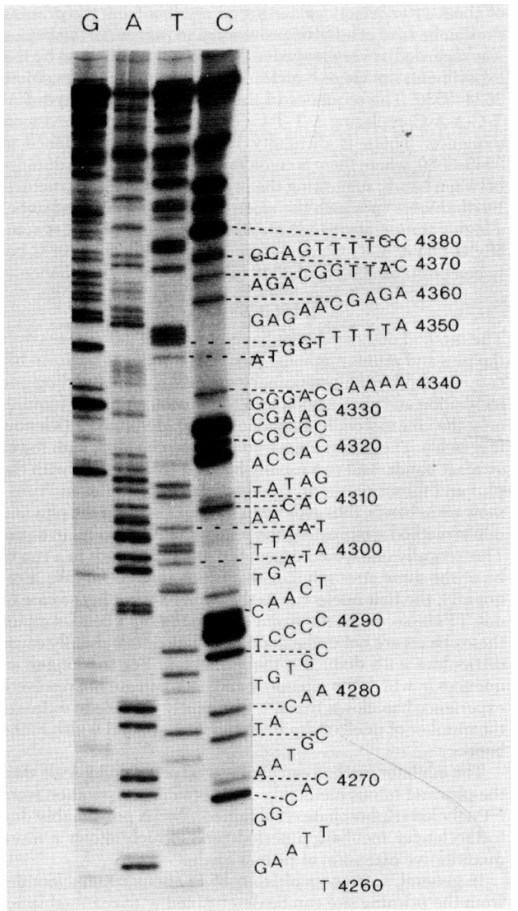
Metagenomics and the development of sequencing technologies

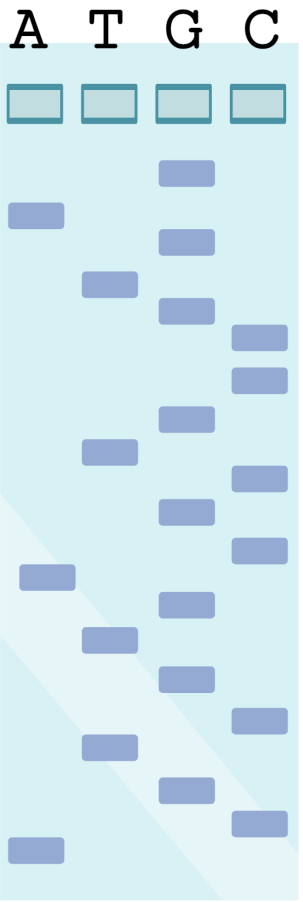
Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

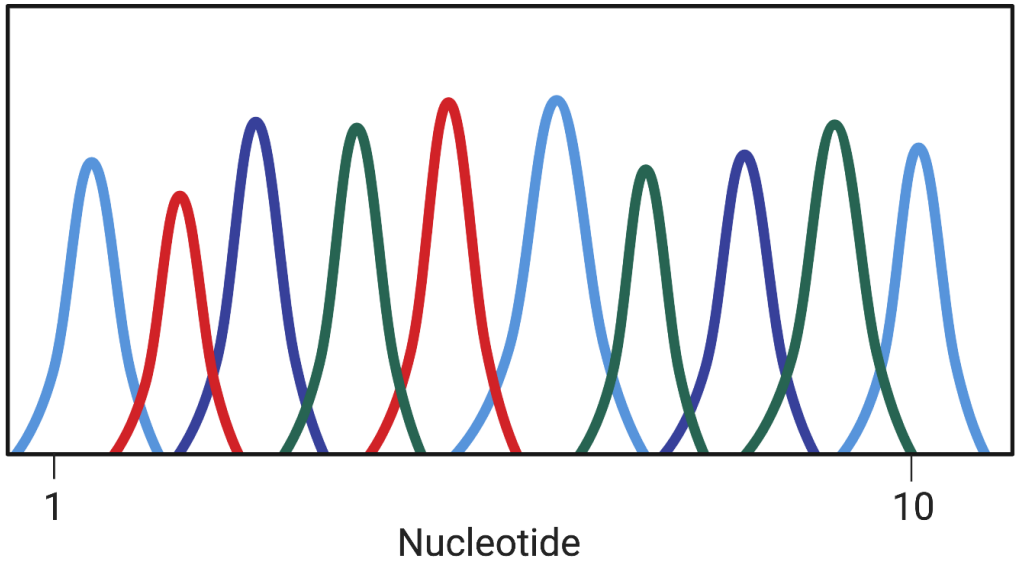
(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON





“High-throughput” Sanger sequencing (a.k.a first-generation sequencing)



96 samples/run

~900 bp/seq

~ 90 Kbp/run

life technologies™

Metagenomics: the early years

RESEARCH ARTICLE

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter[1,*], Karin Remington[1], John F. Heidelberg[3], Aaron L. Halpern[2], Doug Rusch[2], Jonathan A. Eisen[3], Dongying W…
+ See all authors and affiliations

Science 02 Apr 2004:
Vol. 304, Issue 5667, pp. 66-74
DOI: 10.1126/science.1093857

100 Mbp – 1 Gbp

OPEN ACCESS Freely available online PLOS BIOLOGY

The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific

Douglas B. Rusch[1*], Aaron L. Halpern[1], Granger Sutton[1], Karla B. Heidelberg[1,2], Shannon Williamson[1], Shibu Yooseph[1], Dongying Wu[1,3], Jonathan A. Eisen[1,3], Jeff M. Hoffman[1], Karin Remington[1,4], Karen Beeson[1], Bao Tran[1], Hamilton Smith[1], Holly Baden-Tillson[1], Clare Stewart[1], Joyce Thorpe[1], Jason Freeman[1], Cynthia Andrews-Pfannkoch[1], Joseph E. Venter[1], Kelvin Li[1], Saul Kravitz[1], John F. Heidelberg[1,2], Terry Utterback[1], Yu-Hui Rogers[1], Luisa I. Falcón[5], Valeria Souza[5], Germán Bonilla-Rosso[5], Luis E. Eguiarte[5], David M. Karl[6], Shubha Sathyendranath[7], Trevor Platt[7], Eldredge Bermingham[8], Victor Gallardo[9], Giselle Tamayo-Castillo[10], Michael R. Ferrari[11], Robert L. Strausberg[1], Kenneth Nealson[1,12], Robert Friedman[1], Marvin Frazier[1], J. Craig Venter[1]

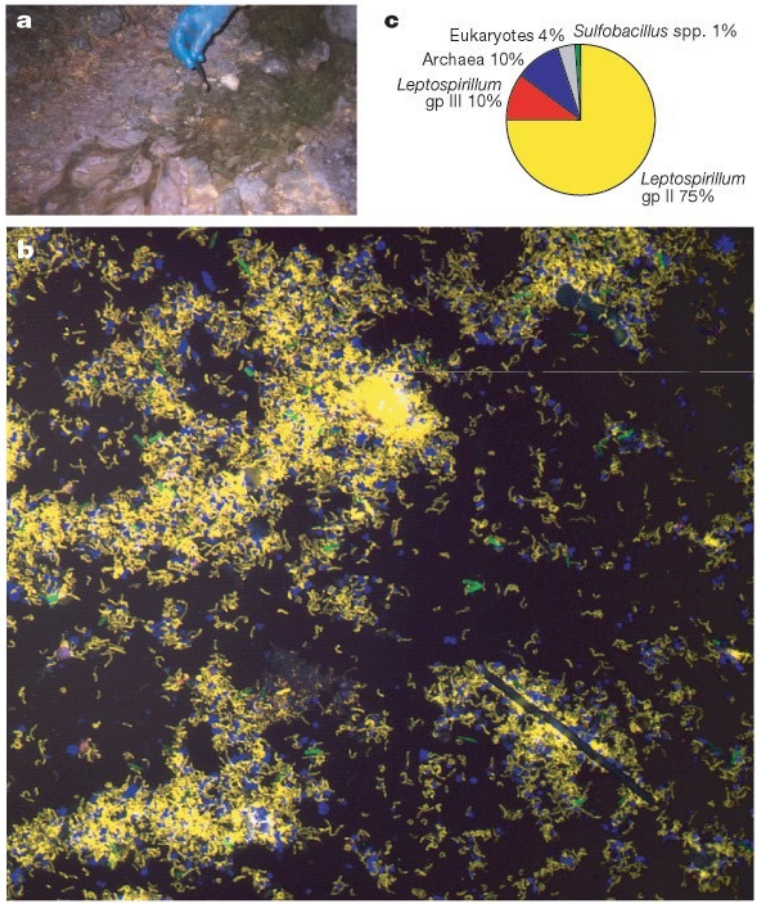
1 Gbp – 10 Gbp

Metagenomics: the early years

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



Acid mine drainage

76.2 Mbp

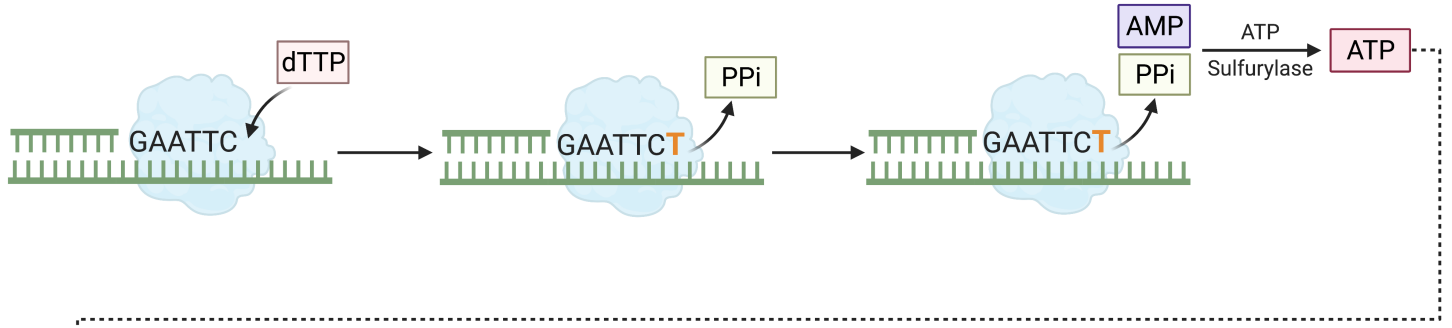
2 MAGs

Next-generation sequencing (a.k.a second-generation sequencing)

Genome sequencing in microfabricated high-density picolitre reactors

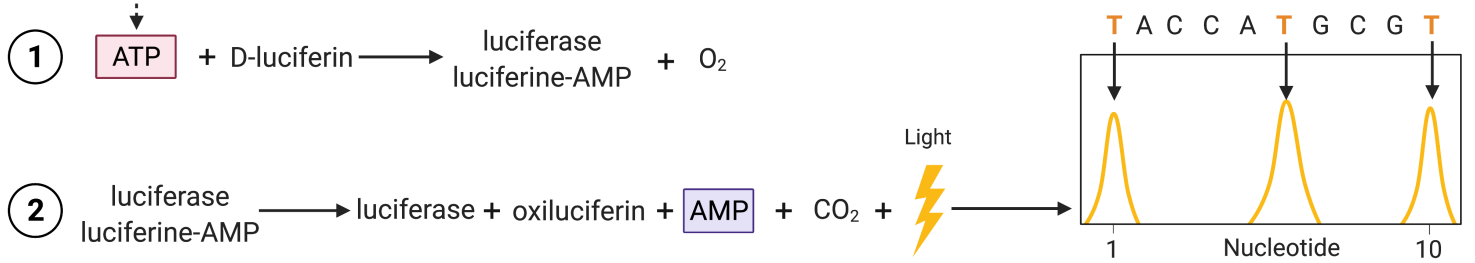
Marcel Margulies, Michael Egholm, [...] Jonathan M. Rothberg

Nature **437**, 376–380(2005) | Cite this article



400–1000 bp/seq

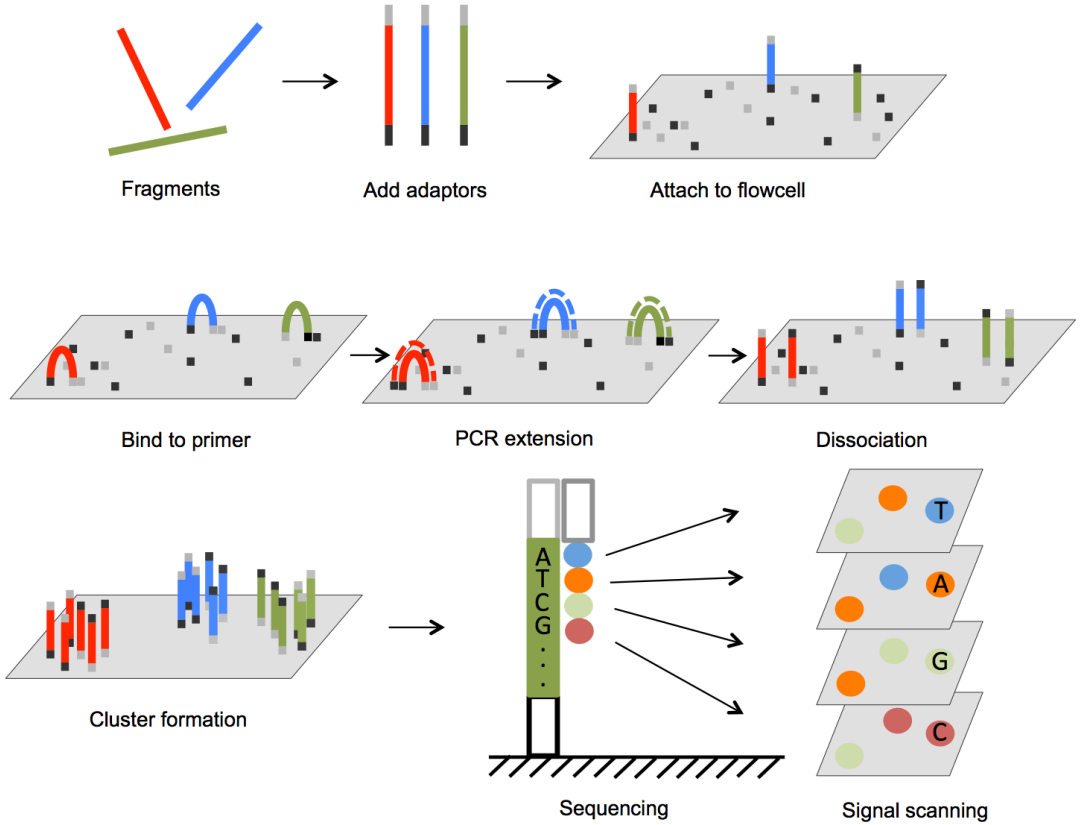
20–600 Mbp/run



Next-generation sequencing (a.k.a second-generation sequencing)

100–300 bp/seq

1–1000 Gbp/run



Metagenomics today

Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes

Mads Albertsen, Philip Hugenholtz, Adam Skarshewski, Kåre L Nielsen, Gene W Tyson & Per H Nielsen

Nature Biotechnology **31**, 533–538(2013) | Cite this article

Activated sludge
~90 Gbp
13 MAGs

Structure and function of the global ocean microbiome

Shinichi Sunagawa[1,*,†], Luis Pedro Coelho[1,*], Samuel Chaffron[2,3,4,*], Jens Roat Kultima[1], Karine Labadie[5], Guillem Salazar[6], ...

+ See all authors and affiliations

Science 22 May 2015:
Vol. 348, Issue 6237, 1261359
DOI: 10.1126/science.1261359

Tara Oceans:
7.2 Tbp

New Results

Follow this preprint

In-depth characterization of denitrifier communities across different soil ecosystems in the tundra

Igor S Pessi, Sirja Viitamaki, Anna-Maria Virkkala, Eeva Eronen-Rasimus, Tom O Delmont, Maija E Marushchak, Miska Luoto, Jenni Hultman

doi: https://doi.org/10.1101/2020.12.21.419267

796 MAGs!
(using Anvi'o)

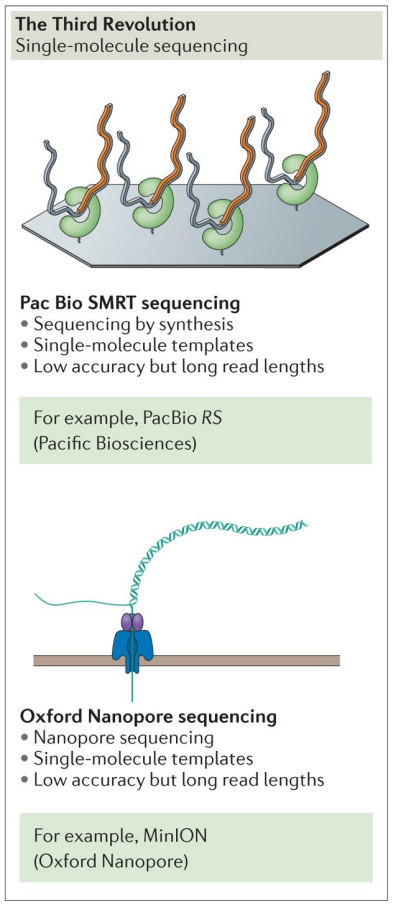
Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes

Tom O. Delmont, Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny TM Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker & A. Murat Eren

Nature Microbiology **3**, 804–813(2018) | Cite this article

957 MAGs!
(using Anvi'o)

Third-generation sequencing
a.k.a. ~~next-next-generation sequencing~~



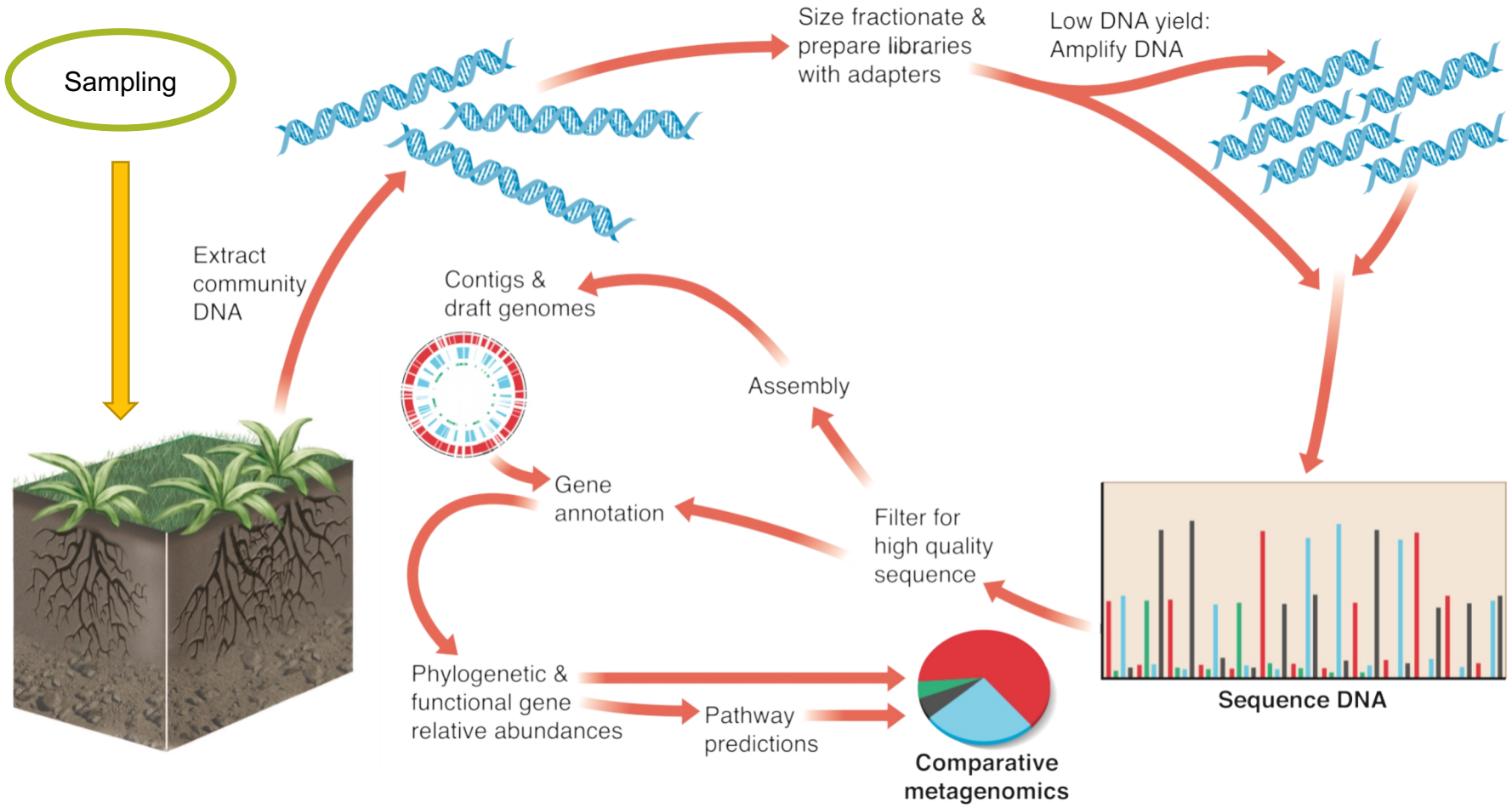
PacBioSequel II

- 20 kb/seq
- **30 Gbp/run**
- 99.92 % accuracy

Nanopore Minion:

- > 4 Mb/seq
- **1–50 Gbp/run**
- 97 % accuracy

Metagenomes:
from samples to genomes


Sampling
Extract community DNA
Size fractionate & prepare libraries with adapters
Low DNA yield: Amplify DNA
Contigs & draft genomes
Assembly
Gene annotation
Filter for high quality sequence
Sequence DNA
Phylogenetic & functional gene relative abundances
Pathway predictions
Comparative metagenomics

Study design and sampling

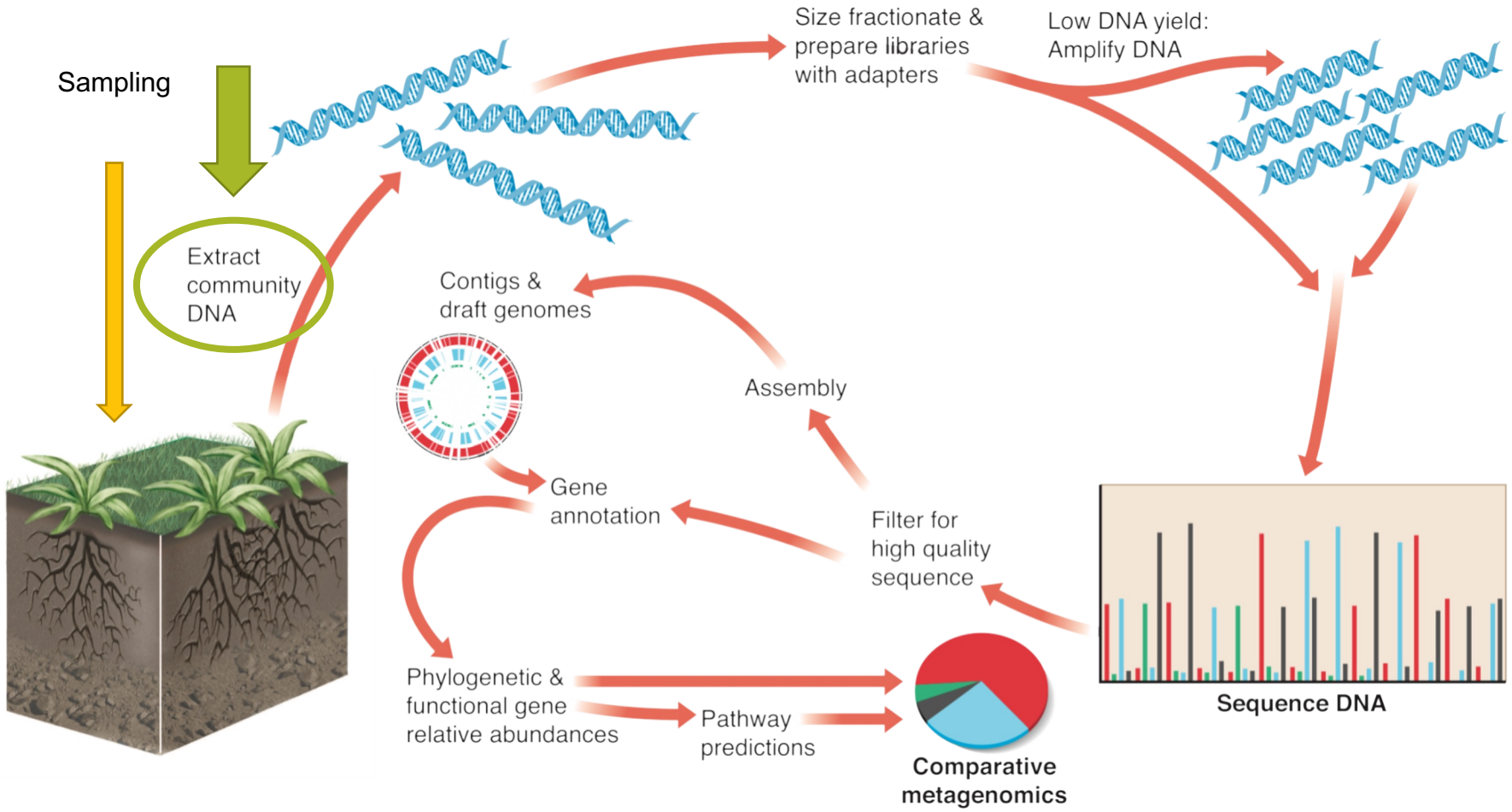
Study design is a critical step in every metagenomic study

- Spatial variation (microhabitats)
- Temporal variation (daily and seasonal)

Sample collection and preservation protocols can affect both the quality and the accuracy of metagenomics data

- Cross-contamination
- Enough biomass

Importance of (proper) metadata!!!


Sampling
Extract community DNA
Size fractionate & prepare libraries with adapters
Low DNA yield: Amplify DNA
Contigs & draft genomes
Assembly
Gene annotation
Filter for high quality sequence
Sequence DNA
Phylogenetic & functional gene relative abundances
Pathway predictions
Comparative metagenomics

DNA extraction

Critical step as DNA extracts have to be:

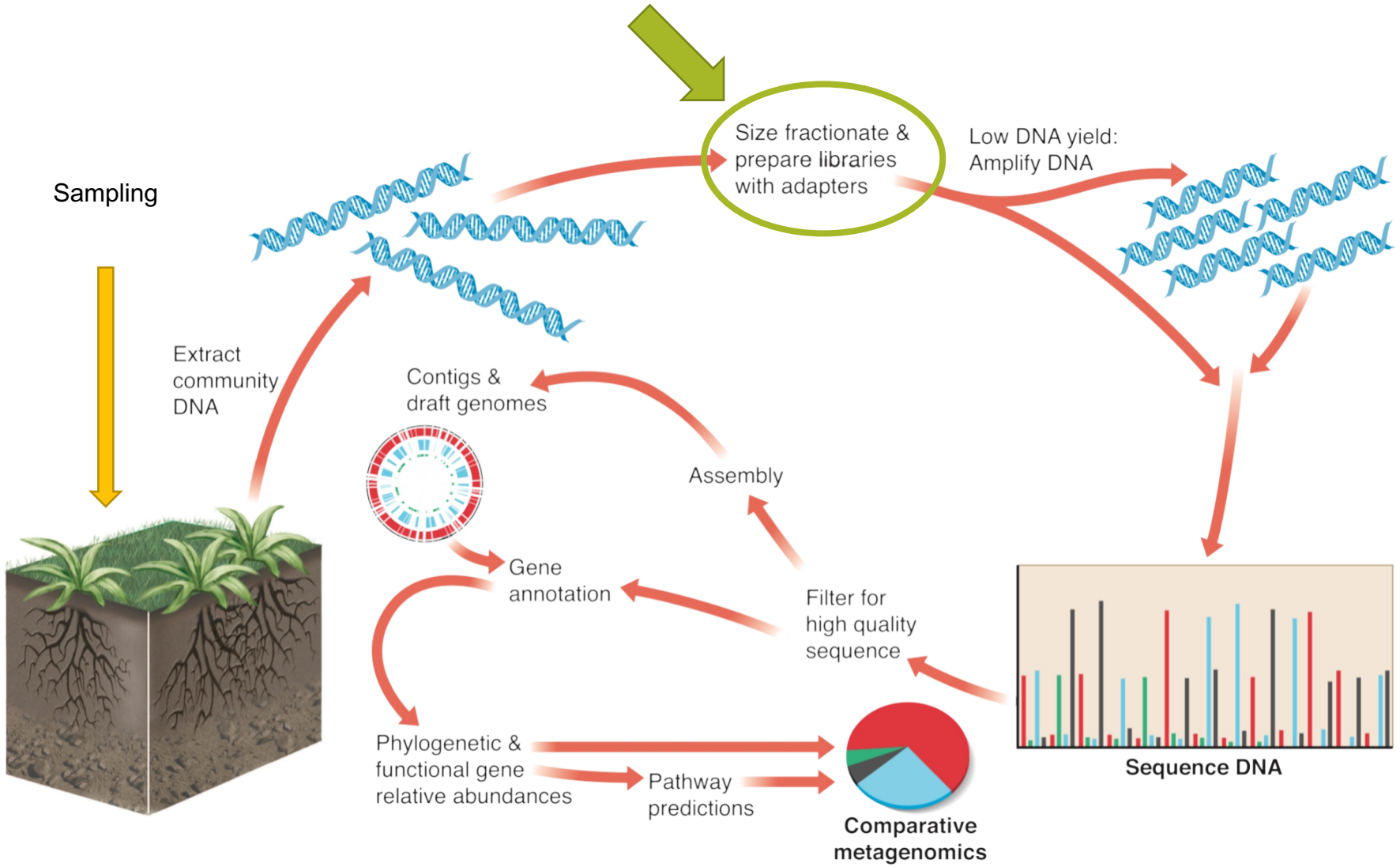
- Representative of the whole community
- Enough amount
- Not too fragmented

Environmental samples are complex

- Different microorganisms with different types of cell walls
- Varying abundances
- Cell aggregates
- Extracellular substances and inhibitors

DNA extraction from low-biomass samples is tricky

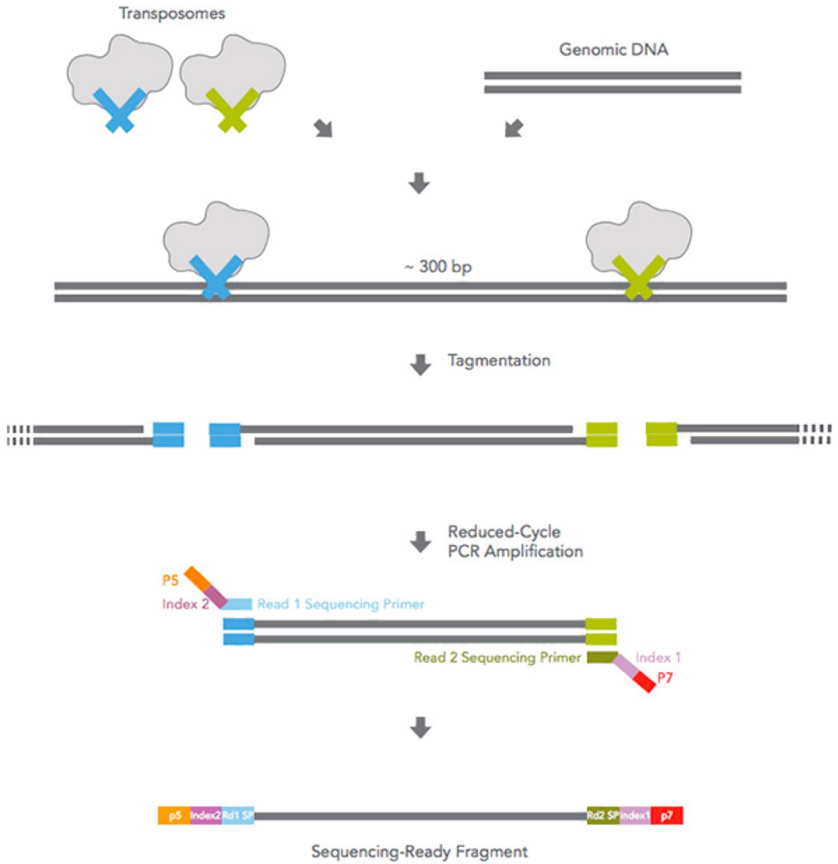
- Contamination risks become critical
- Blank controls

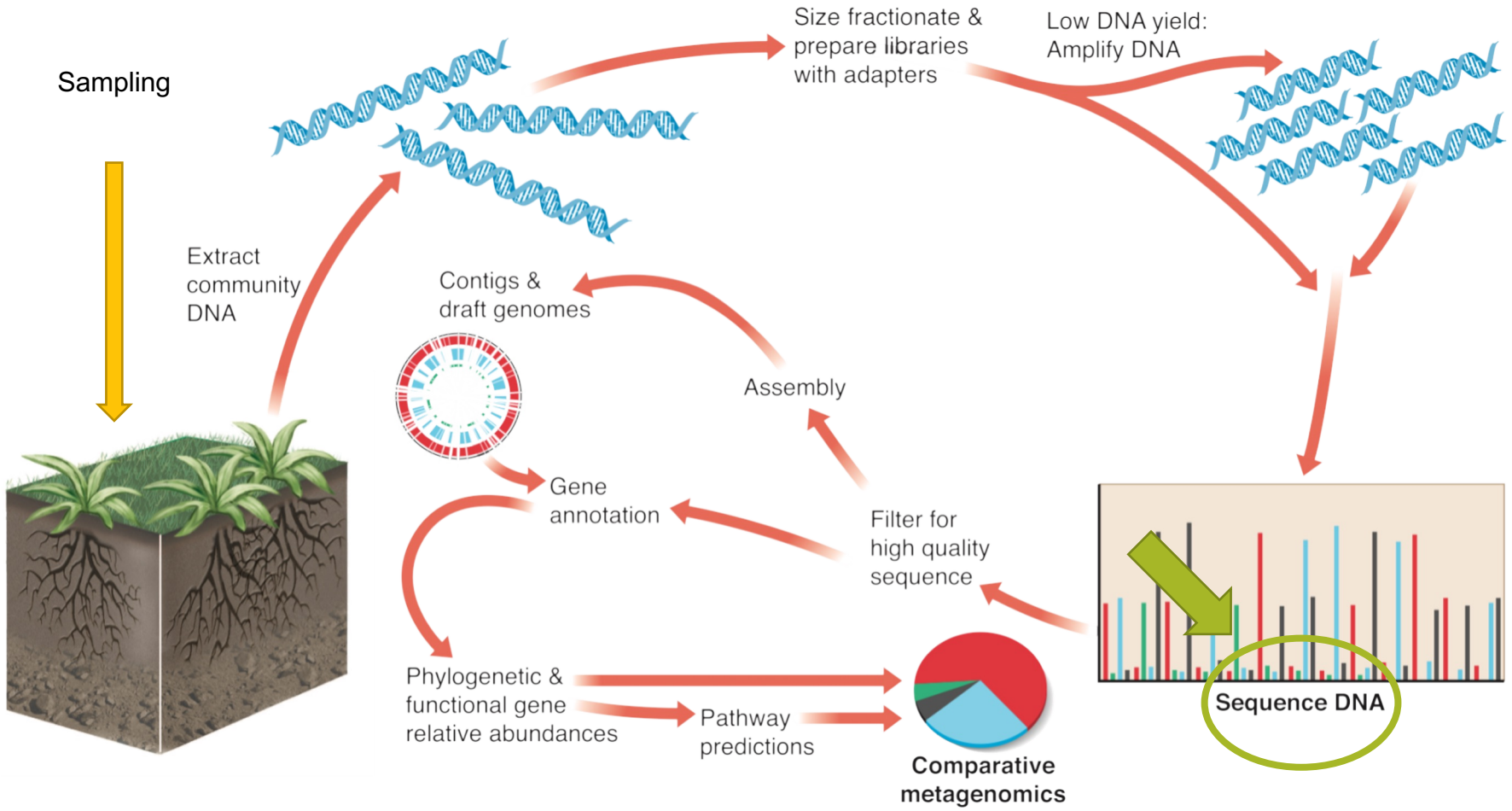

Sampling
Extract community DNA
Size fractionate & prepare libraries with adapters
Low DNA yield: Amplify DNA
Contigs & draft genomes
Assembly
Gene annotation
Filter for high quality sequence
Sequence DNA
Phylogenetic & functional gene relative abundances
Pathway predictions
Comparative metagenomics

Library preparation

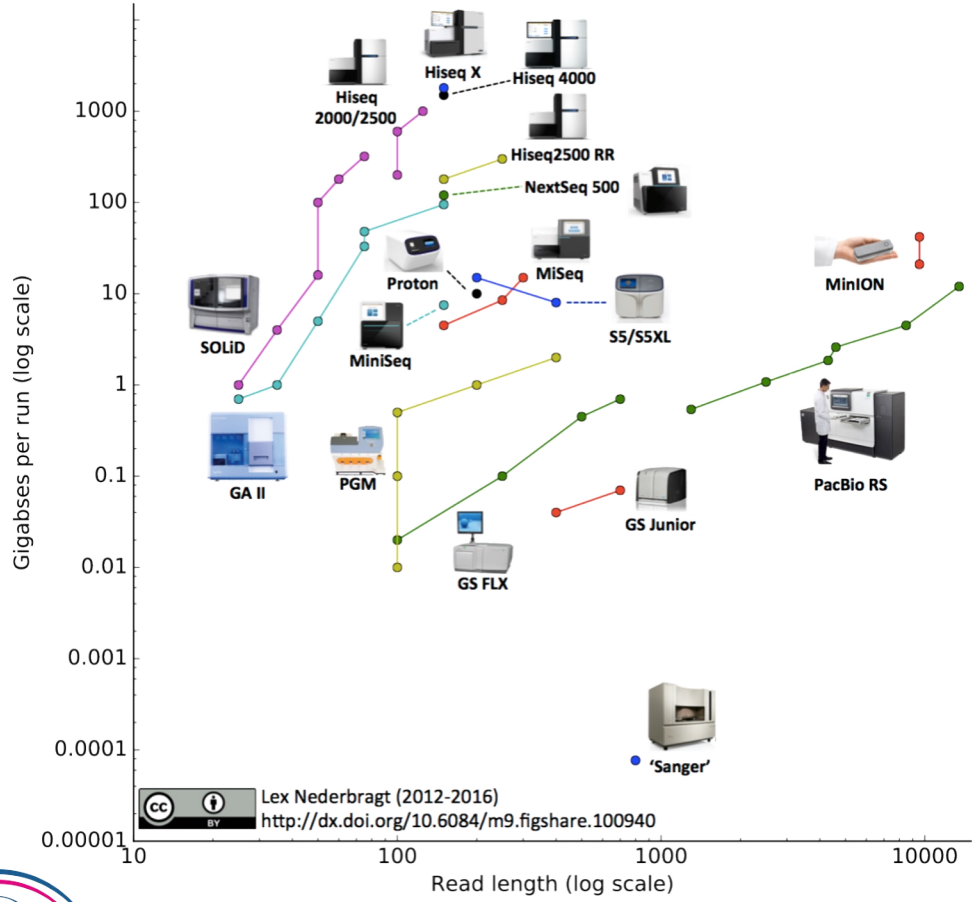
Size selection

Ligation of sequencing adaptors



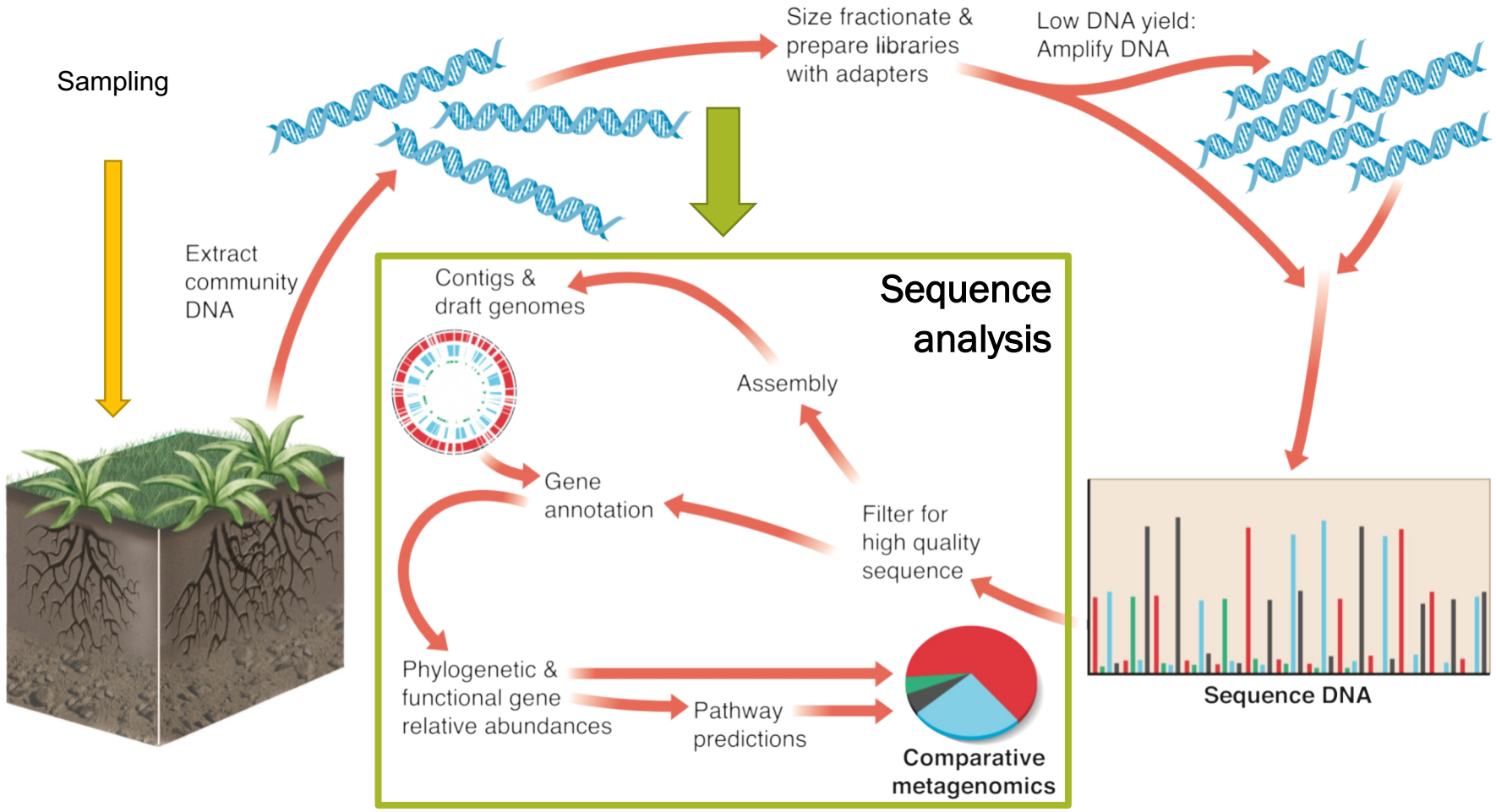

Sampling
Extract community DNA
Size fractionate & prepare libraries with adapters
Low DNA yield: Amplify DNA
Contigs & draft genomes
Assembly
Gene annotation
Filter for high quality sequence
Sequence DNA
Phylogenetic & functional gene relative abundances
Pathway predictions
Comparative metagenomics

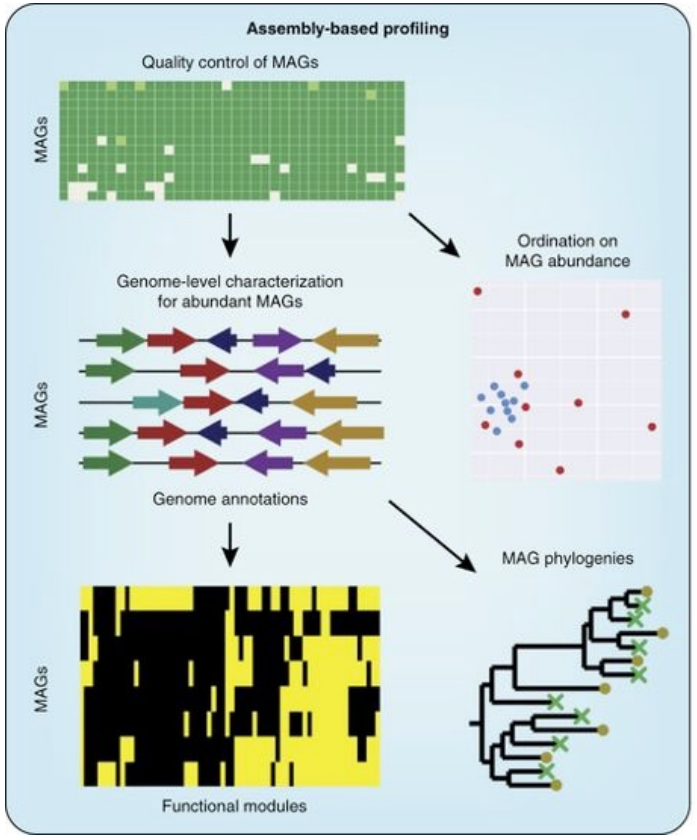
Sequencing

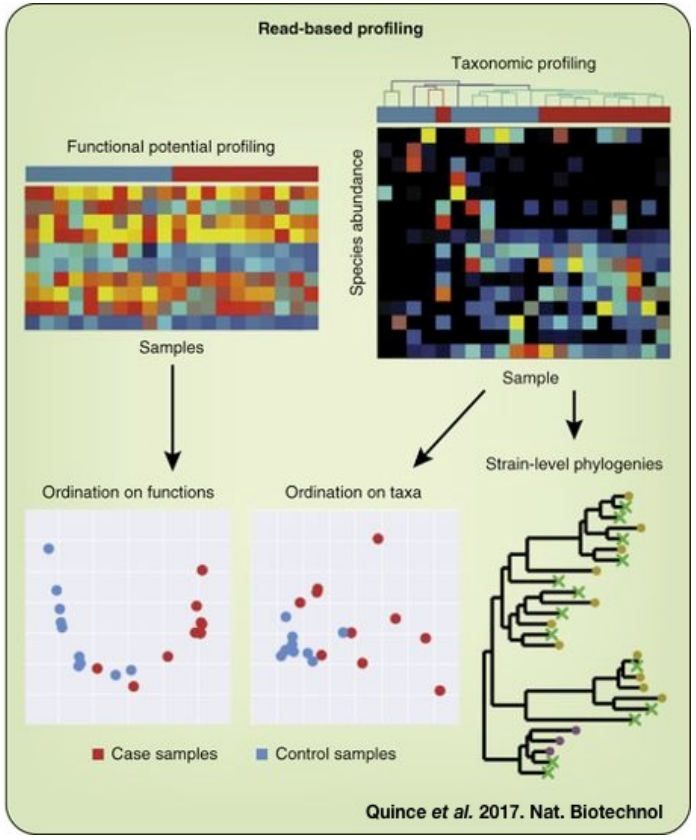


Features

- Read size
- Read depth (output)
- Error rates
- Price


Sampling
Extract community DNA
Size fractionate & prepare libraries with adapters
Low DNA yield: Amplify DNA
Sequence analysis
Contigs & draft genomes
Assembly
Gene annotation
Filter for high quality sequence
Phylogenetic & functional gene relative abundances
Pathway predictions
Comparative metagenomics
Sequence DNA



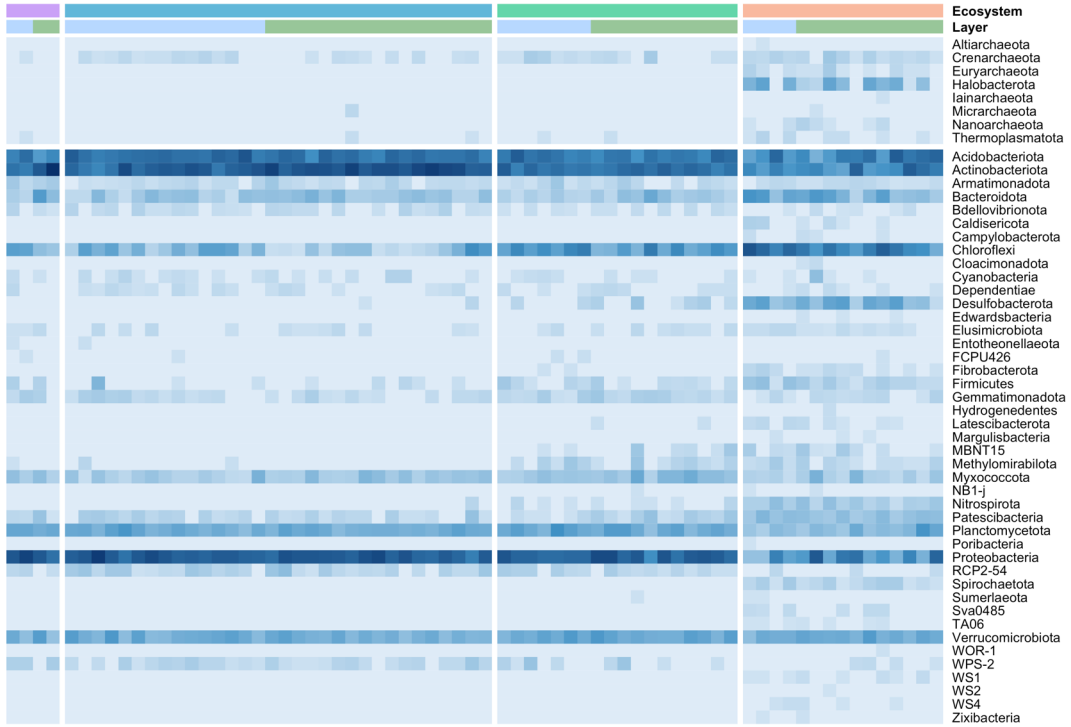


Read- *vs.* assembly-based metagenomics

Read-based metagenomics

Raw reads -> Similarity search (BLAST)

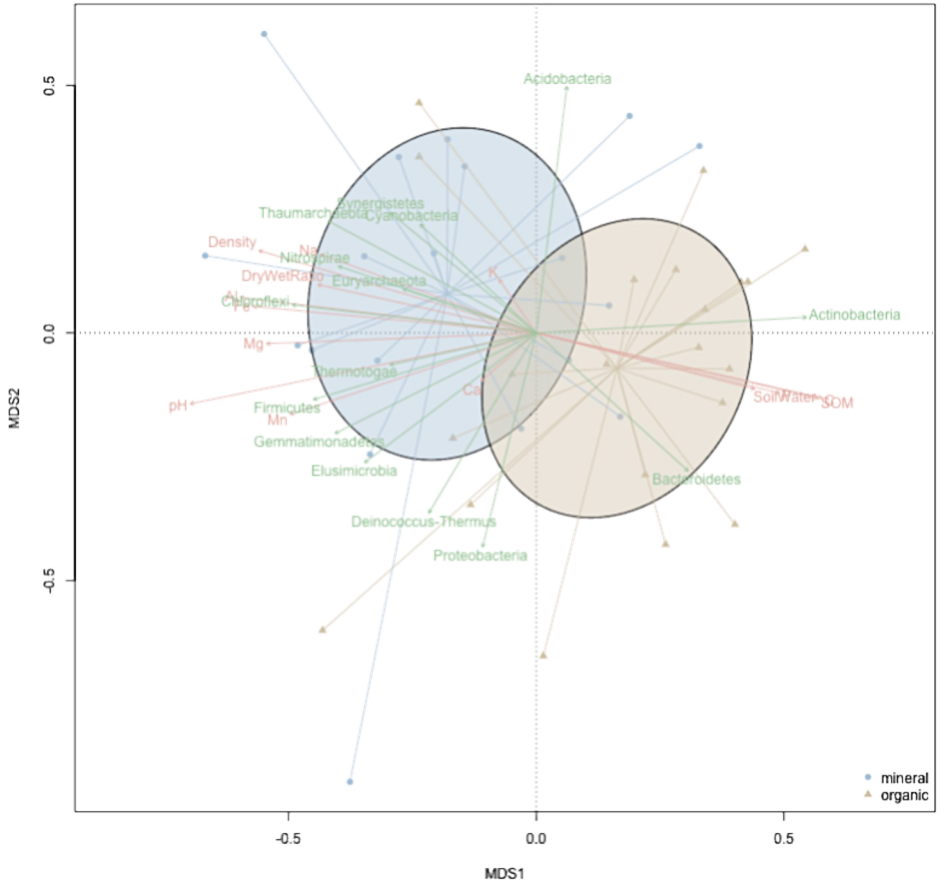
- Taxonomy: SILVA, Greengenes, RDP...
- Functional: KEGG, COG, SEED...



Read-based metagenomics

Profiling of taxa and functions

- Environmental gradients
- Time series
- Controls vs. treatments



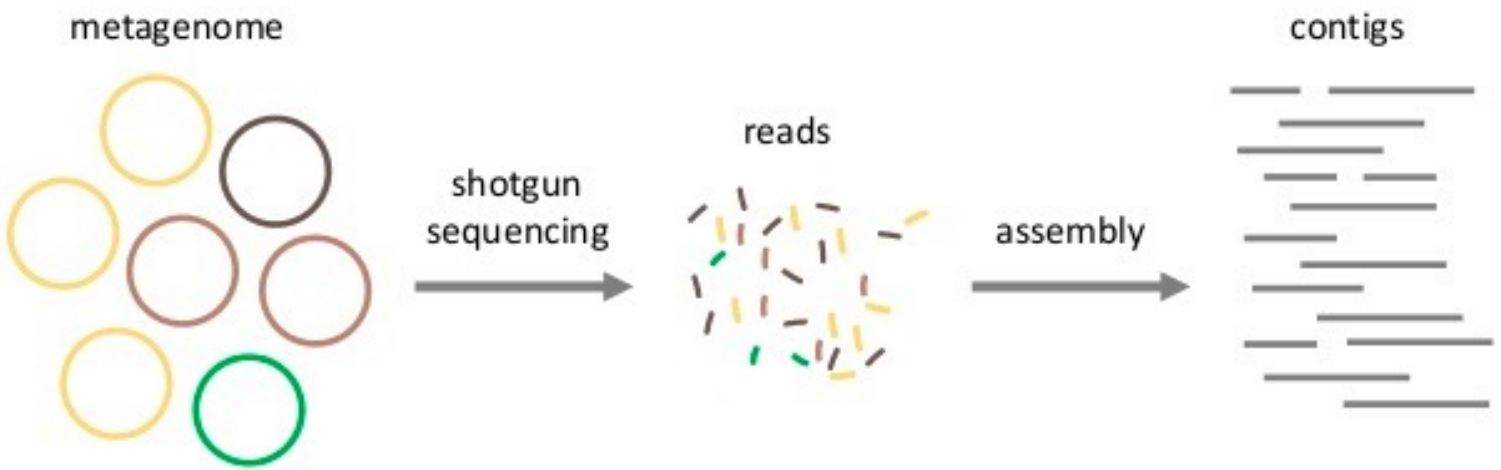
Strenghts and weaknesses of read-based metagenomics

Comprehensiveness	Can provide an aggregate picture of community function or structure, but is based only on the fraction of reads that map effectively to reference dbs
Community complexity	Can deal with communities of arbitrary complexity given sufficient sequencing depth and satisfactory reference database coverage
Novelty	Cannot resolve organisms for which genomes of close relatives are unknown
Computational burden	Can be performed efficiently, enabling large meta-analyses
Genome-resolved metabolism	Can typically resolve only the aggregate metabolism of the community, and links with phylogeny are only possible in the context of known reference genomes
Expert manual supervision	Usually does not require manual curation, but selection of reference genomes to use could involve human supervision
Integration with microbial genomics	Obtained profiles cannot be directly put into the context of genomes derived from pure cultured isolates

Quince *et al*, 2017, Nat Biotechnol

Assembly-based metagenomics

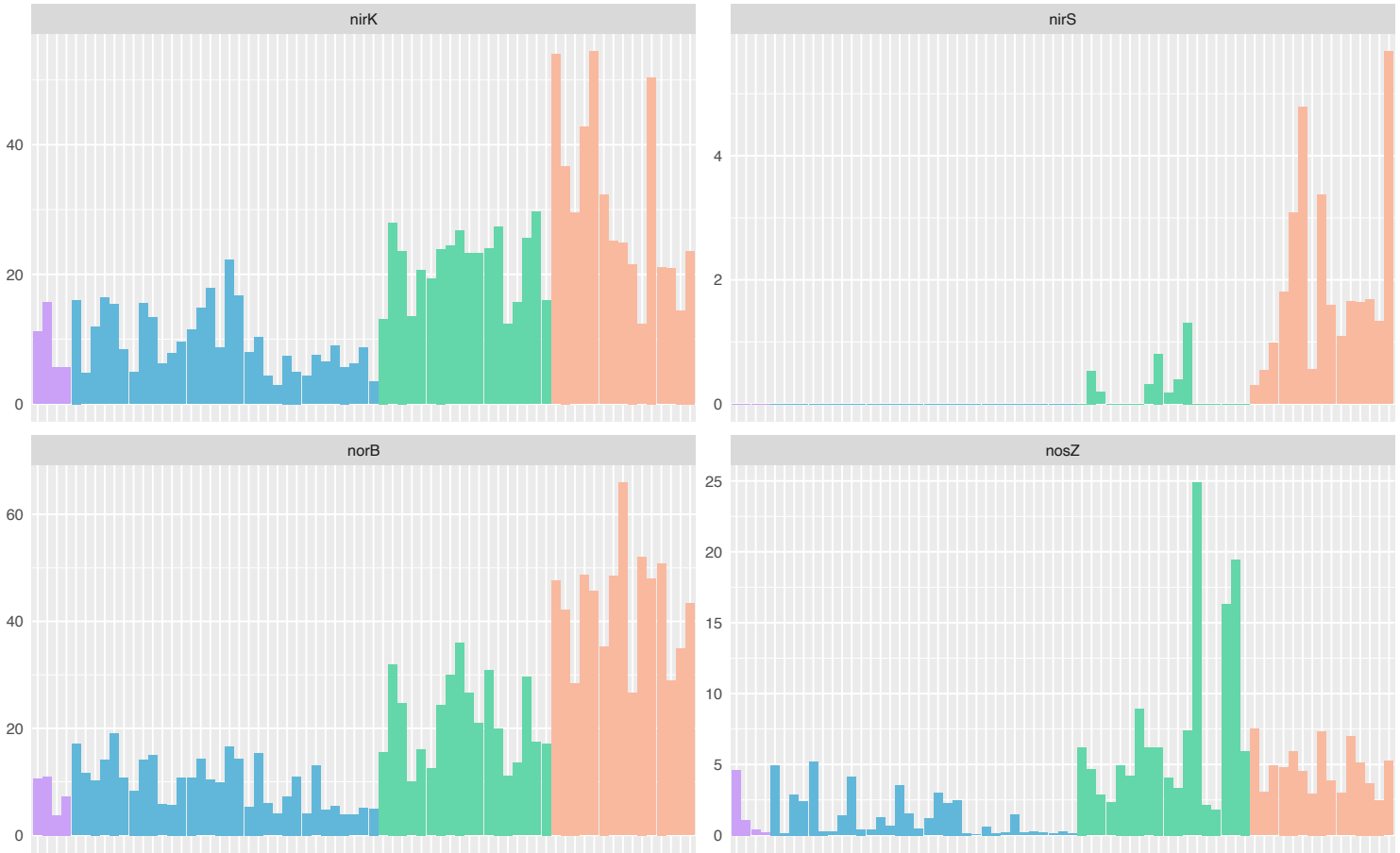
Reads are assembled into larger contiguous segments (contigs)



Assembly-based metagenomics

Profiling of taxa and functions

- Environmental gradients
- Time series
- Controls vs. treatments

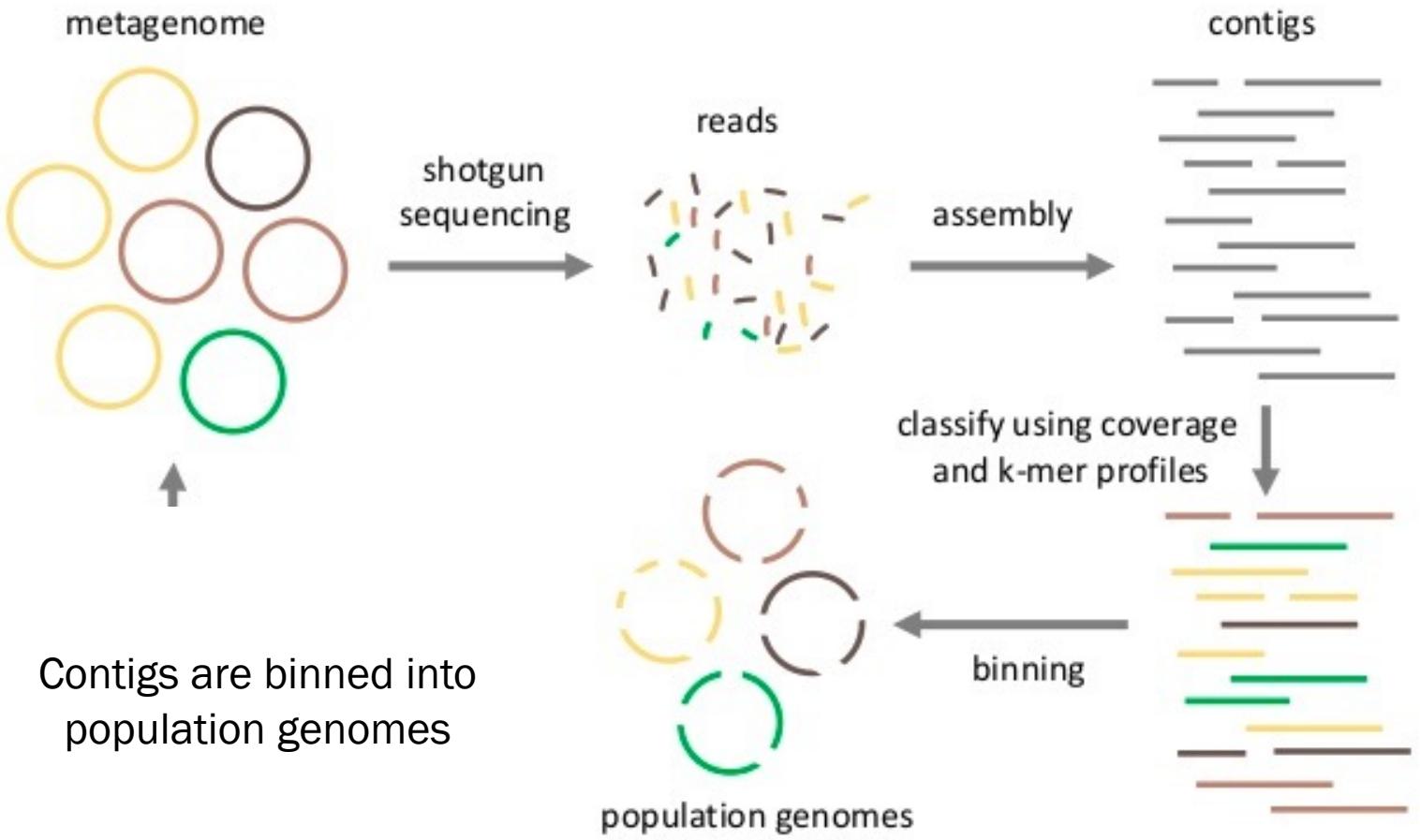


Assembly-based metagenomics

Challenges of metagenome assembly

- Metagenomes are complex: 1 g of soil typically contain ~10^9 genomes
- High frequency of polymorphisms and genome variations
- Coverage (abundance) of individual genomes vary
- Sequencing coverage still low
- Repetitive (low-complexity) regions
- Relatively good assemblies only possible for low-complexity samples

Genome-resolved metagenomics



Contigs are binned into population genomes

Binning of genomes

Based on e.g.:

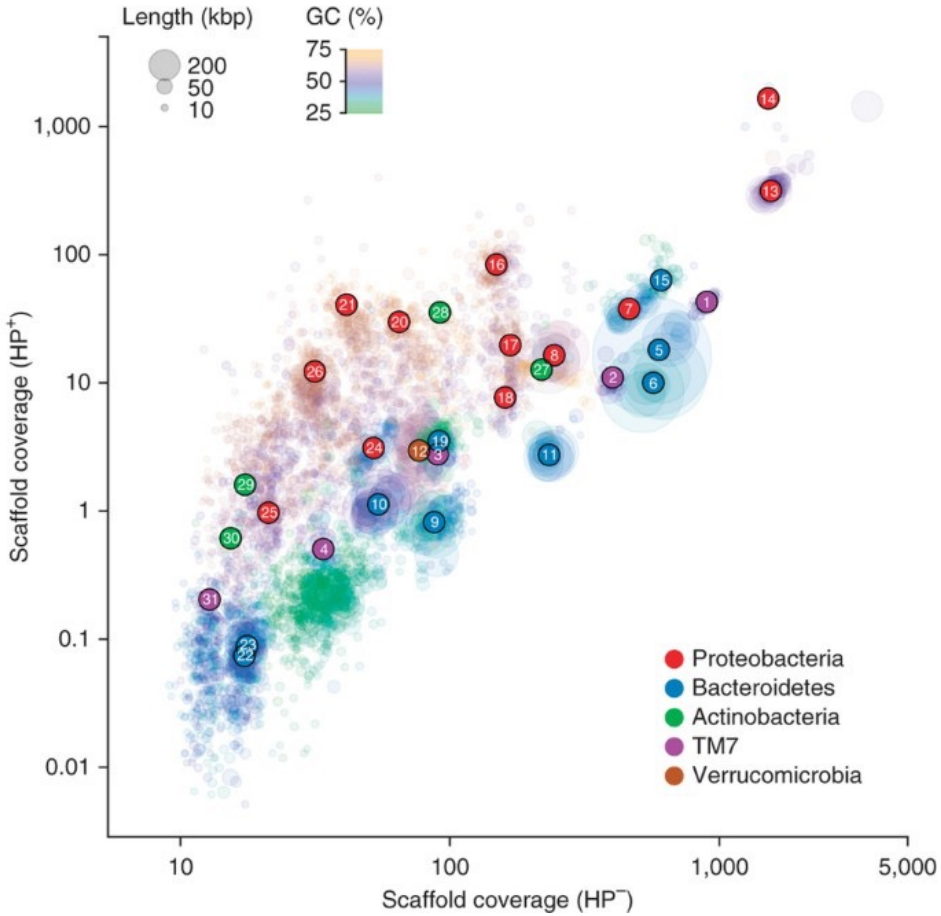
- Coverage
- Tetranucleotide frequency

Automated

- CONCOCT, metaBAT, etc.

Manual

- Anvi'o



Manual binning with anvi'o

PeerJ

Anvi'o: an advanced analysis and visualization platform for 'omics data

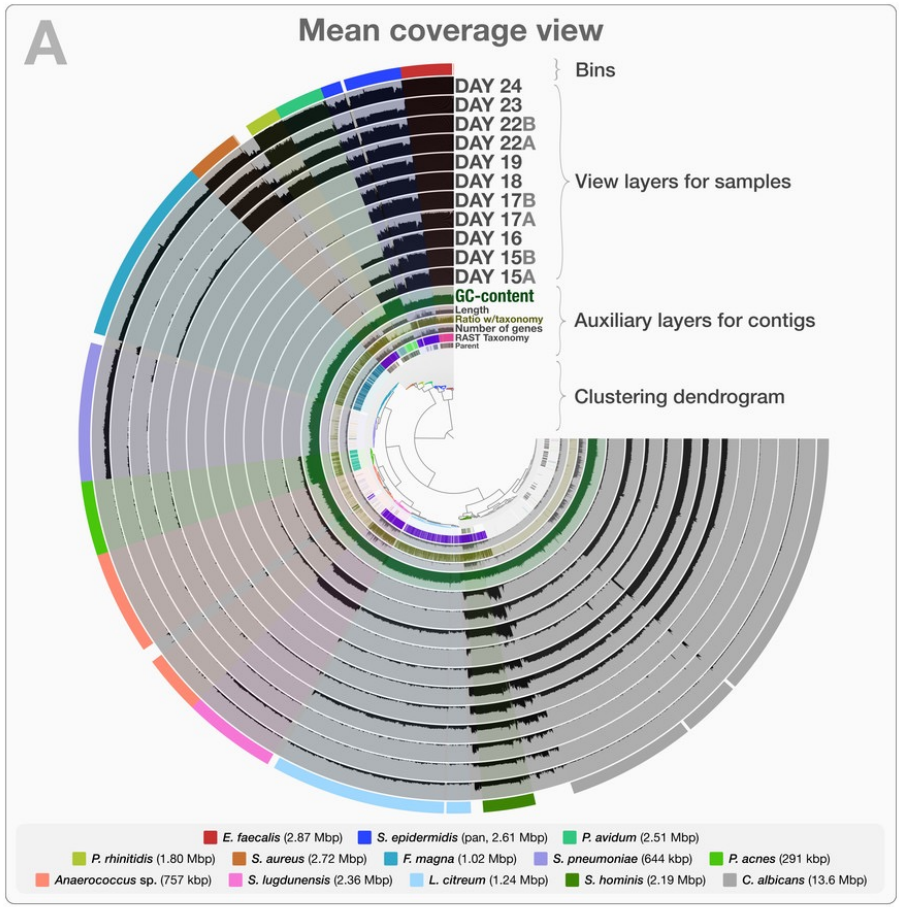
A. Murat Eren[1,2], Özcan C. Esen[1], Christopher Quince[3], Joseph H. Vineis[1], Hilary G. Morrison[1], Mitchell L. Sogin[1] and Tom O. Delmont[1]

[1] Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, United States
[2] Department of Medicine, The University of Chicago, Chicago, IL, United States
[3] Warwick Medical School, University of Warwick, Coventry, United Kingdom

Community-led, integrated, reproducible multi-omics with anvi'o

Big data abound in microbiology, but the workflows designed to enable researchers to interpret data can constrain the biological questions that can be asked. Five years after anvi'o was first published, this community-led multi-omics platform is maturing into an open software ecosystem that reduces constraints in 'omics data analyses.

A. Murat Eren, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E. Miller, Matthew S. Schechter, Isaac Fink, Jessica N. Pan, Mahmoud Yousef, Emily C. Fogarty, Florian Trigodet, Andrea R. Watson, Özcan C. Esen, Ryan M. Moore, Quentin Clayssen, Michael D. Lee, Veronika Kivenson, Elaina D. Graham, Bryan D. Merrill, Antti Karkman, Daniel Blankenberg, John M. Eppley, Andreas Sjödin, Jarrod J. Scott, Xabier Vázquez-Campos, Luke J. McKay, Elizabeth A. McDaniel, Sarah L. R. Stevens, Rika E. Anderson, Jessika Fuessel, Antonio Fernandez-Guerra, Lois Maignien, Tom O. Delmont and Amy D. Willis



Strenghts and weaknesses of assembly-based metagenomics

Comprehensiveness	Can construct multiple whole genomes, but only for organisms with enough coverage to be assembled and binned
Community complexity	In complex communities, only a fraction of the genomes can be resolved by assembly
Novelty	Can resolve genomes of entirely novel organisms with no sequenced relatives
Computational burden	Requires computationally costly assembly, mapping and binning
Genome-resolved metabolism	Can link metabolism to phylogeny through completely assembled genomes, even for novel diversity
Expert manual supervision	Manual curation required for accurate binning and scaffolding and for misassembly detection
Integration with microbial genomics	Assemblies can be fed into microbial genomic pipelines designed for analysis of genomes from pure cultured isolates

Quince *et al*, 2017, Nat Biotechnol

Current bottlenecks in metagenomic analyses

Production of data has dramatically increased over the past year

- Reduction in price, robust protocols/kits are available

Processing and analysis steps are the main bottleneck

- Availability of computational resources
- Bioinformatics expertise
- Large diversity of specialized software are available, but how to choose the most appropriate?
- The “bioinformatics middle-class”

Comprehensiveness of genome catalogues

- Available microbial genomes are biased toward model organisms, pathogens and easily cultivable bacteria

Live or dead dilemma

Low-biomass samples

- Extreme environments
- Host-associated environments

Data sharing and reproducibility