Metagenomics

Luke 2022

Jenni Hultman

Learning goals

- Foundational skills to work with metagenomic data
- Familiarity and practice with bioinformatics tools
- Perspective and confidence to apply these skills in your own work
- Empower you to ask and answer the questions you have of your own data

https://aindustryreports.com/2019/07/23/metagenomics-market/

This course

- Hands-on
- Materials available during and after the course
  - Github
- Mix of lectures, tutorials and practice
- Ask questions but if there is an error read the error message first
- Learn from each other as well as instructors

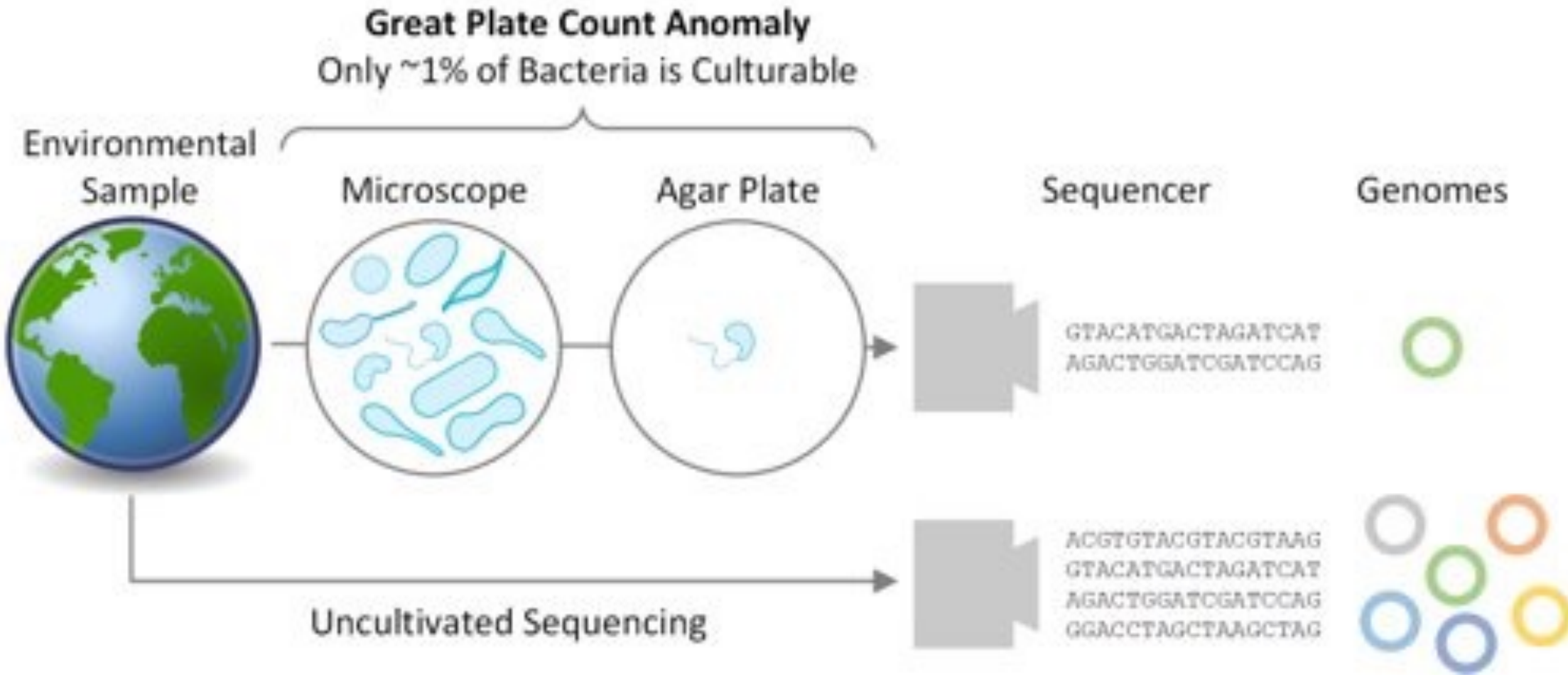
Material

- Course material will be available at GitHub
  - Constructed from material from various MBDP courses by Jenni Hultman, Antti Karkman and Igor S. Pessi
  - Free to use, copy and modify, GNU General Public License v3.0

- Part one on read based analysis of metagenomes, part two on assembly based methods

You learn programming by programming

Various authors at CSC


Great Plate Count Anomaly
Only ~1% of Bacteria is Culturable
Environmental Sample
Microscope
Agar Plate
Sequencer
Genomes
GTACATGACTAGATCAT
AGACTGGATCGATCCAG
ACGTGTACGTACGTAAG
GTACATGACTAGATCAT
AGACTGGATCGATCCAG
GGACCTAGCTAAGCTAG
Uncultivated Sequencing

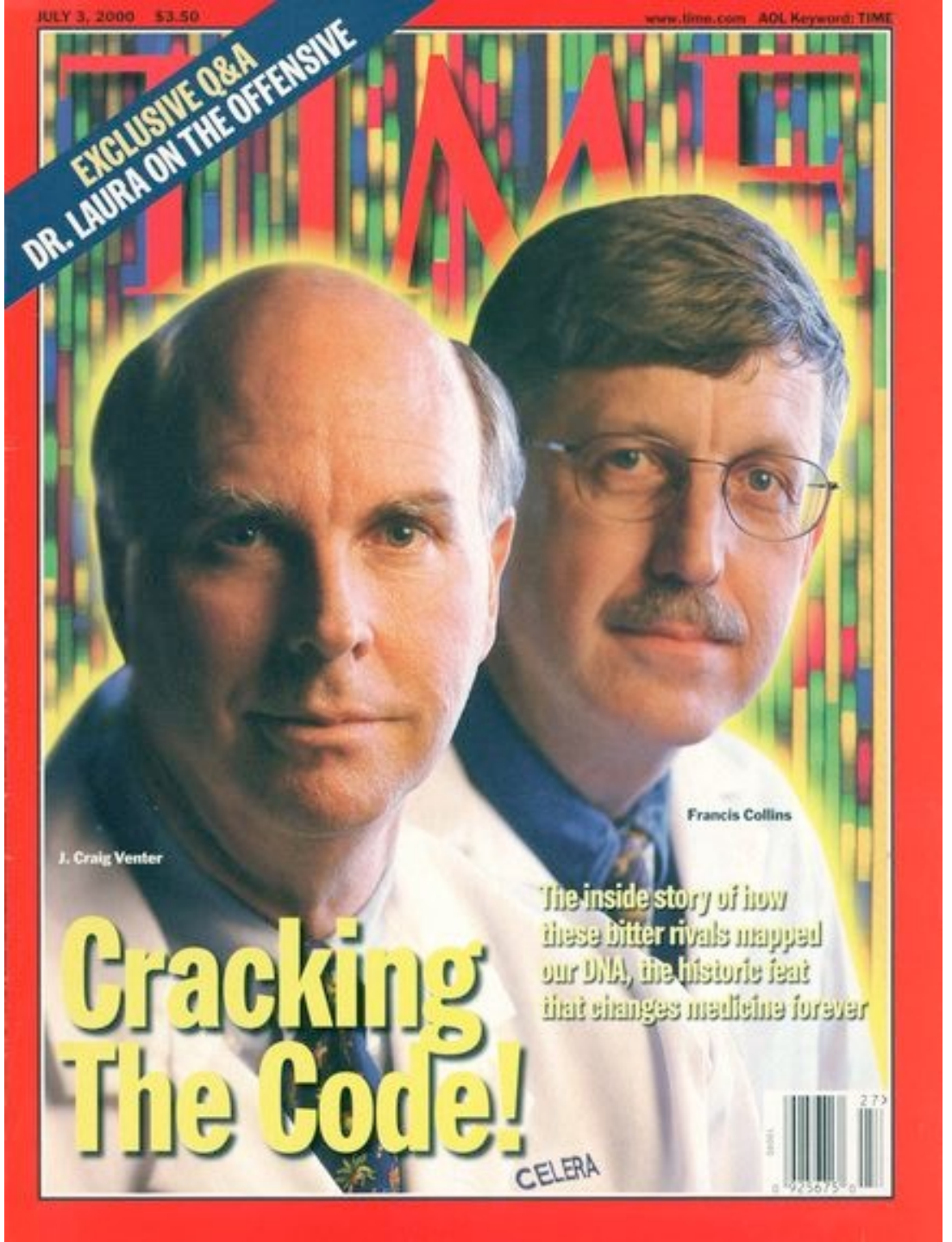
Metagenomics

- Jo Handelsman 1998

”the application of modern genomics techniques to the study of communities of microbial **organisms directly in their natural environments**, bypassing the need for isolation and lab cultivation of individual species”

H CS

- 19 ea of clo env
- 19 A ge
- 20 sec rent vir
- 20 me



RESEARCH ARTICLE

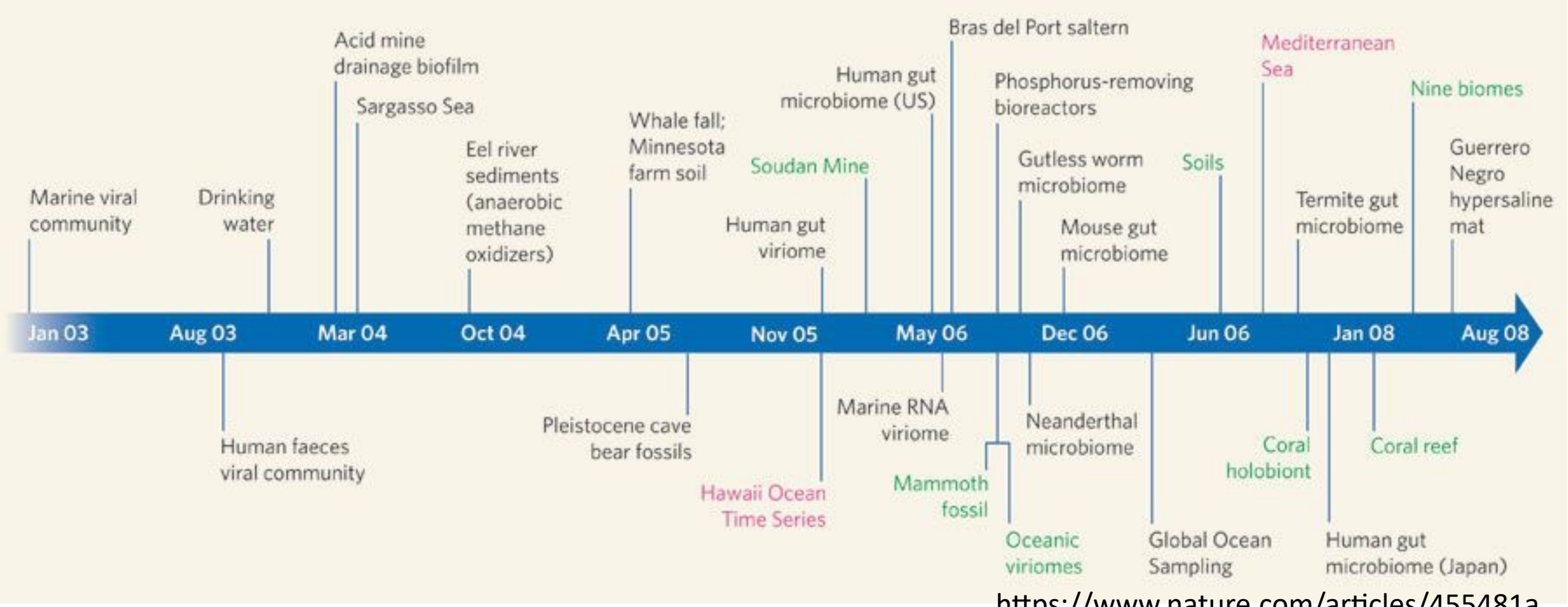
Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1*] Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

We have applied "whole-genome shotgun sequencing" to microbial populations collected en masse on tangential flow and impact filters from seawater samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. We have identified over 1.2 million previously unknown genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.

- 148 previously unknown bacterial phylotypes
- 1.2 million previously unknown genes
  - 782 new rhodopsin like genes

Timeline of sequence-based metagenomic projects showing the variety of environments sampled since 2002

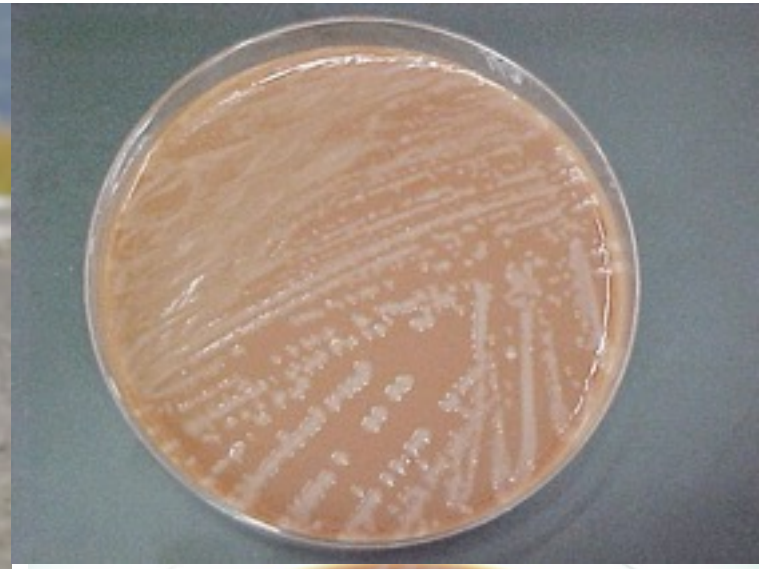


https://www.nature.com/articles/455481a

Metagenomes and microbes

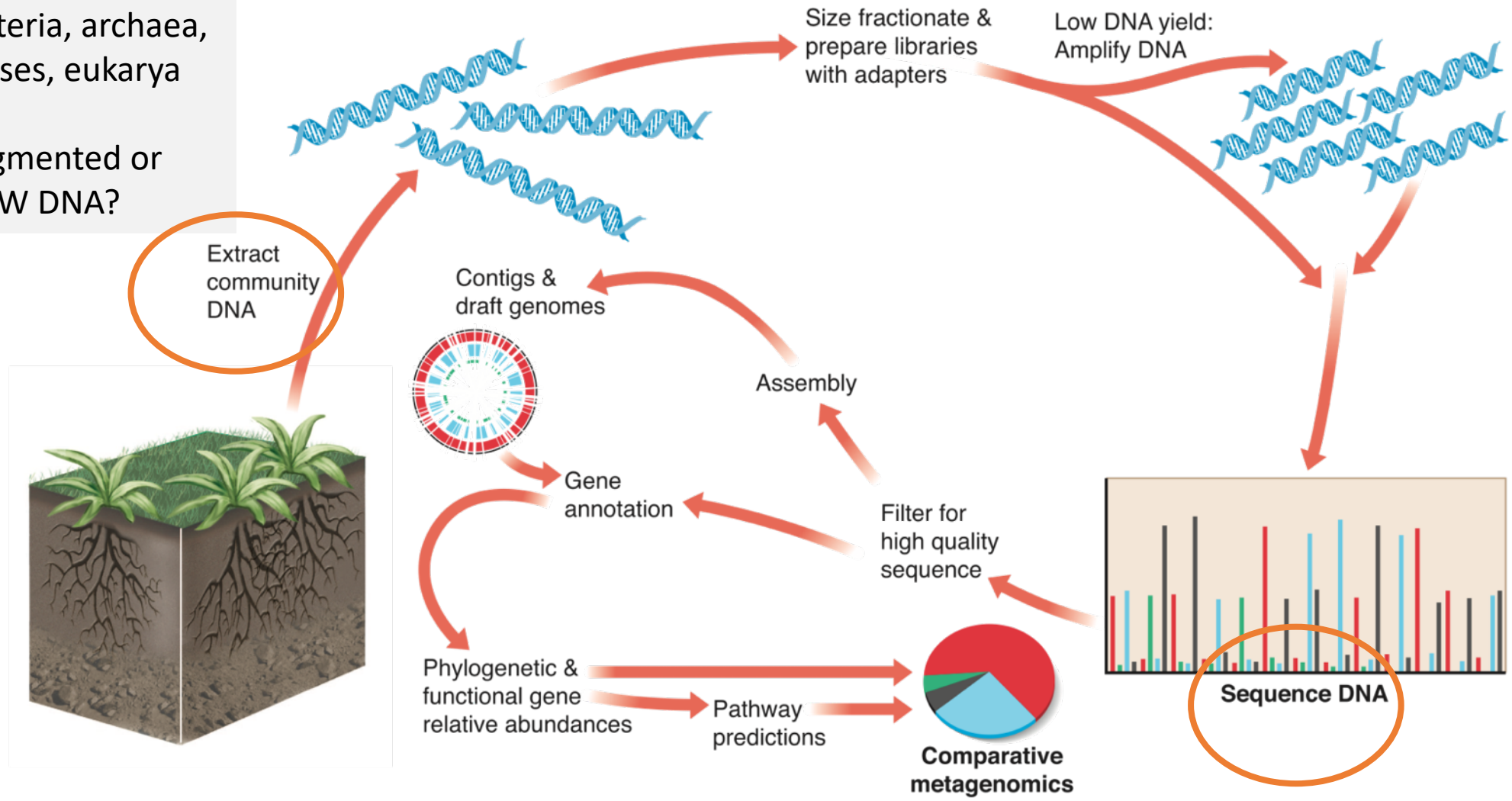
- Often metagenome is mostly microbial (Bacteria, Archaea, Fungi, some protist)
- Why metagenomes are high in microbes?
- Microbes are hard to study: small, diversity and numbers are high
- Phenotype

Metagenomes and microbes



DNA from all bacteria, archaea, viruses, eukarya

Fragmented or HMW DNA?



Soil metagenomics methodology.

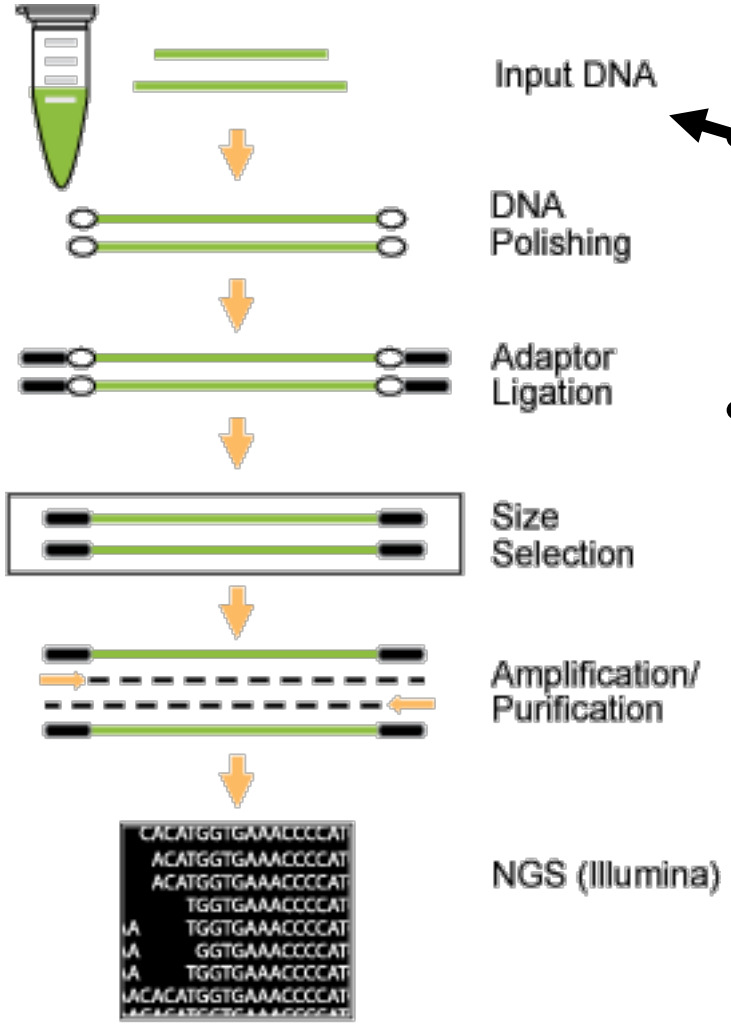
Illumina, PacBio, Oxford Nanopore, Ion Torrent

Paired end, single read, size of library

Sequencing depth, replicates

Jansson 2011: Towards tera-terra. Microbe June 2011

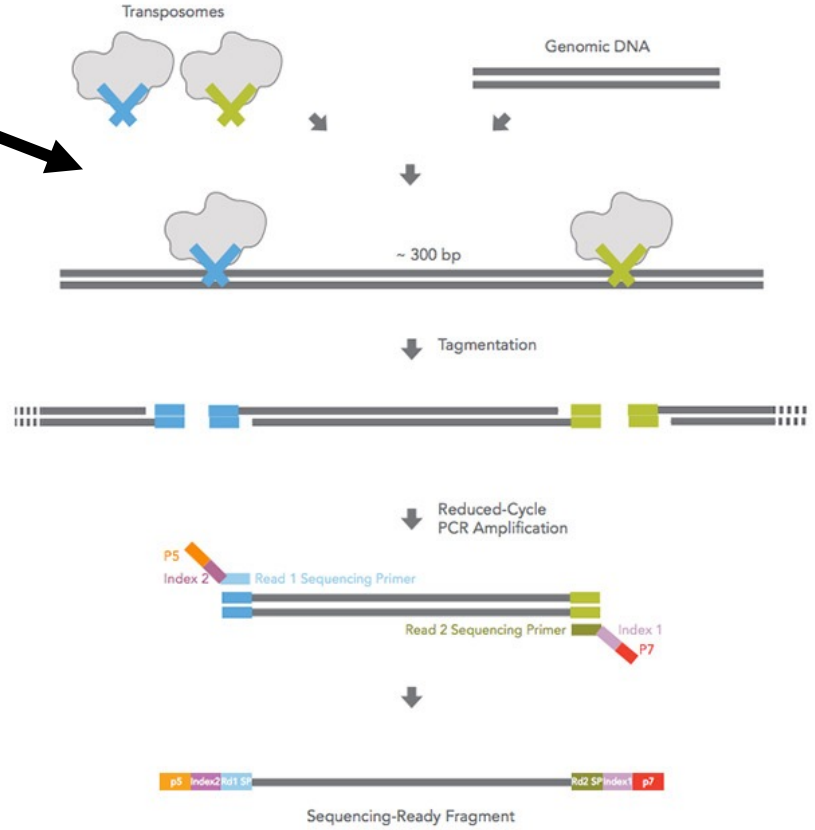
Library preparation

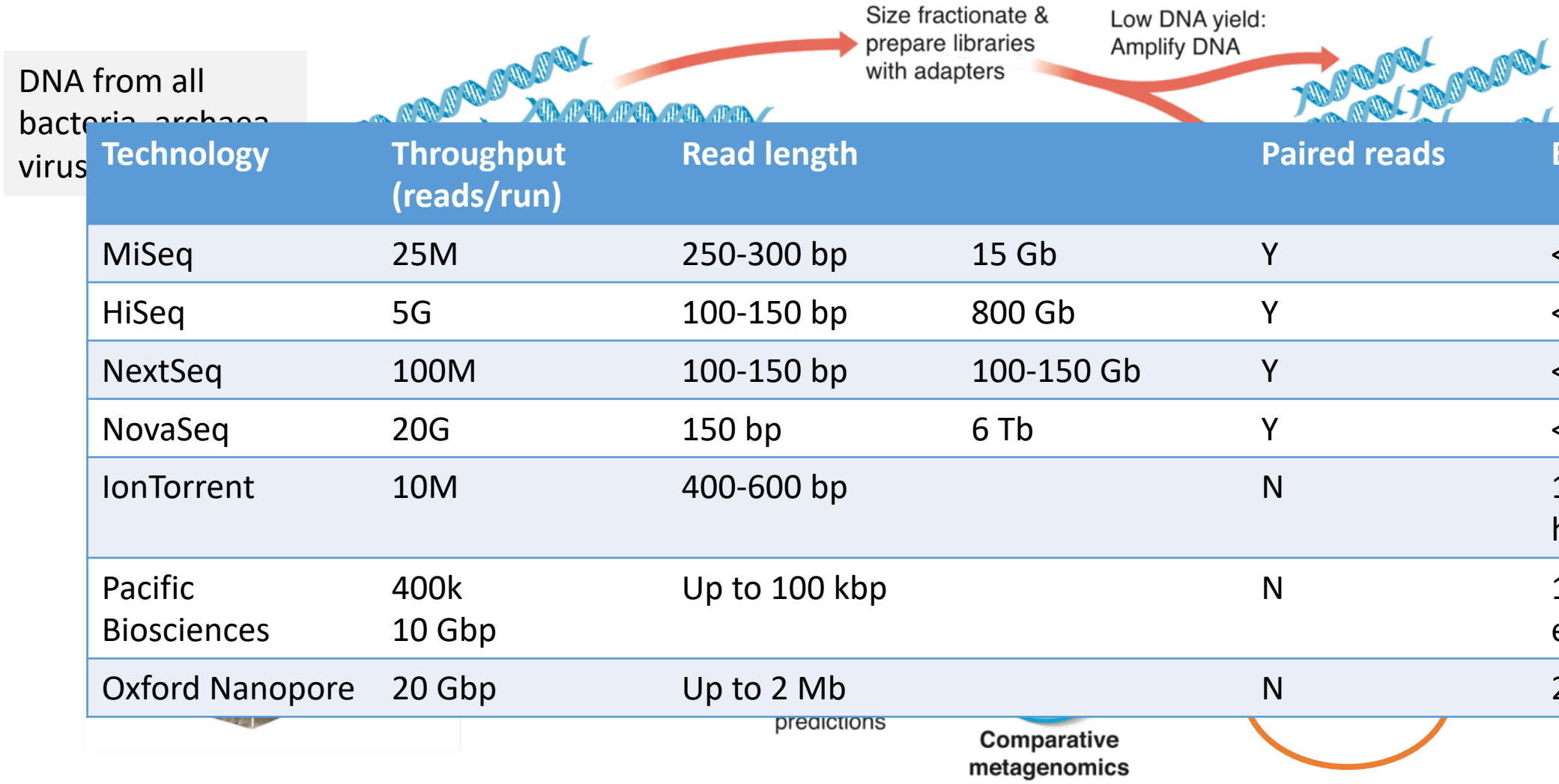


DNA shearing and adapter ligation

- DNA quantity

- **Nextera transposases**
  - Inhibitors (in soil)
  - 1-10 ng of DNA
  - Transposase:DNA ratio



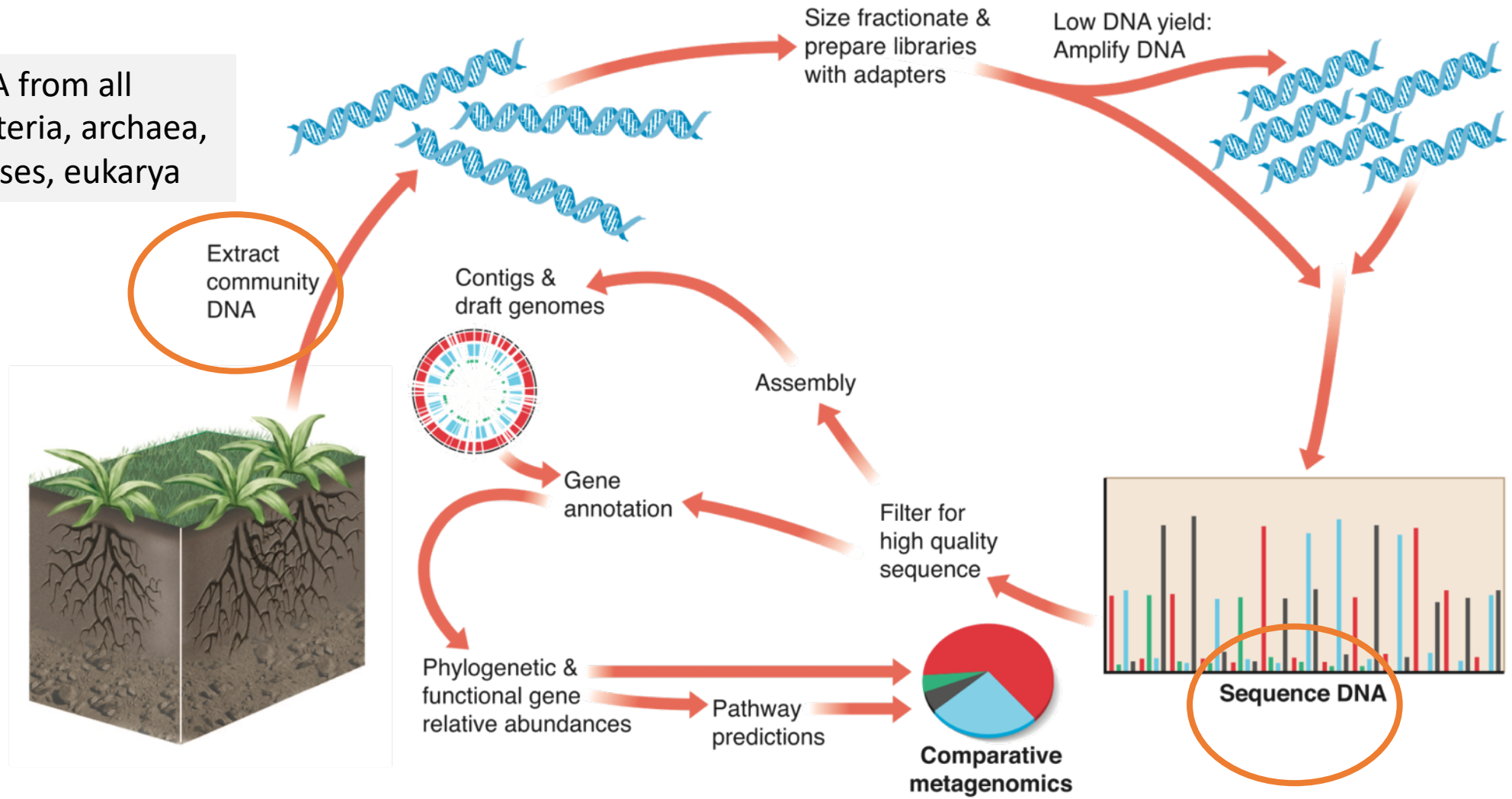


Technology	Throughput (reads/run)	Read length		Paired reads	Errors?
MiSeq	25M	250-300 bp	15 Gb	Y	<1%
HiSeq	5G	100-150 bp	800 Gb	Y	<1%
NextSeq	100M	100-150 bp	100-150 Gb	Y	<1%
NovaSeq	20G	150 bp	6 Tb	Y	<1%
IonTorrent	10M	400-600 bp		N	1-2% (indels, homopolymers)
Pacific Biosciences	400k 10 Gbp	Up to 100 kbp		N	14% (random error)
Oxford Nanopore	20 Gbp	Up to 2 Mb		N	2-13%

Soil metagenomics methodology.

Jansson 2011: Towards tera-terra. Microbe June 2011

DNA from all bacteria, archaea, viruses, eukarya



Soil metagenomics methodology.

Illumina, PacBio, Oxford Nanopore, Ion Torrent

Paired end, single read, size of library

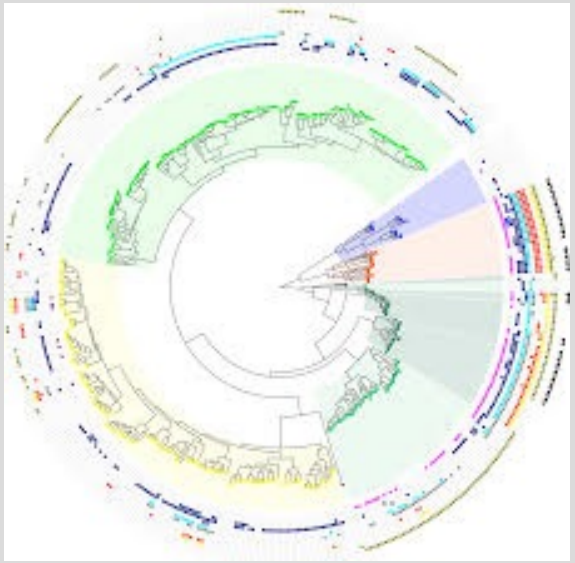
Sequencing depth, replicates

Jansson 2011: Towards tera-terra. Microbe June 2011

With metagenomics

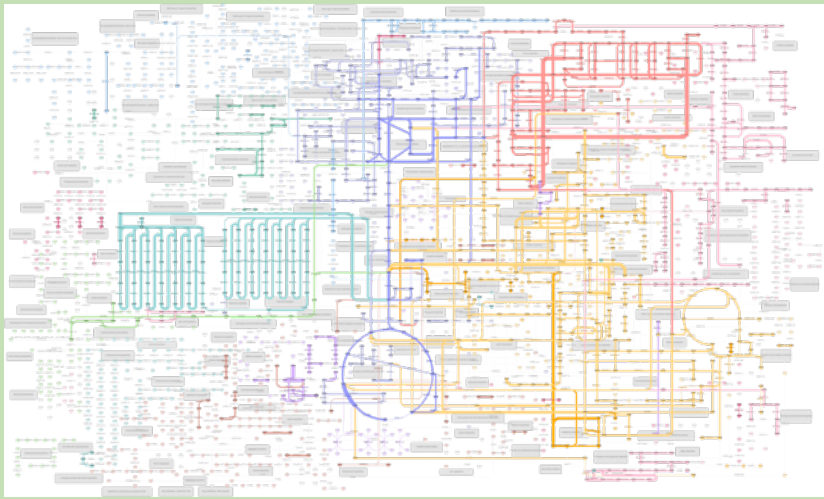
- Microbial community composition – who is there

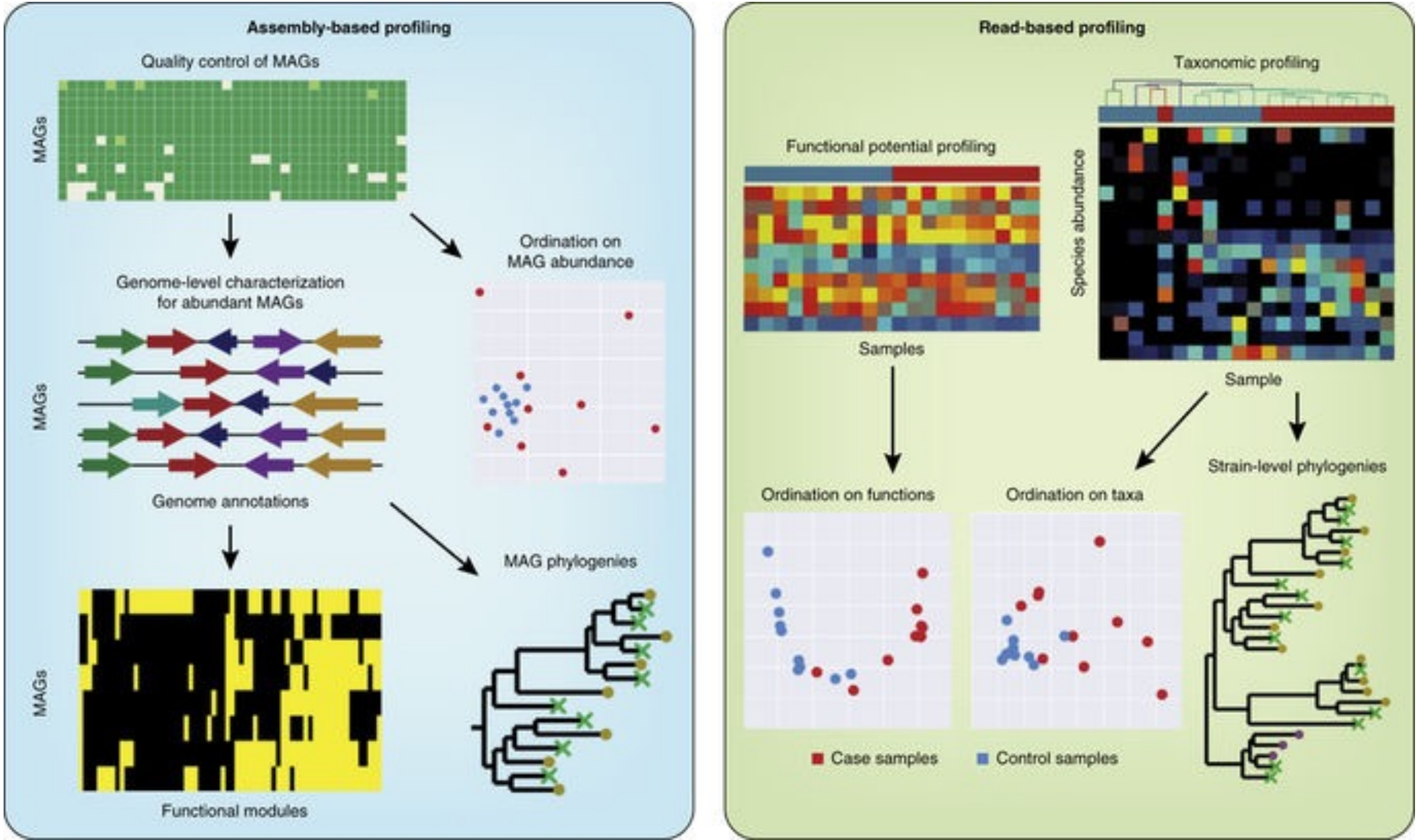
  16S/18S, ITS



- Microbial community function – what can they do

  Protein coding sequences

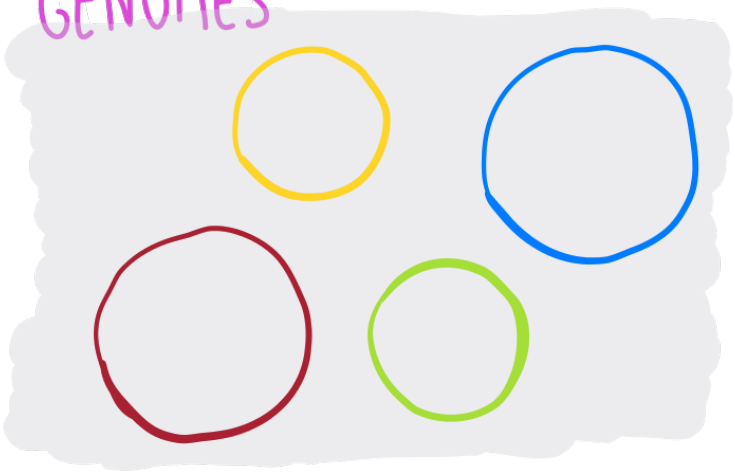


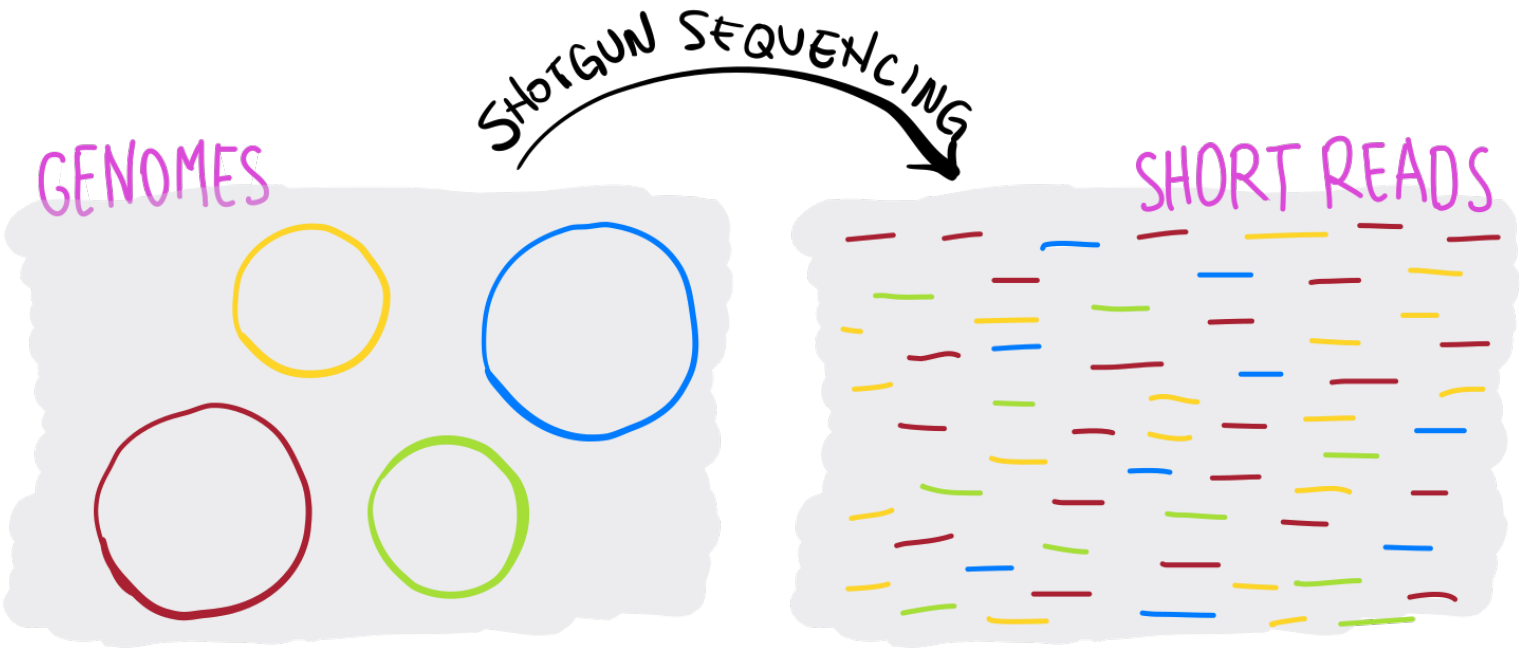

Assembly-based profiling
Quality control of MAGs
MAGs
Ordination on
MAG abundance
Genome-level characterization
for abundant MAGs
MAGs
Genome annotations
MAG phylogenies
MAGs
Functional modules
Read-based profiling
Taxonomic profiling
Functional potential profiling
Species abundance
Samples
Sample
Strain-level phylogenies
Ordination on functions
Ordination on taxa
Case samples
Control samples

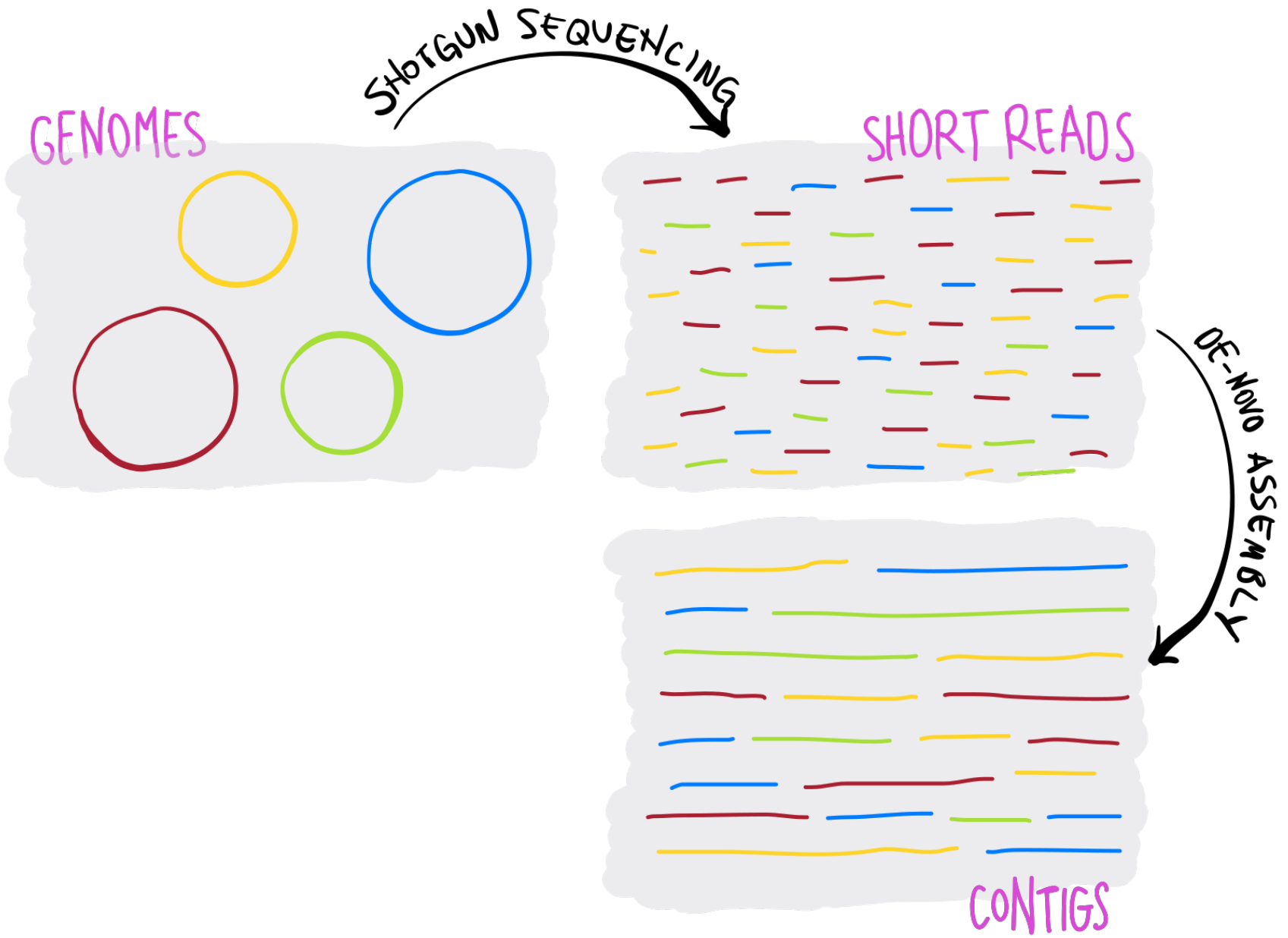
Reconstructing genomes from metagenomes

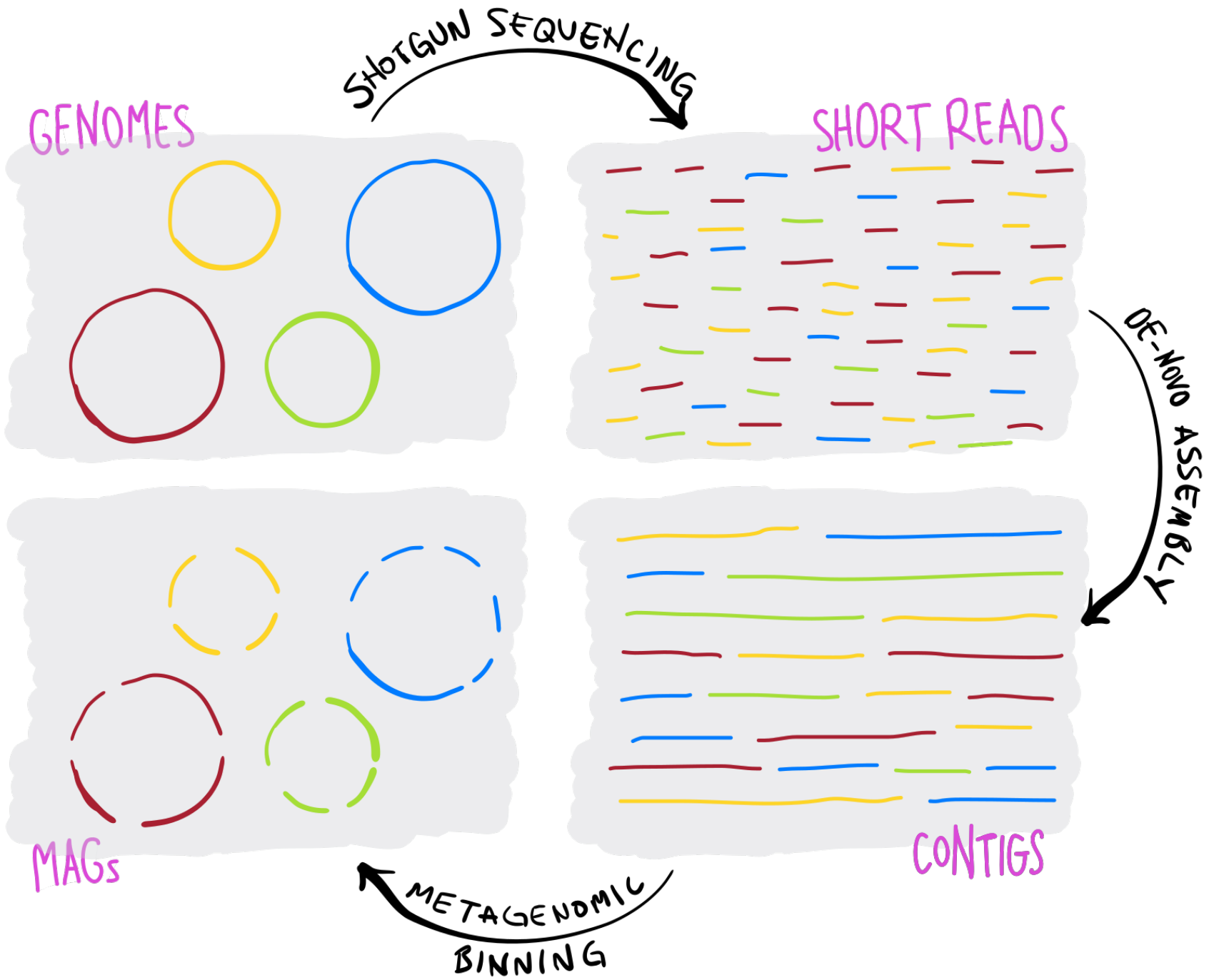
http://merenlab.org/momics

GENOMES




SHOTGUN SEQUENCING
GENOMES
SHORT READS

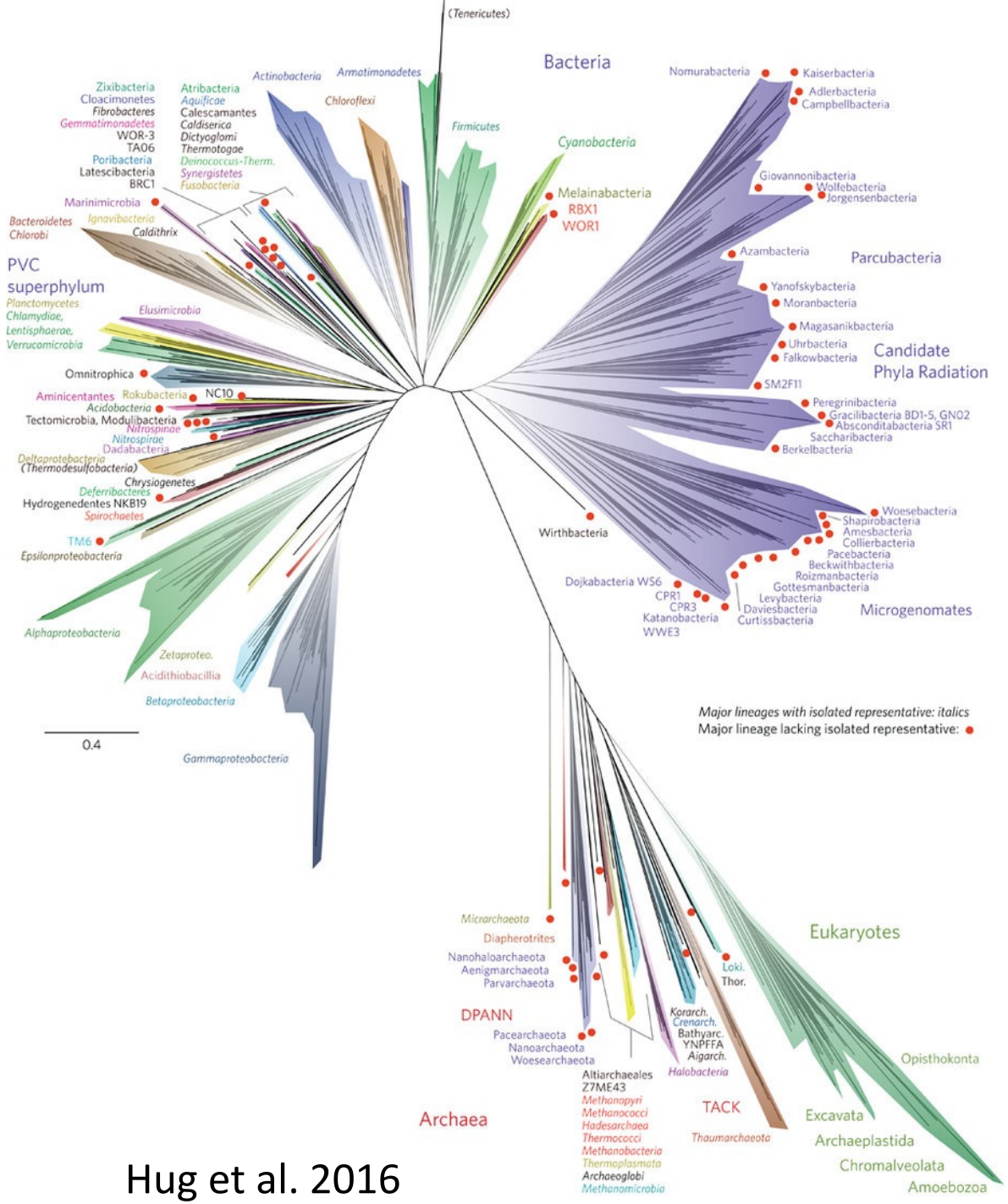

SHOTGUN SEQUENCING
GENOMES
SHORT READS
DE-NOVO ASSEMBLY
CONTIGS



Same community without cultivation of the members

Omics

- Metagenomics
- Metatranscriptomics
- Metaproteomics
- Meta-metabolomics

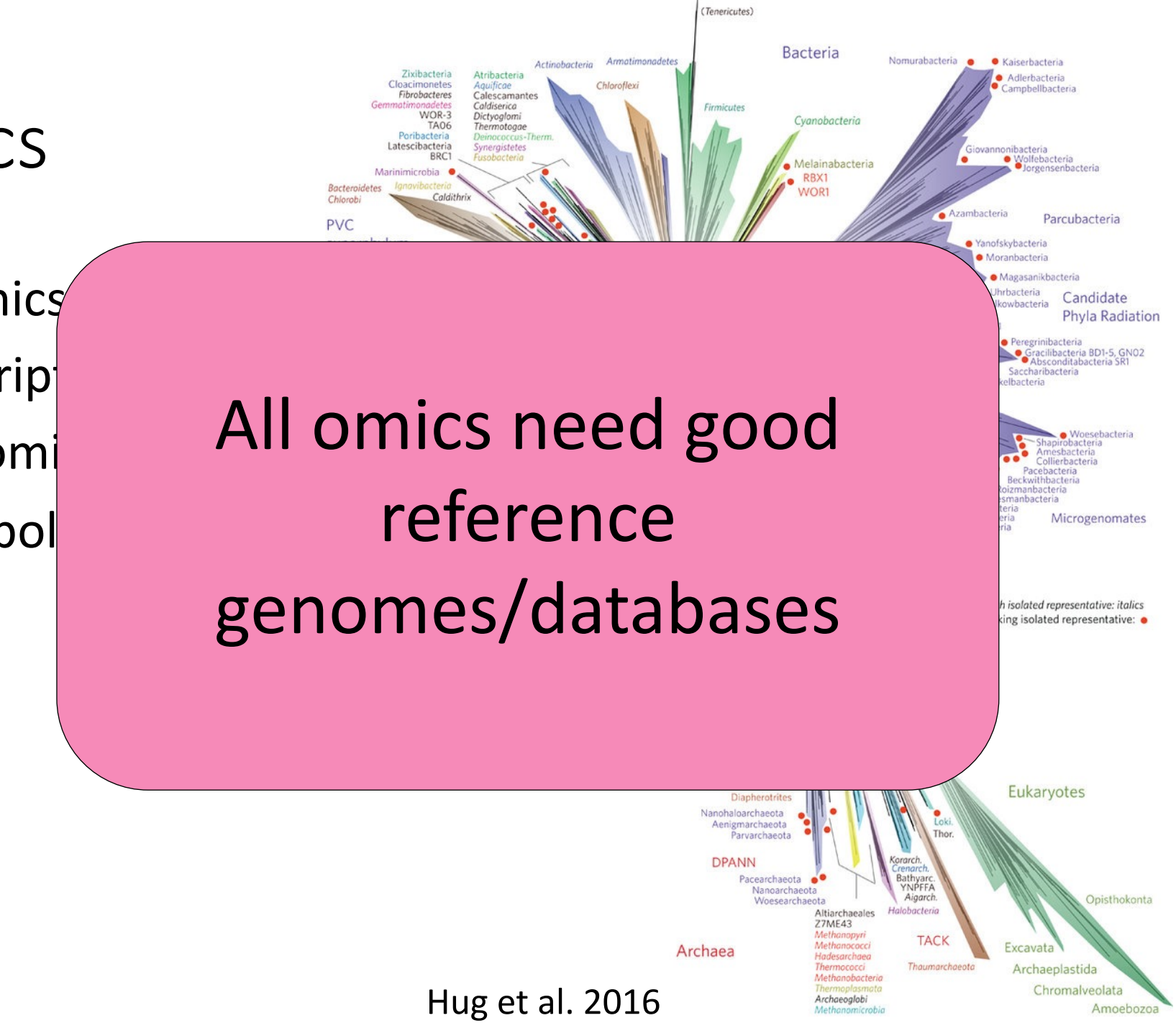


Hug et al. 2016

Omics

- Metagenomics
- Metatranscript
- Metaproteomi
- Meta-metabol

All omics need good reference genomes/databases



Hug et al. 2016

Overview of generic* metagenomics workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → fastq files → demultiplex (split samples by barodes) → quality filter/trim (remove adapters/**primers**) → fasta files

might be done by sequencing facility

fastqc/multiqc

@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

fasta files → read-based analysis (no-assembly path)

fasta files → assembly path → consider testing assemblies with and w/o → digital normalization

Some tools:
- bbnorm
- diginorm

Some tools:
- TIPP/SEPP
- metaphlan2
- humann2
- sourmash
- kraken

Count Table

	Sample_A	Sample_B	...
obj_1	0	428	...
obj_2	306	323	...
obj_3	217	1	...
...	...	...	...

MetaQUAST is a great tool for comparing assemblies

Analysis

Some tools:
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpiecEasi
- MaAsLin
- DESeq2

(co)-assembly

map individual sample reads to (co)-assembly → Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Recovering genomes from metagenomes → Gene calling Functional/taxonomic profiling

A note on MAGs:
MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

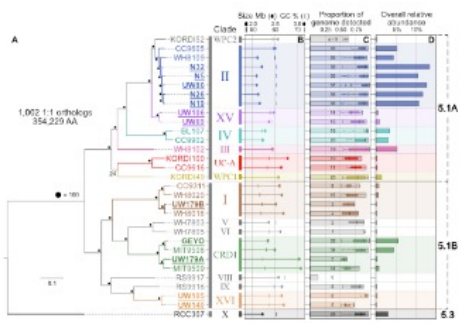
Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (primarily coverage-based, optional second round kmer-based binning tool)
- checkm (genome-level taxonomy; and much more)
- DASTool (a tool for evaluating bins recovered by different methods)
- DESMAN (tool aimed at resolving strains)

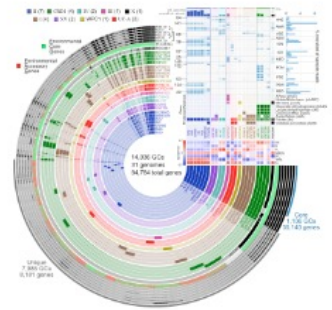
Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

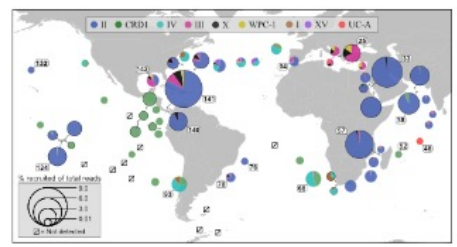
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- StrainPhlAn/PanPhlAn (tools for strain-level analyses)
- MUSCLE (alignment software)
- FastTree (very fast, pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignment)







astrobiomike.github.io

CSC greetings

- You all have CSC account with 1000 billing units
  - But not project where to do more intensive computing
  - You can run out of billing units
    - **`saldo`**, should not be negative
- For this course we have a project Metagenomics2019, Jenni will add you
- Make sure that when you work with **real data** you have **a PI who has a project** with enough billing units and you are member of that project

Some basic Unix commands

pwd: print working directory (“where am I?”)

ls: list (“show folder contents”)

mkdir: make directory (a.k.a. folder)

cd: change directory (“go to folder”)

cp: copy

mv: move

rm: remove

Some additional notes

Running commands

- Space after each “word” in the command
- Typed in a single line, one at a time
- After each command, hit “Enter” to execute it
- Lines starting with “#” are comments

Directory navigation:

- One dot (.) means “here”
- To go up one folder: ../
- To go up two folders: ../../

How to learn UNIX?

By using it!

- Trial and error
- Don’t copy and paste it, type it

Ask the internet

- http://stackoverflow.com/
- http://stackexchange.com/
- http://askubuntu.com/
- Google!

Cheat sheets

- https://www.guru99.com/linux-commands-cheat-sheet.html

Manual (“man”) pages

- `man ls`

Online courses/tutorials

- http://codecademy.com

Let’s start practicing!