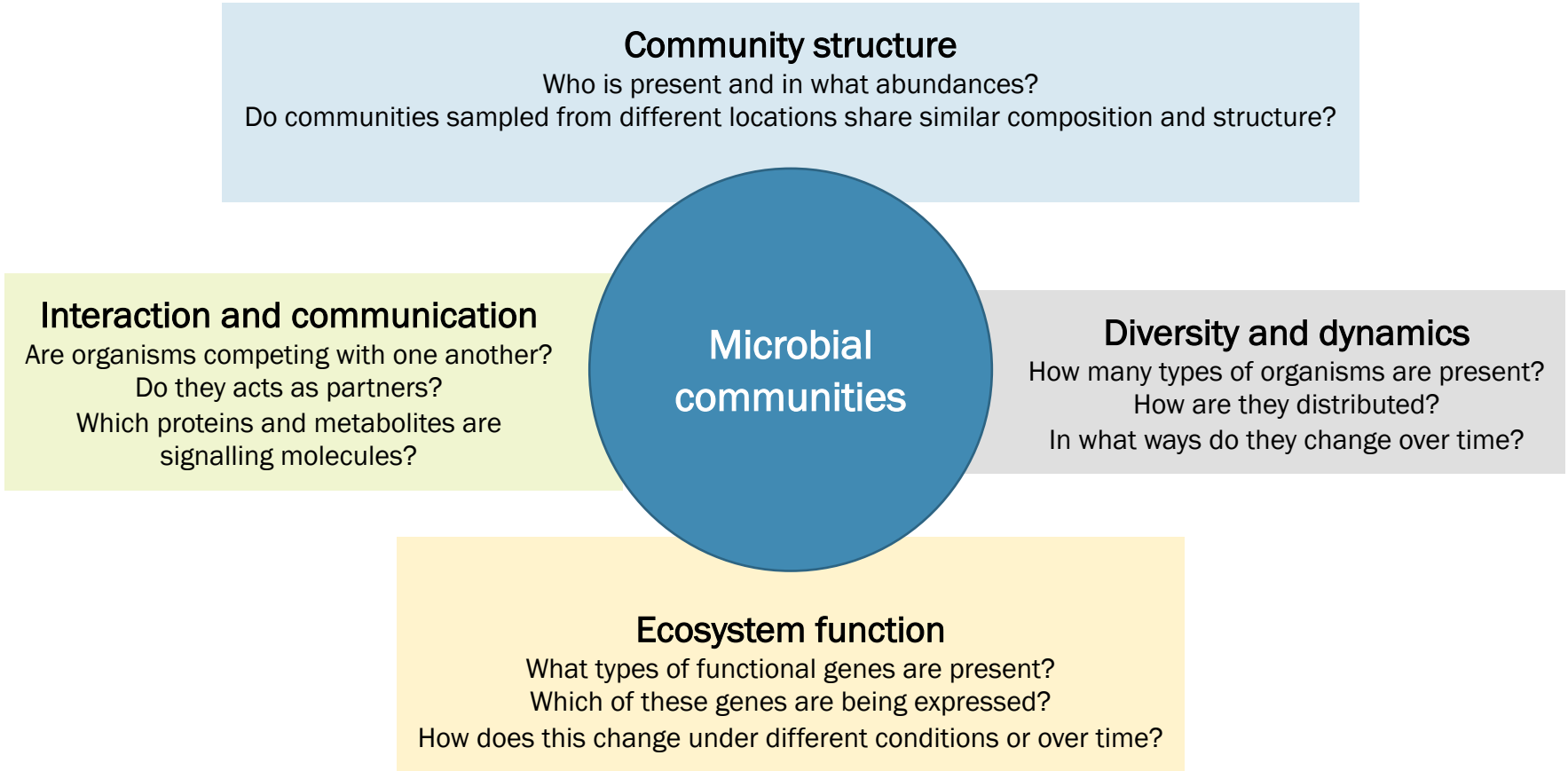
Environmental metagenomics

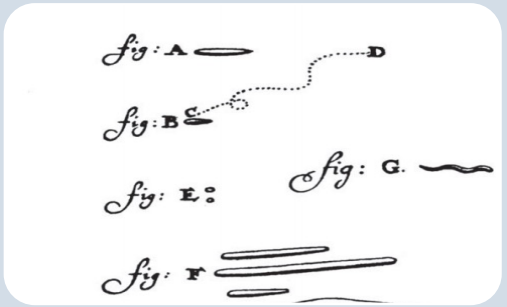
Introduction to metagenomics

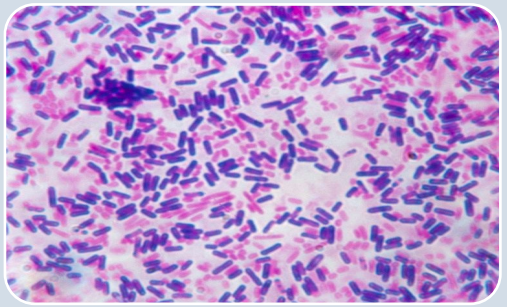
Igor S. Pessi & Antti Karkman

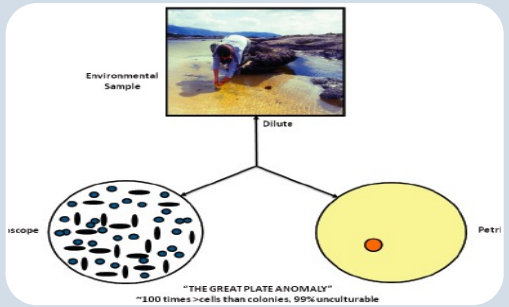
Metagenomics is the ultimate way to study microbial communities



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

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"

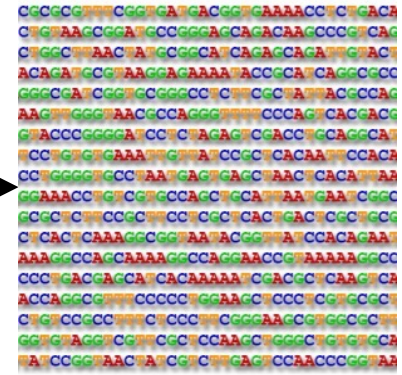
"[cloning of environmental DNA into *E. coli* for phenotype screening] has been made possible by advances in molecular biology and Eukaryotic genomics, which have laid the groundwork for cloning and functional analysis of the collective genomes of soil microflora, which we term the metagenome of the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

What is metagenomics?

Environmental sample

DNA extraction

Sequencing

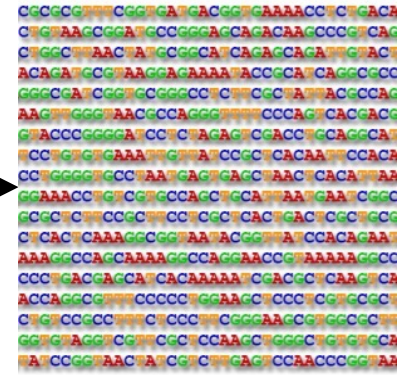


What is NOT metagenomics?

Environmental sample

DNA extraction

Sequencing



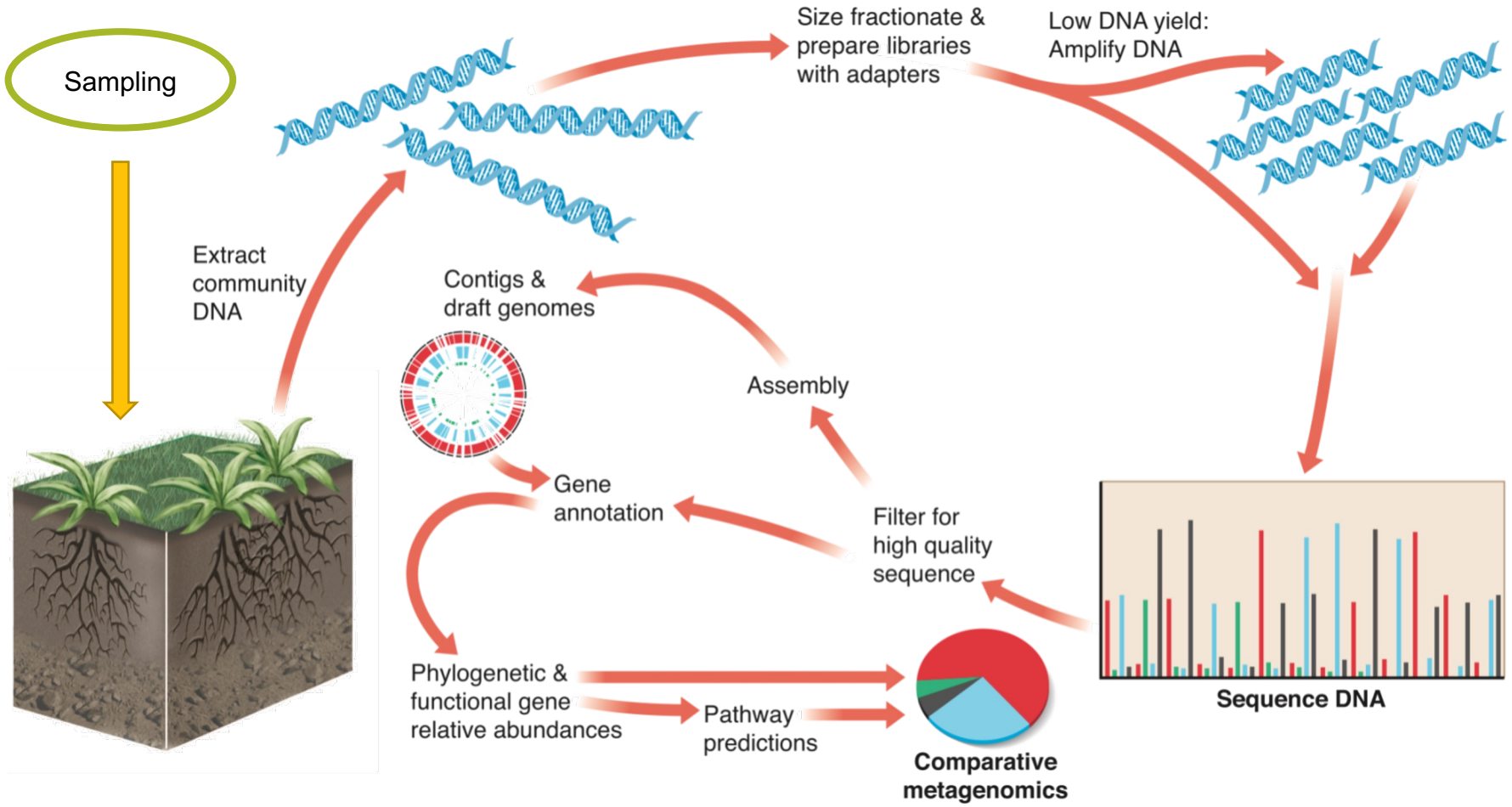
Sequencing

PCR
(e.g. 16S rRNA gene)

Metagenomics *vs.* sequencing technologies

www.pacb.com/blog/the-evolution-of-dna-sequencing-tools/

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

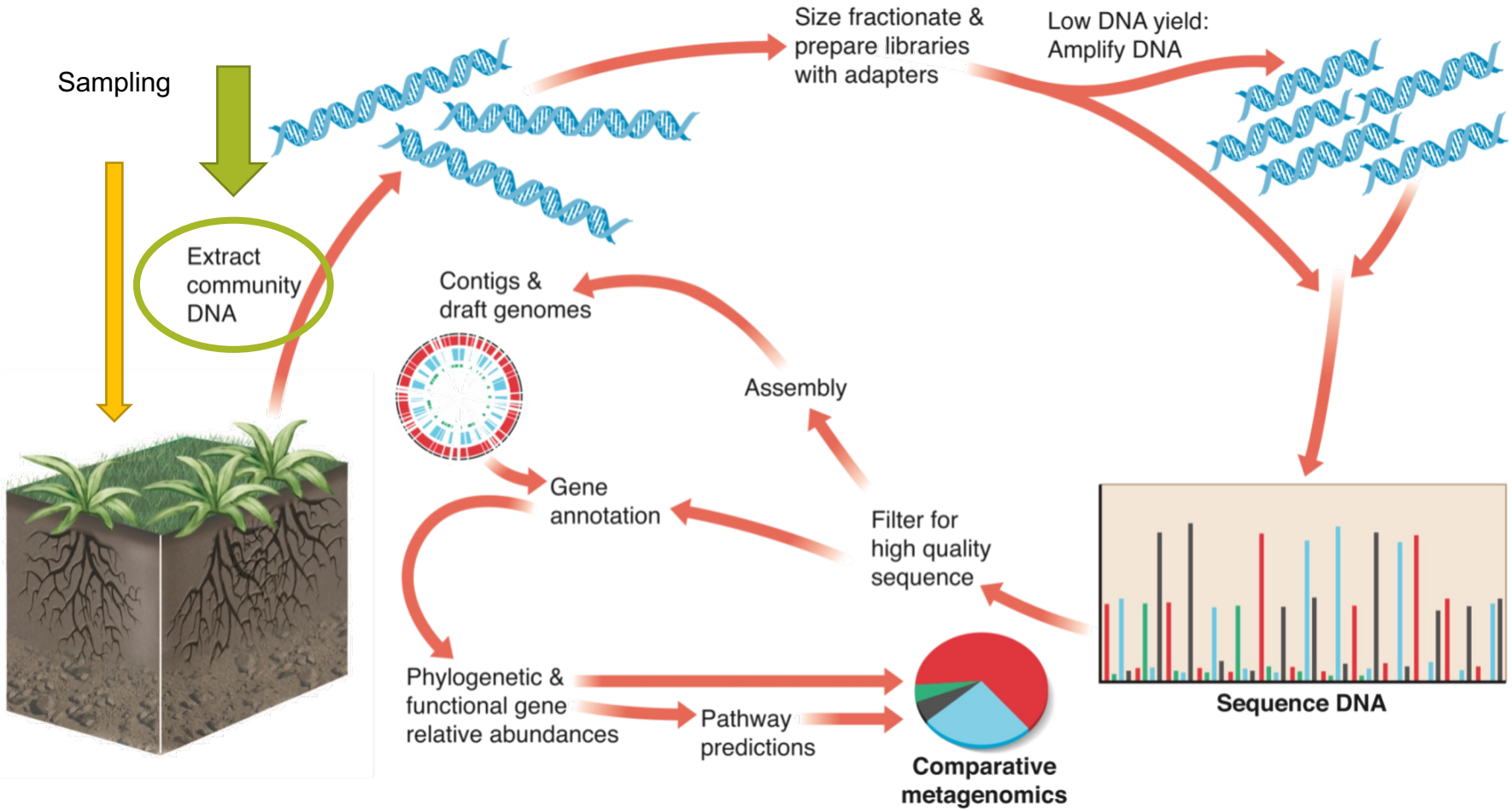
Study design is a critical step in every metagenomic study

- 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:

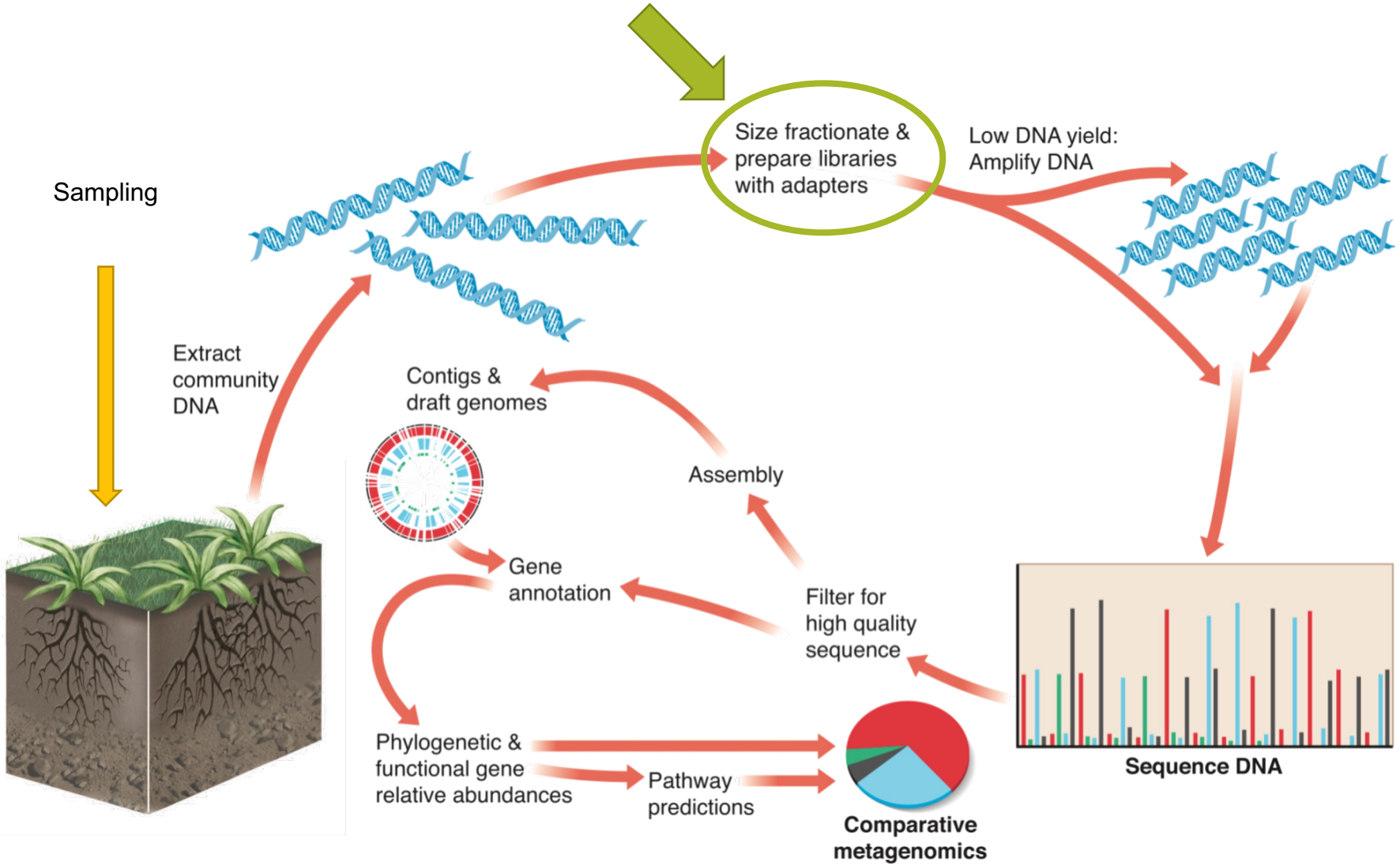
- Representative of the whole community
- Enough amount
- Not too fragmented

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

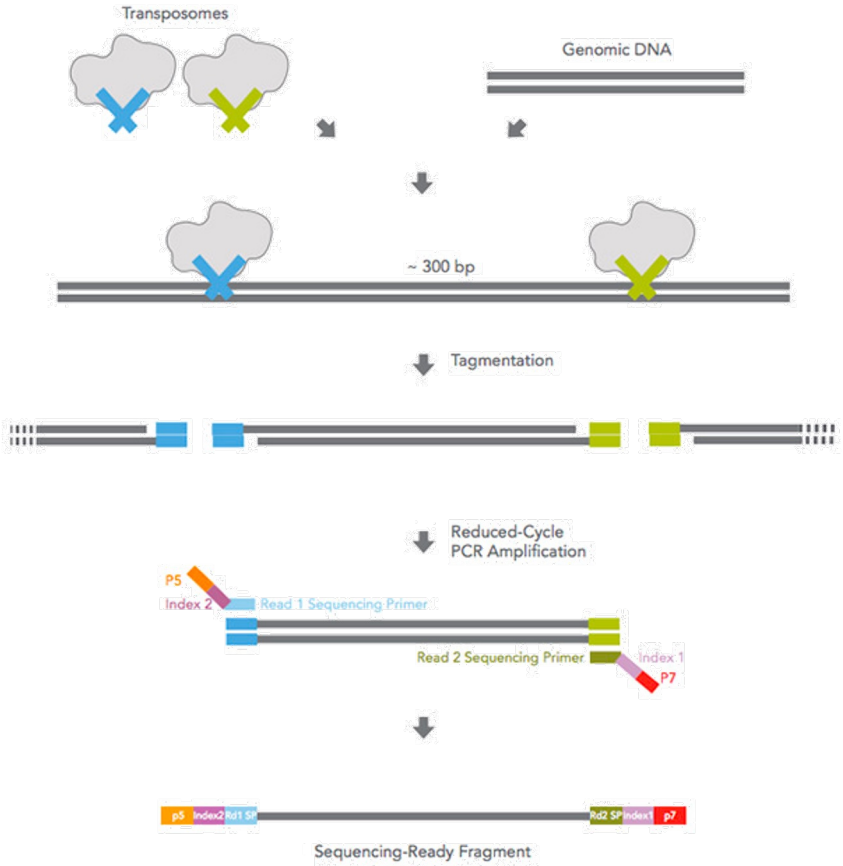
- Contamination risks become critical
- Blank controls

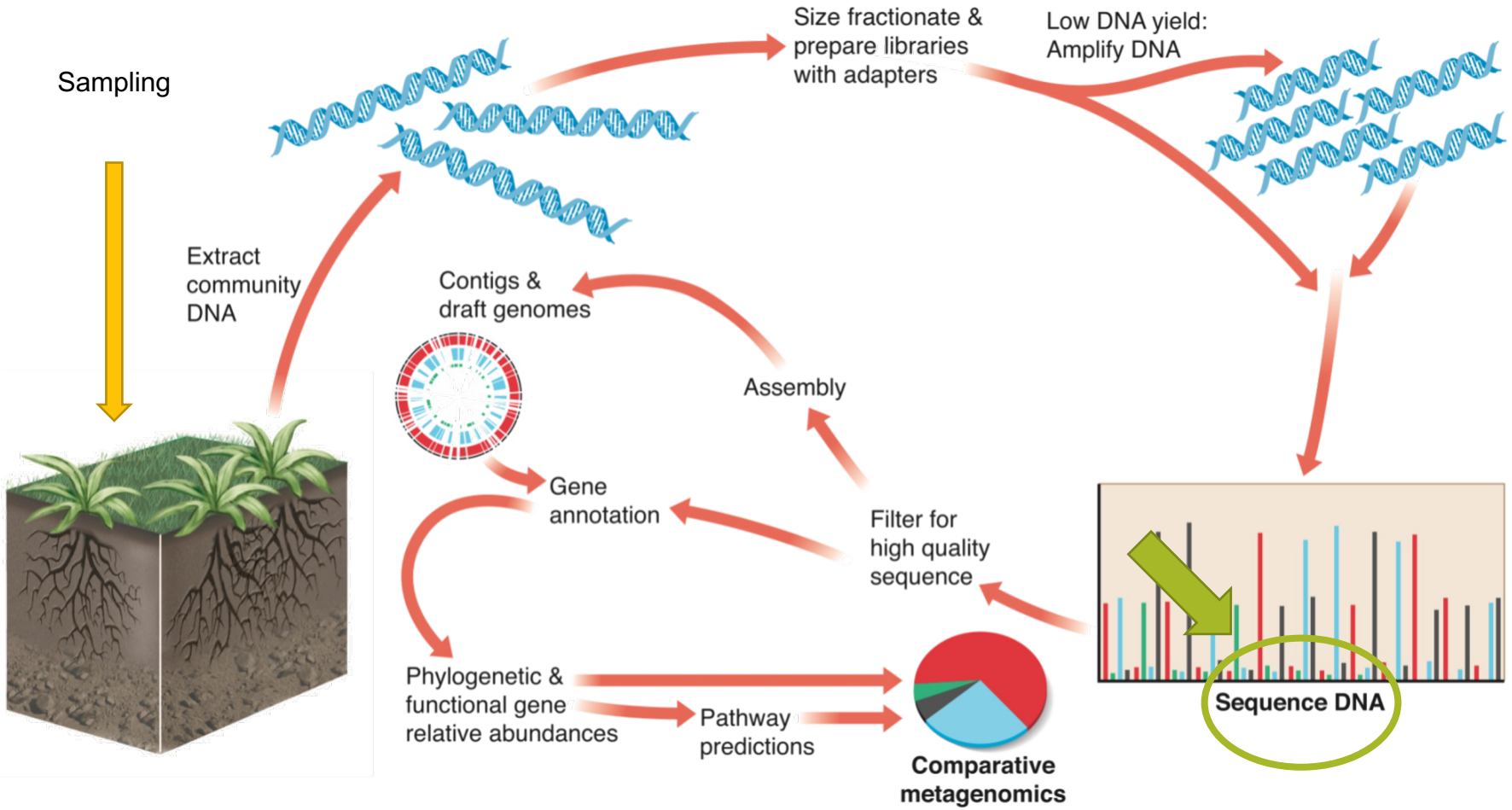

Sampling
Extract community DNA
Size fractionate & prepare libraries with adapters
Low DNA yield: Amplify DNA
Contigs & draft genomes
Assembly
Gene annotation
Filter for high quality sequence
Sequence DNA
Phylogenetic & functional gene relative abundances
Pathway predictions
Comparative metagenomics

Library preparation

Size selection

Ligation of sequencing adaptors

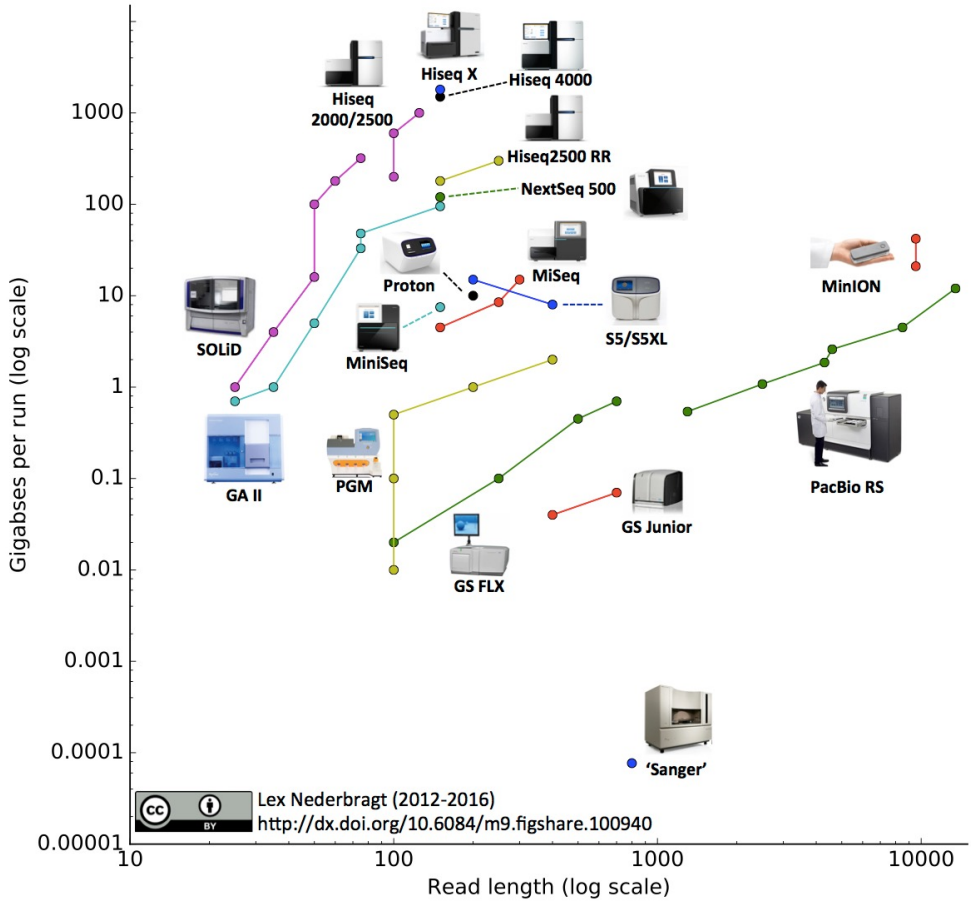



Sampling
Extract community DNA
Size fractionate & prepare libraries with adapters
Low DNA yield: Amplify DNA
Contigs & draft genomes
Assembly
Gene annotation
Filter for high quality sequence
Sequence DNA
Phylogenetic & functional gene relative abundances
Pathway predictions
Comparative metagenomics

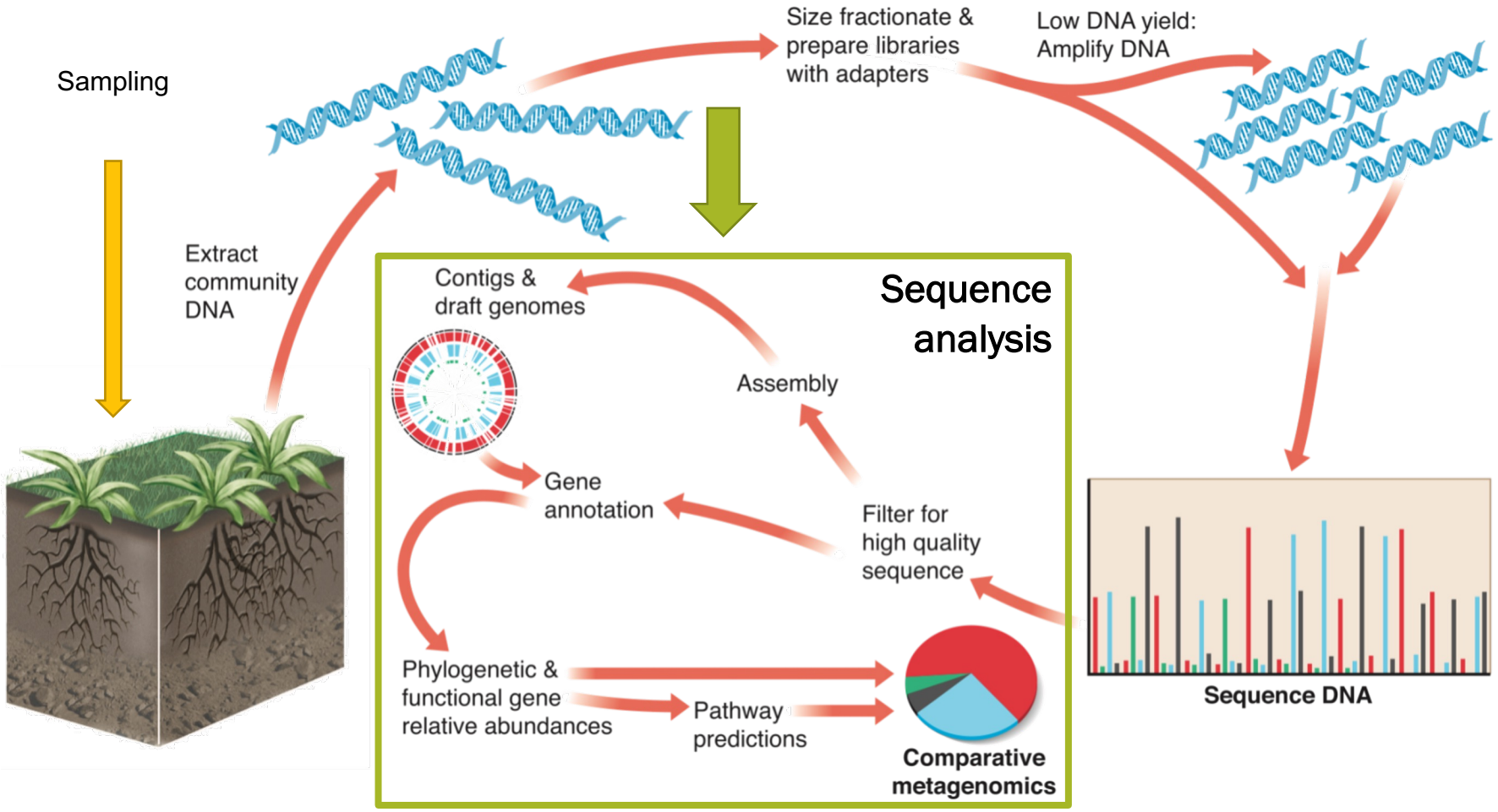
Sequencing technologies

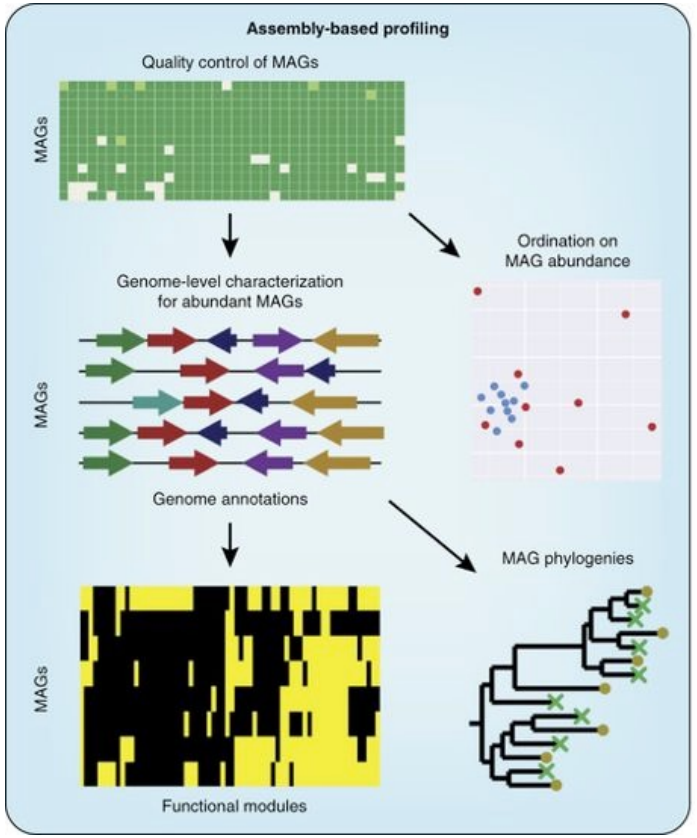
Important features

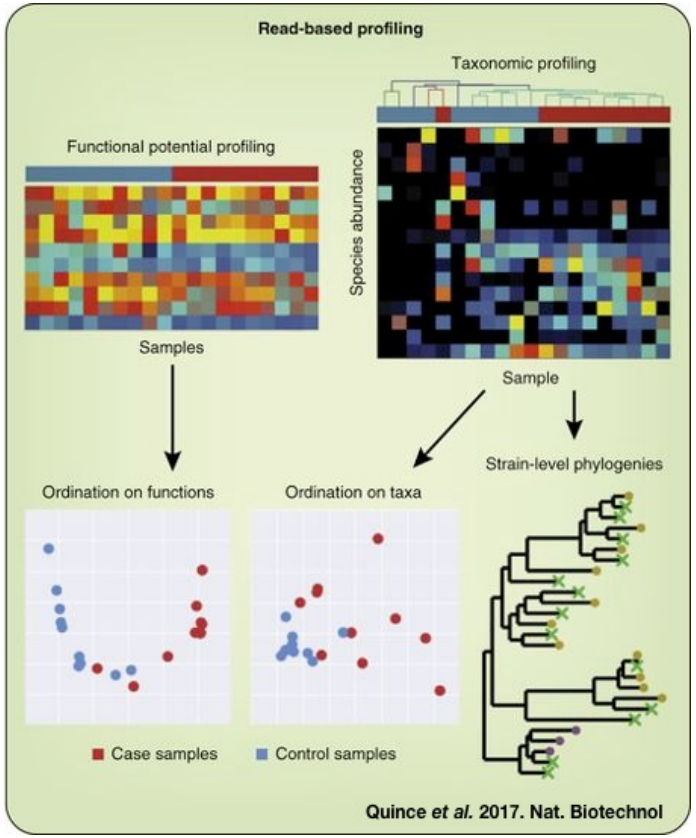
- Read size
- Read depth
- Error rate
- Price



doi.org/10.6084/m9.figshare.100940


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…

Profiling of taxa and functions

- Environmental gradients
- Time series
- Controls vs. treatments

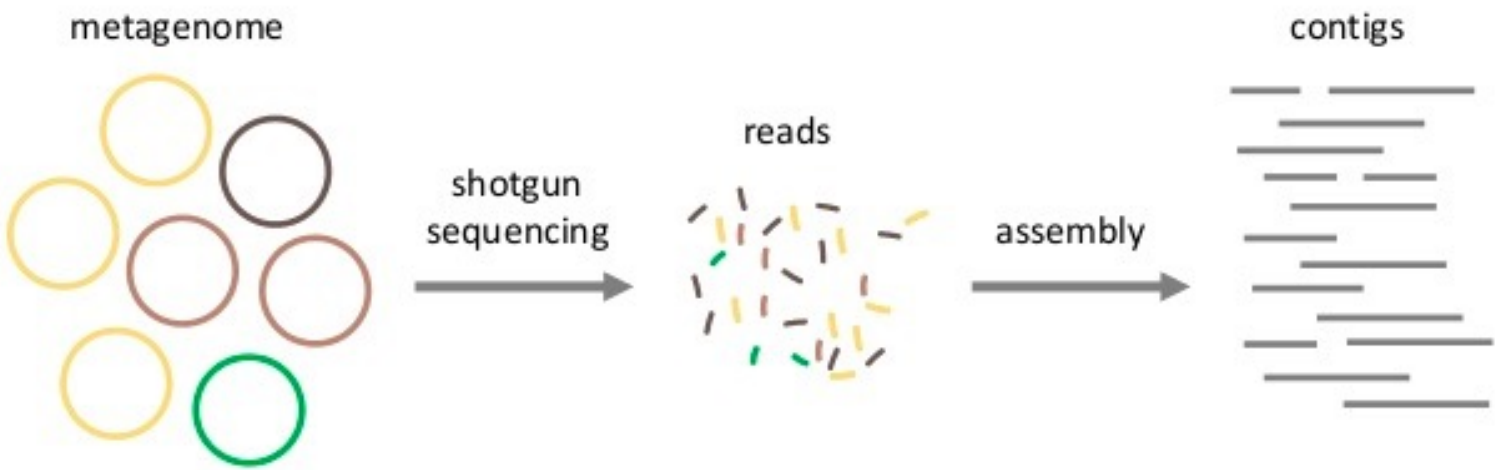
Strenghts and weaknesses of read-based metagenomics

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

doi.org/10.1038/nbt.3935

Assembly-based metagenomics

Reads are assembled into larger contiguous segments (contigs)

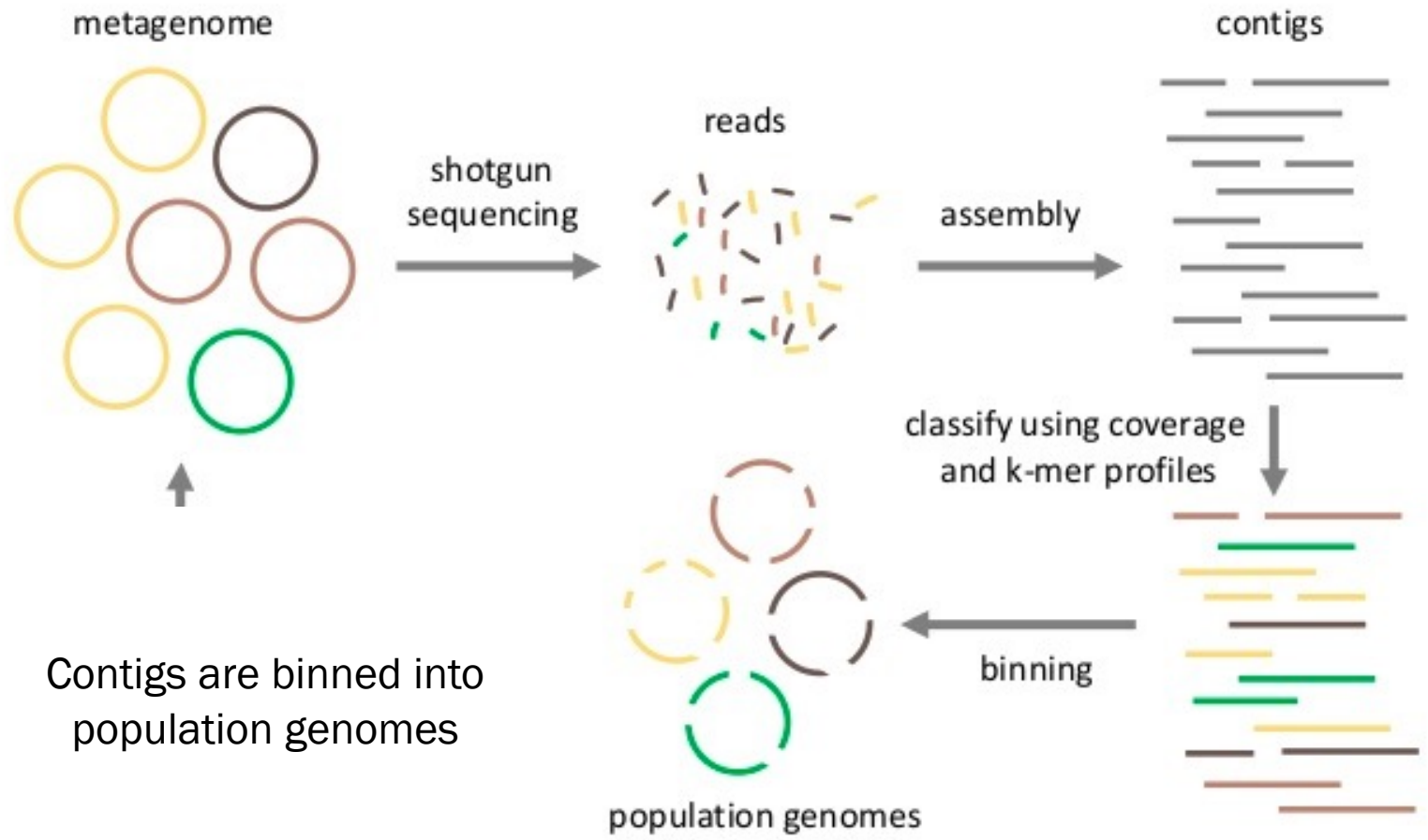


Assembly-based metagenomics

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.:

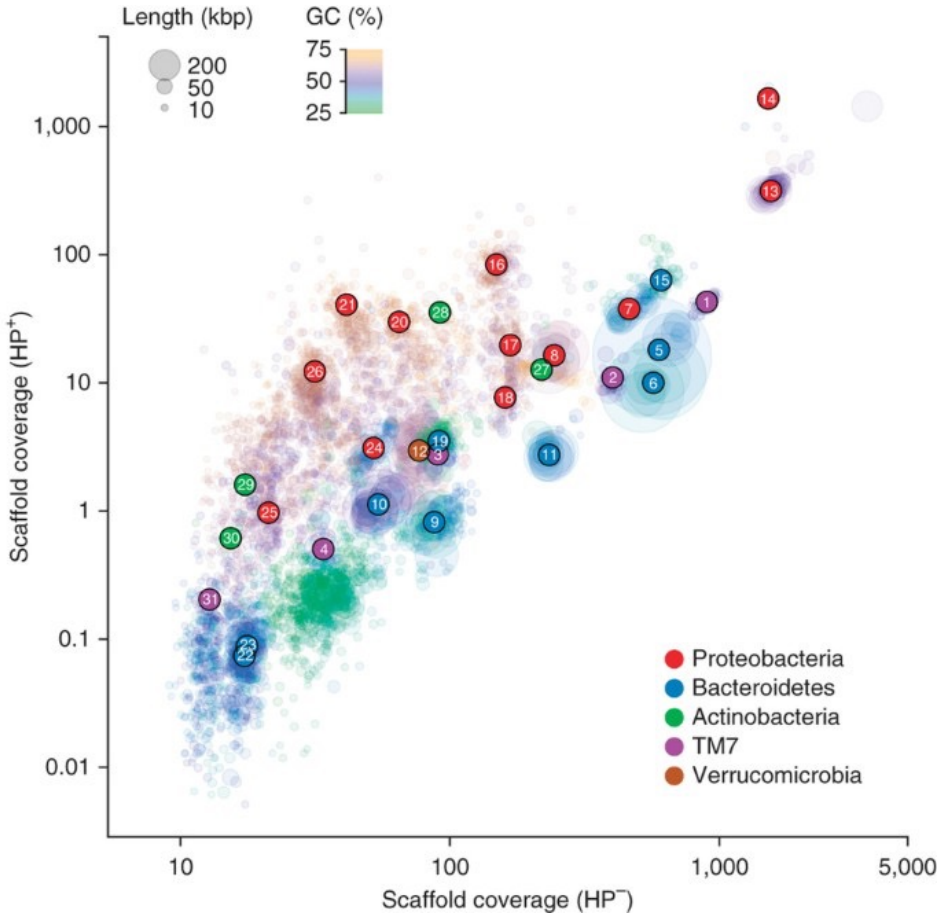
- Coverage
- Tetranucleotide frequency

Automated

- CONCOCT, metaBAT, etc.

Manual

- Anvi'o



Manual binning with anvi'o

PeerJ

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

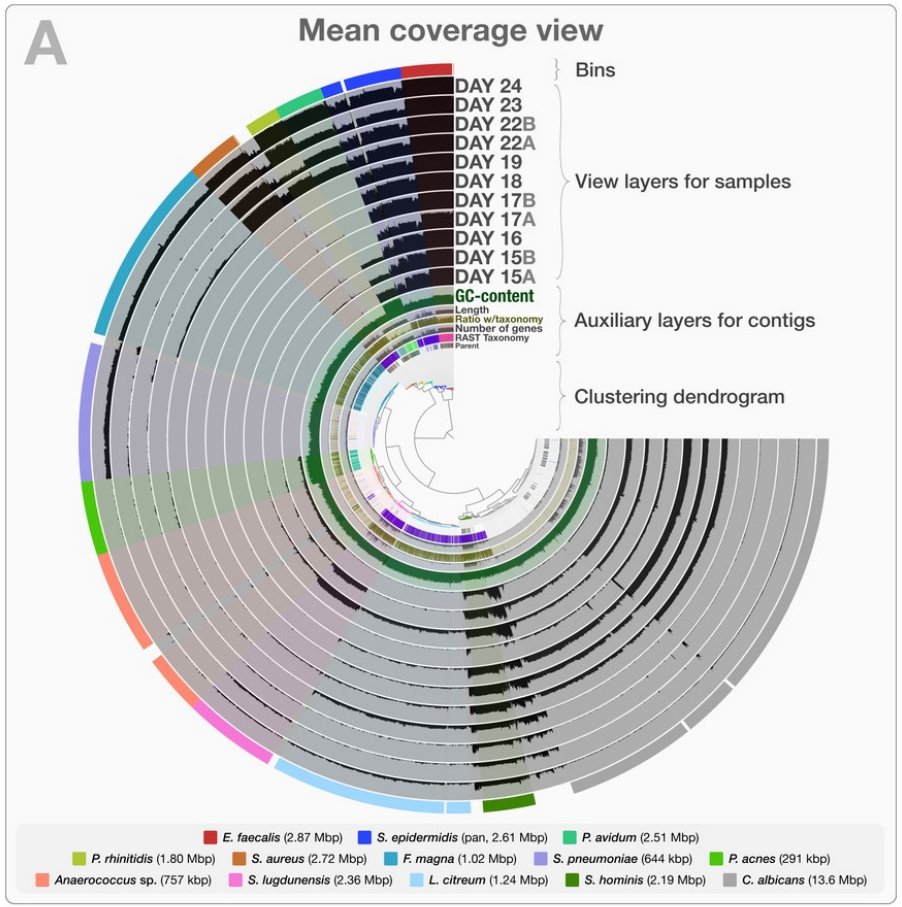
A. Murat Eren[1,2], Özcan C. Esen[1], Christopher Quince[3], Joseph H. Vineis[1], Hilary G. Morrison[1], Mitchell L. Sogin[1] and Tom O. Delmont[1]

[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



Strengths 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

doi.org/10.1038/nbt.3935

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