

Metagenomics

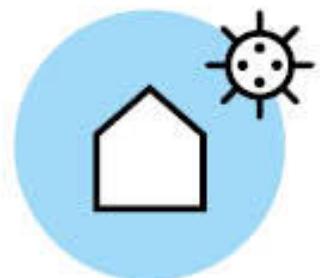
MBDP-102

Jenni Hultman, Antti Karkman, Igor Pessi

Covid specific rules

- Wear mask all the time in the lecture room and at UH premises
 - Provided by MBDP
- If you have any of the Covid symptoms/have been in contact with covid+ stay at home and get tested
 - We will organize possibility to follow teaching from home
- Unicafe is closed but Ladonlukko is open
- Do not move tables or chairs

HOW TO PREVENT THE SPREAD OF CORONAVIRUS IN THE UNIVERSITY COMMUNITY



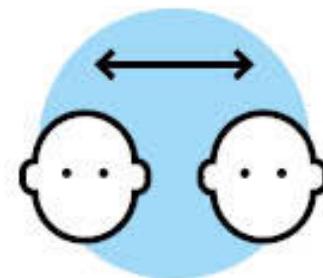
Stay at home
when
feeling ill.



Practise good
hand hygiene and
coughing etiquette.



Avoid
close
contact.



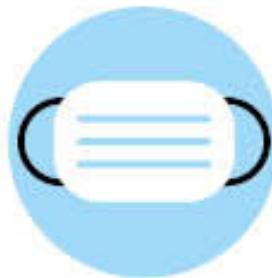
Maintain
safe
distances.



Do not
shake hands or
hug people.



Clean
work equipment
after use.



Use
a protective
mask.



Carefully
consider the need
for travel.

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



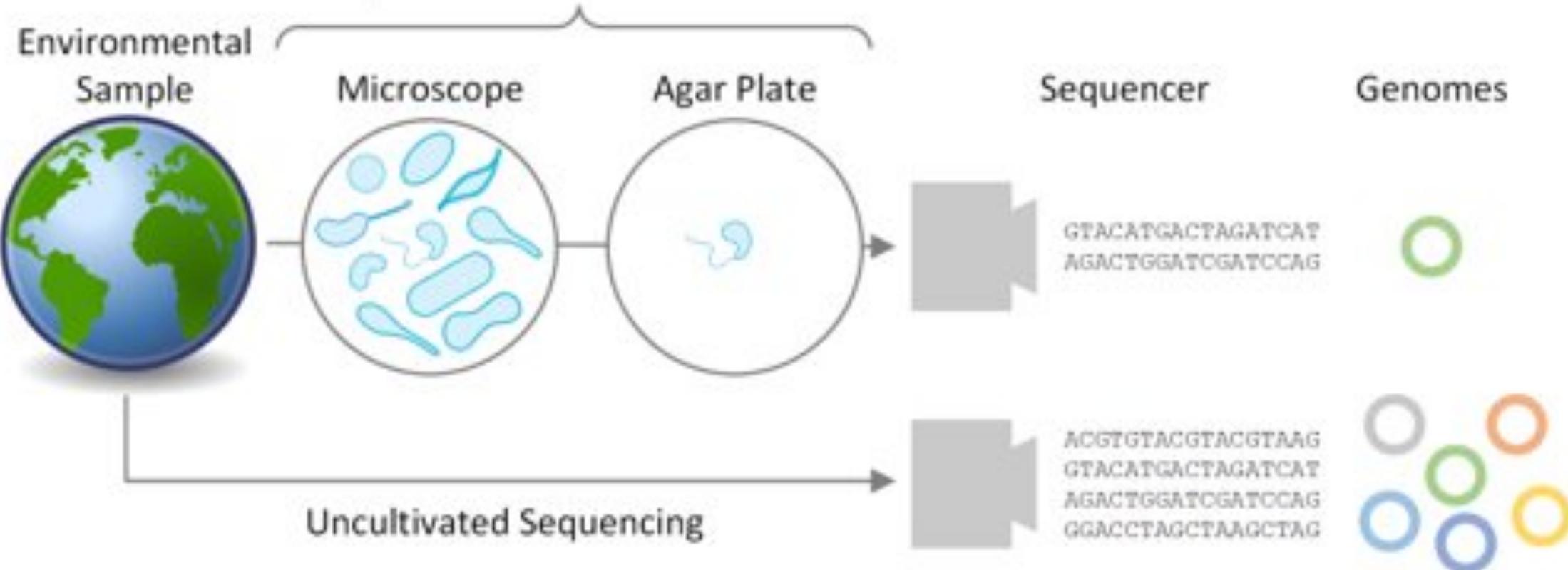
This course

- 
- 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

You learn programming by
programming

Various authors at CSC

Great Plate Count Anomaly
Only ~1% of Bacteria is Culturable



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

- 198
close
envir
- 199
geno
- 200
sec
virus
- 200
me

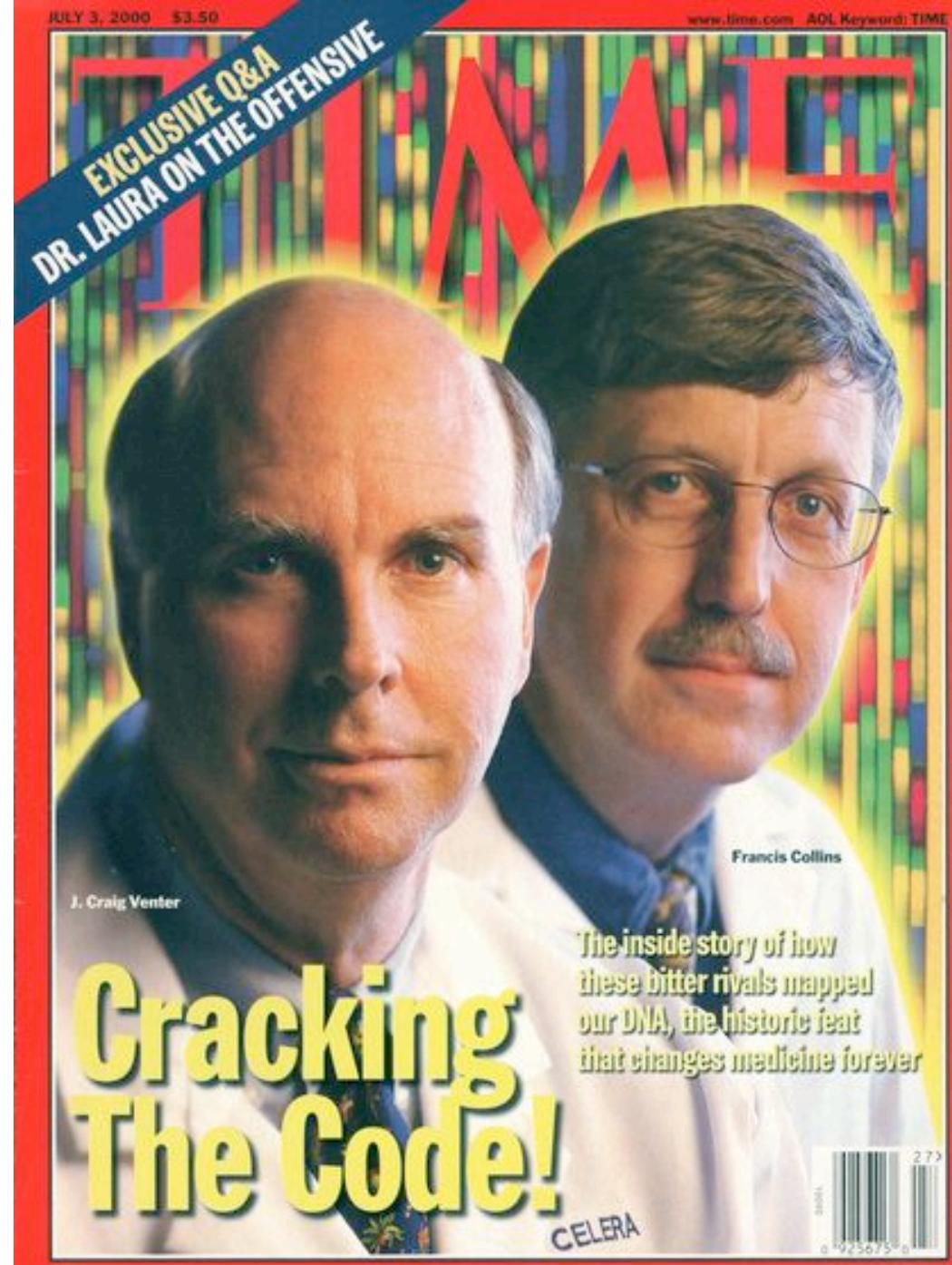


CS

ea of

A

rent



RESEARCH ARTICLE

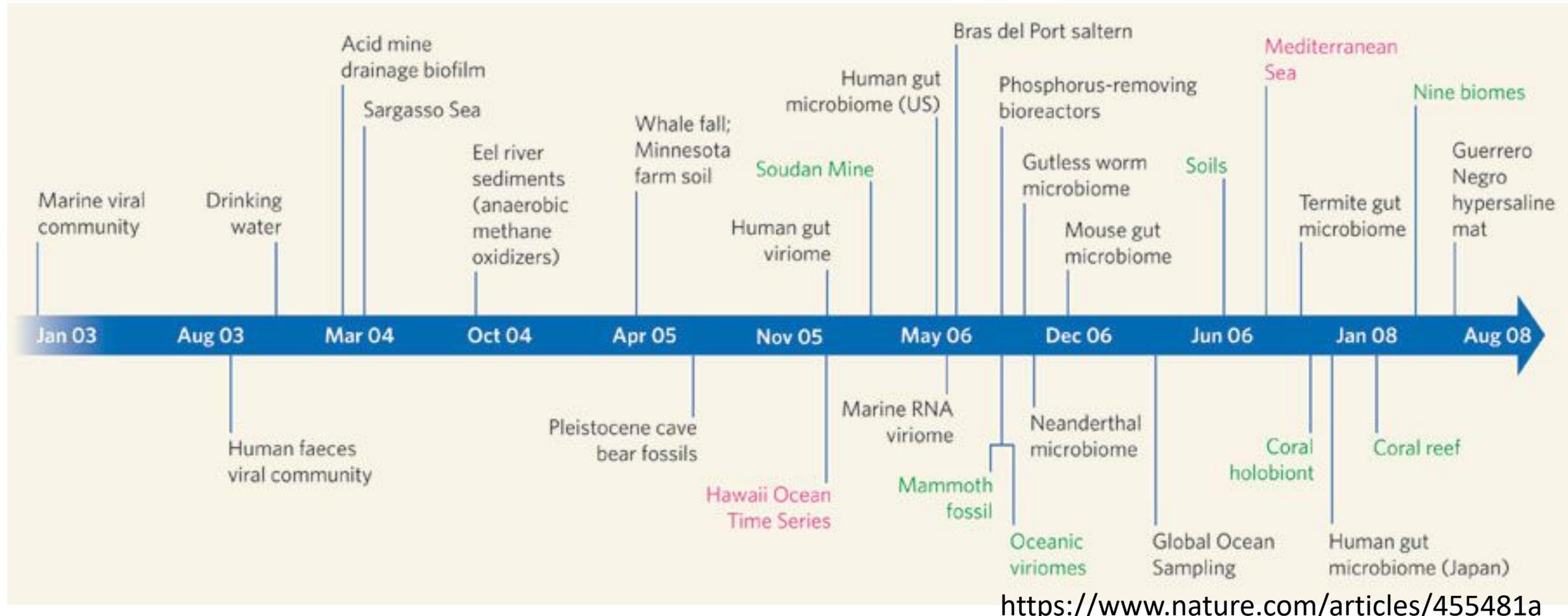
Environmental Genome Shotgun Sequencing of the Sargasso Sea

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. Smith¹

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



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

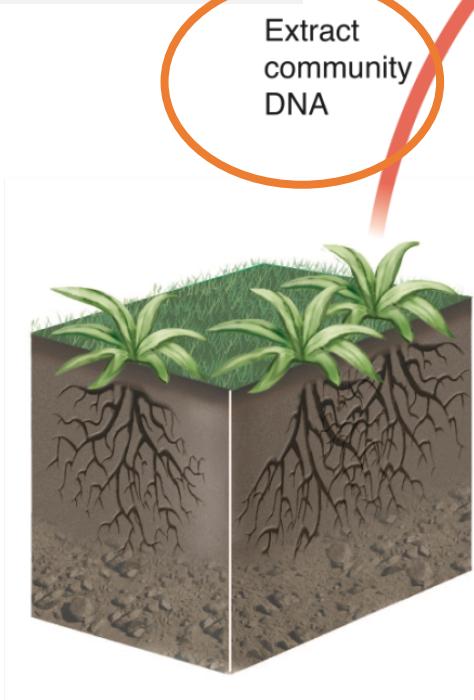
Metagenomes and microbes

Photo By Jim Floyd. Bear Viewing Client



DNA from all
bacteria, archaea,
viruses, eukarya

Fragmented or
HMW DNA?



Soil metagenomics methodology.

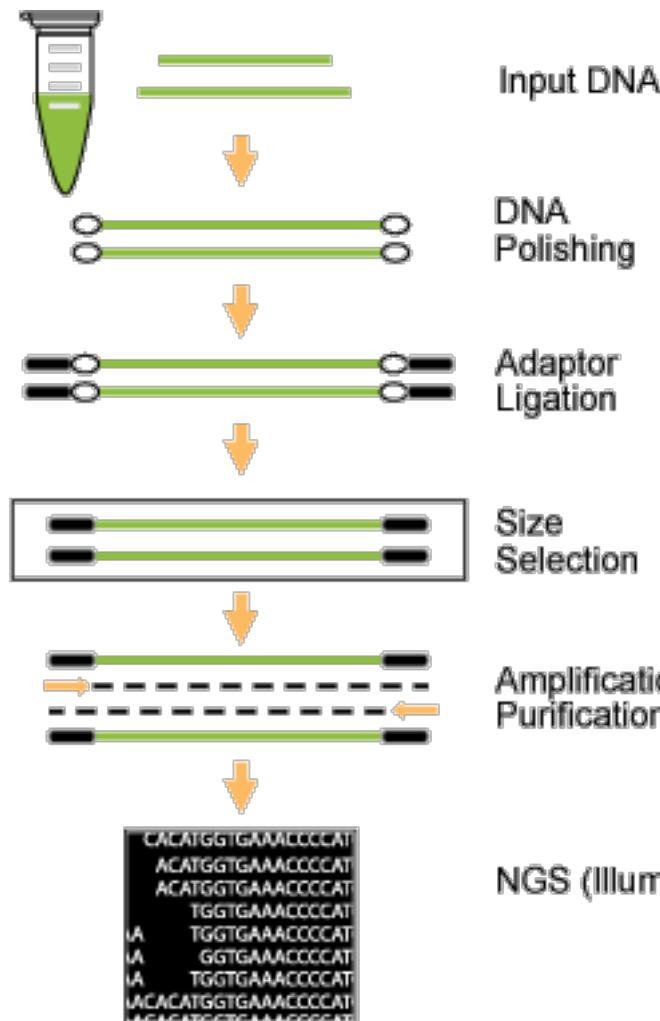
Jansson 2011: Towards tera-terra. Microbe June 2011

Illumina, PacBio,
Oxford Nanopore,
Ion Torrent

Paired end, single
read, size of library

Sequencing depth,
replicates

Library preparation

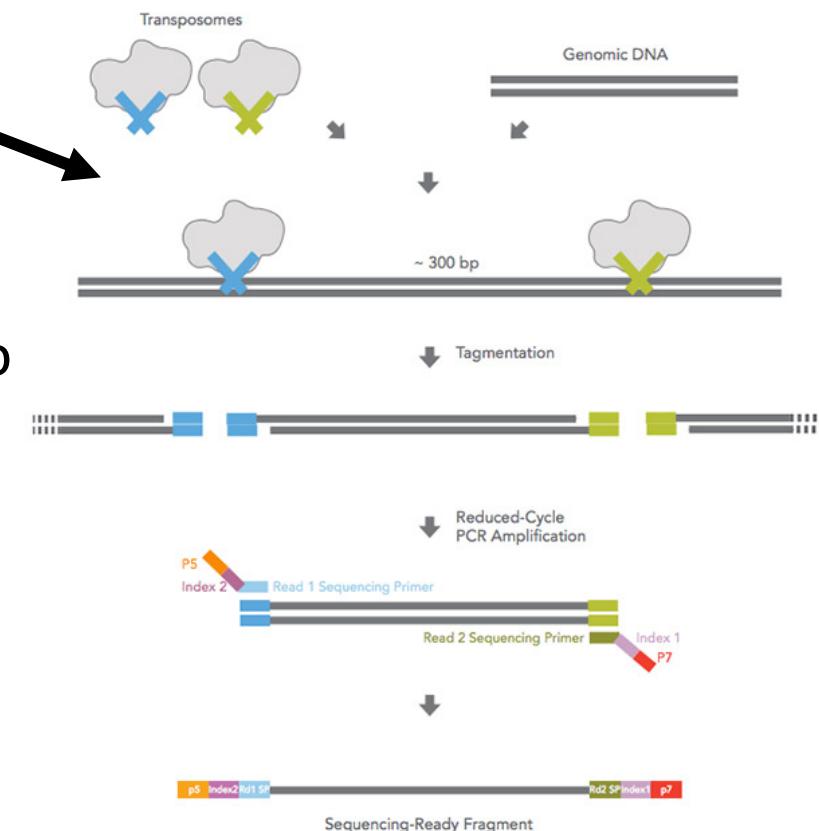


DNA shearing and adapter ligation

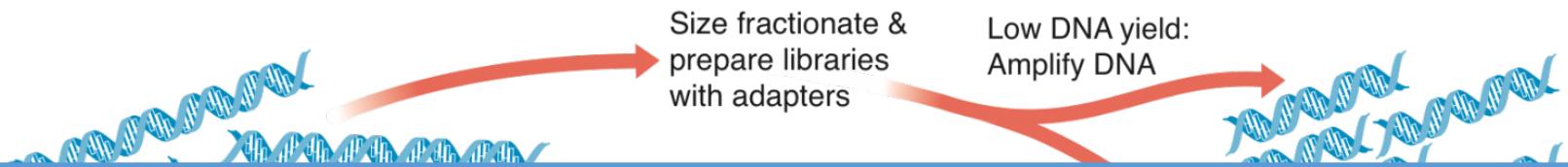
- DNA quantity

NexTera transposases

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



DNA from all
bacteria, archaea,
virus



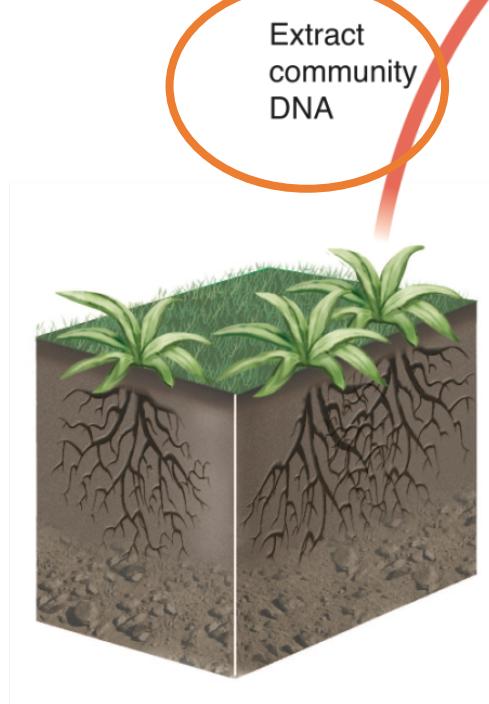
| Technology | Throughput (reads/run) | Read length | | Paired reads | Errors? |
|---------------------|------------------------|---------------|------------|--------------|-----------------------------|
| MiSeq | 25M | 250-300 bp | 15 Gb | Y | <1% |
| HiSeq | 5G | 100-150 bp | 800 Gb | Y | <1% |
| NextSeq | 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% |

predictions

Comparative
metagenomics

Soil metagenomics methodology.

DNA from all
bacteria, archaea,
viruses, eukarya



Soil metagenomics methodology.

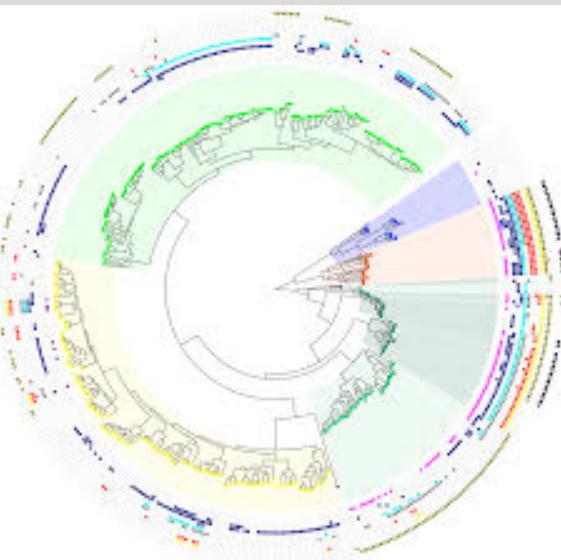
Illumina, PacBio,
Oxford Nanopore,
Ion Torrent

Paired end, single
read, size of library

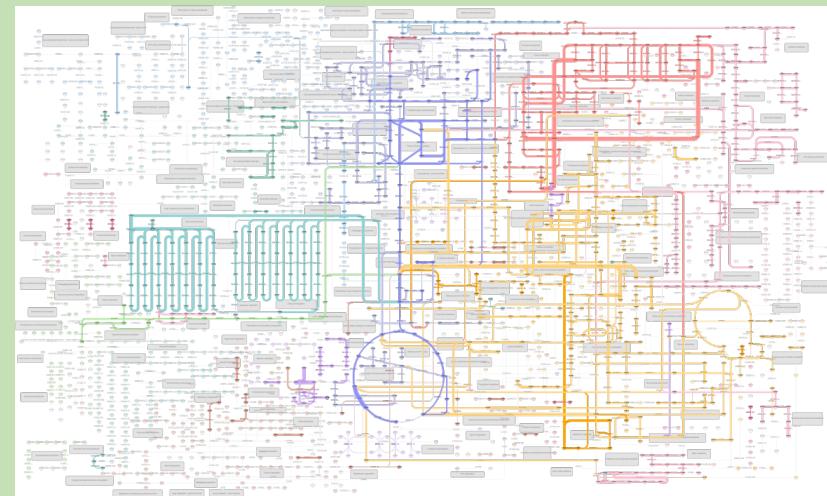
Sequencing depth,
replicates

With metagenomics

- Microbial community composition – who is there
16S/18S, ITS



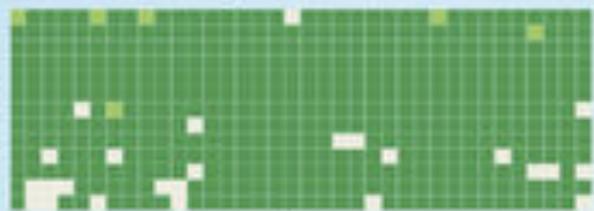
- Microbial community function – what can they do
Protein coding sequences



Assembly-based profiling

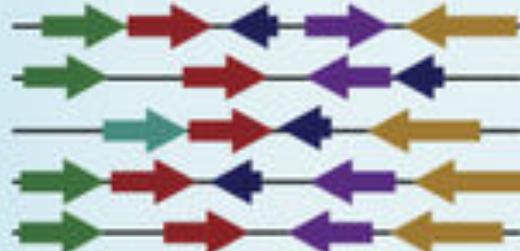
Quality control of MAGs

MAGs



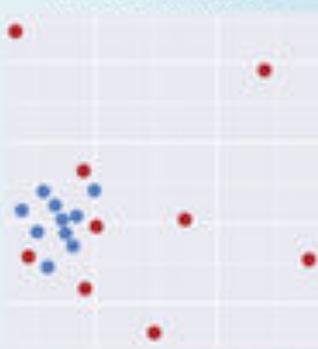
Genome-level characterization for abundant MAGs

MAGs



Genome annotations

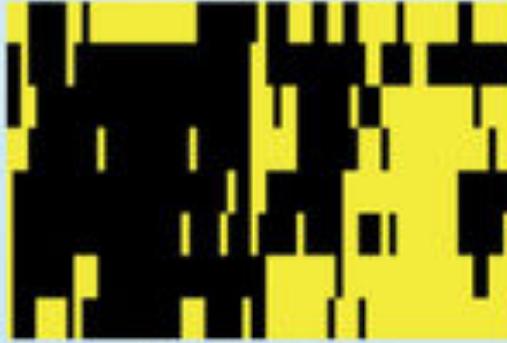
Ordination on MAG abundance



MAG phylogenies



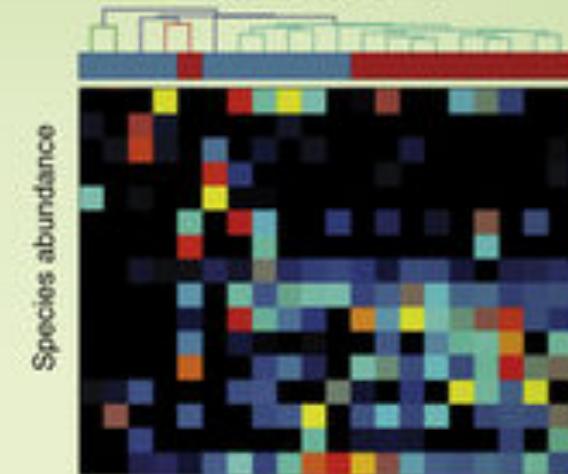
MAGs



Functional modules

Read-based profiling

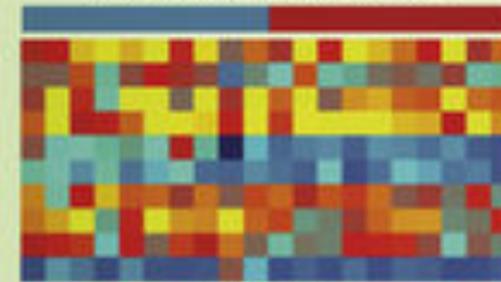
Taxonomic profiling



Species abundance

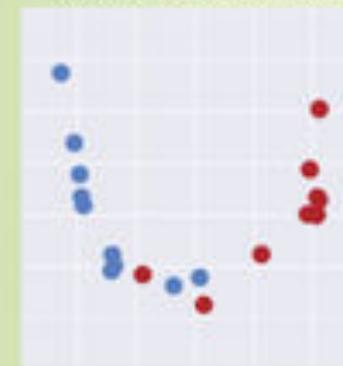
Sample

Functional potential profiling

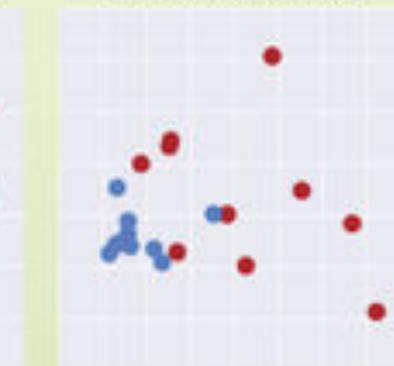


Samples

Ordination on functions

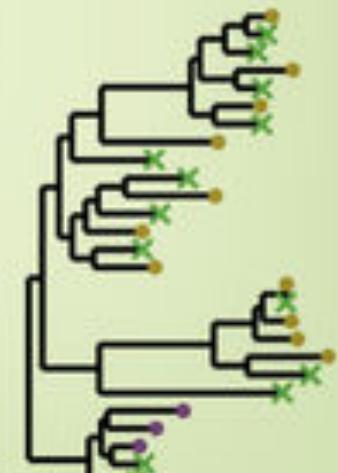


Ordination on taxa



■ Case samples

■ Control samples

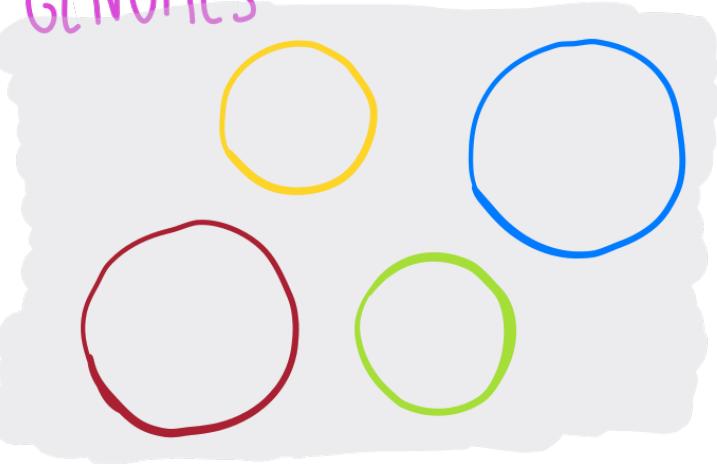


Sample
Strain-level phylogenies



Reconstructing genomes from metagenomes

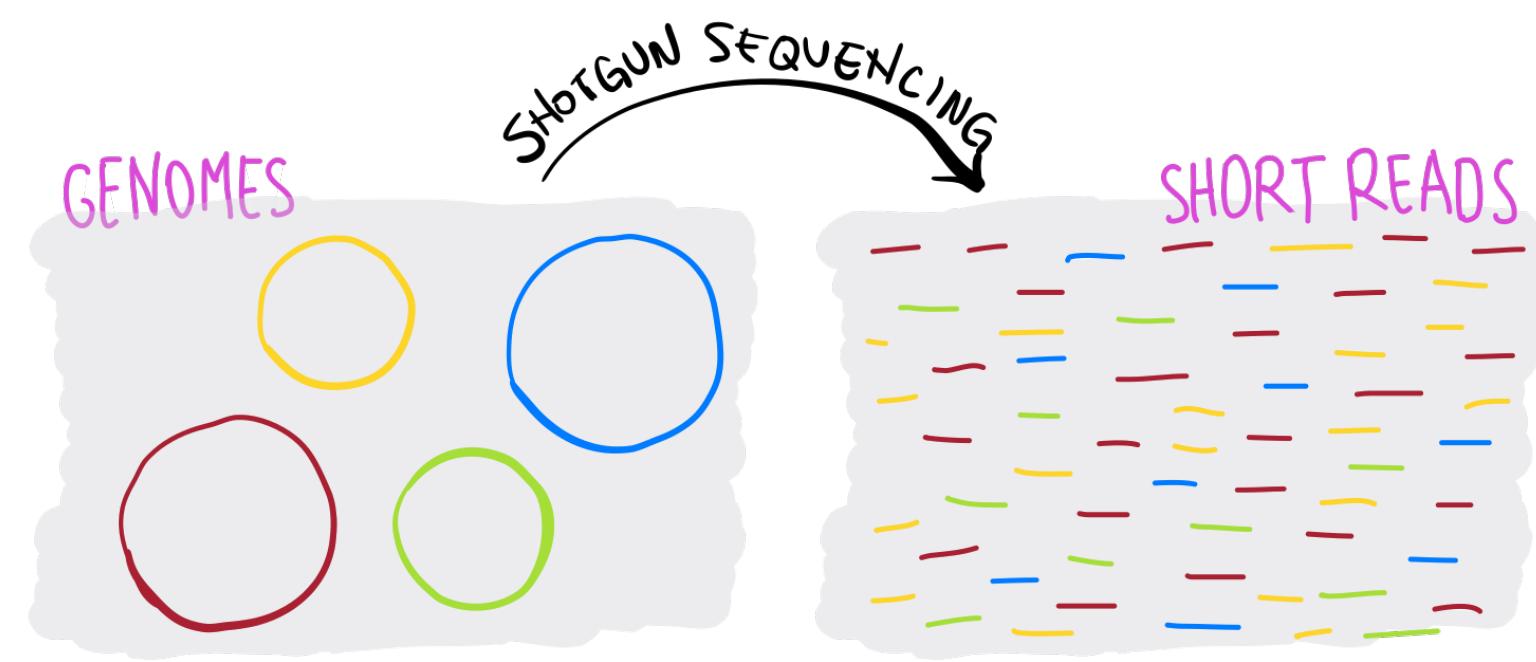
GENOMES



GENOMES

SHOTGUN SEQUENCING

SHORT READS



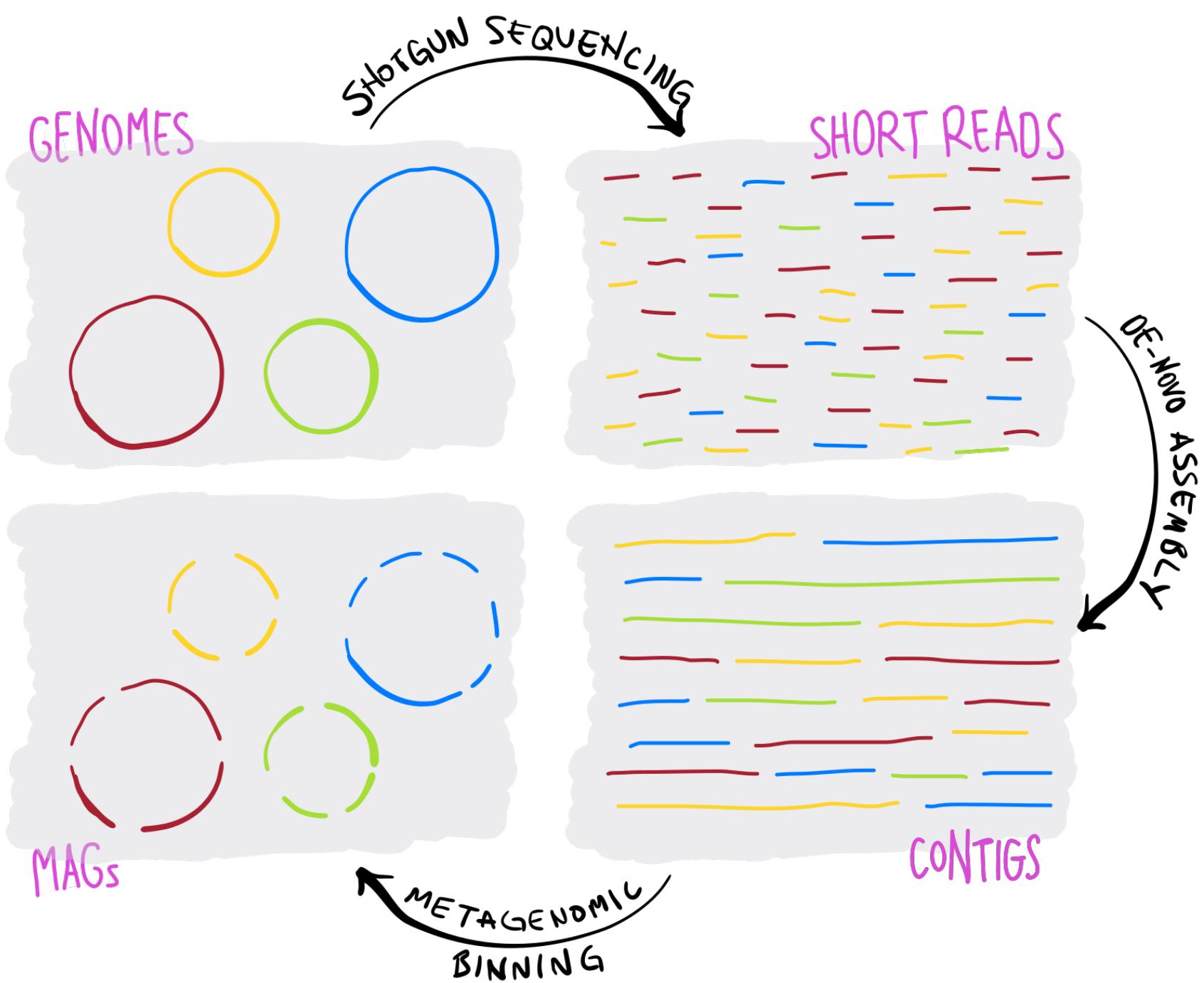
GENOMES

SHOTGUN SEQUENCING

SHORT READS

DE-NOVO ASSEMBLY

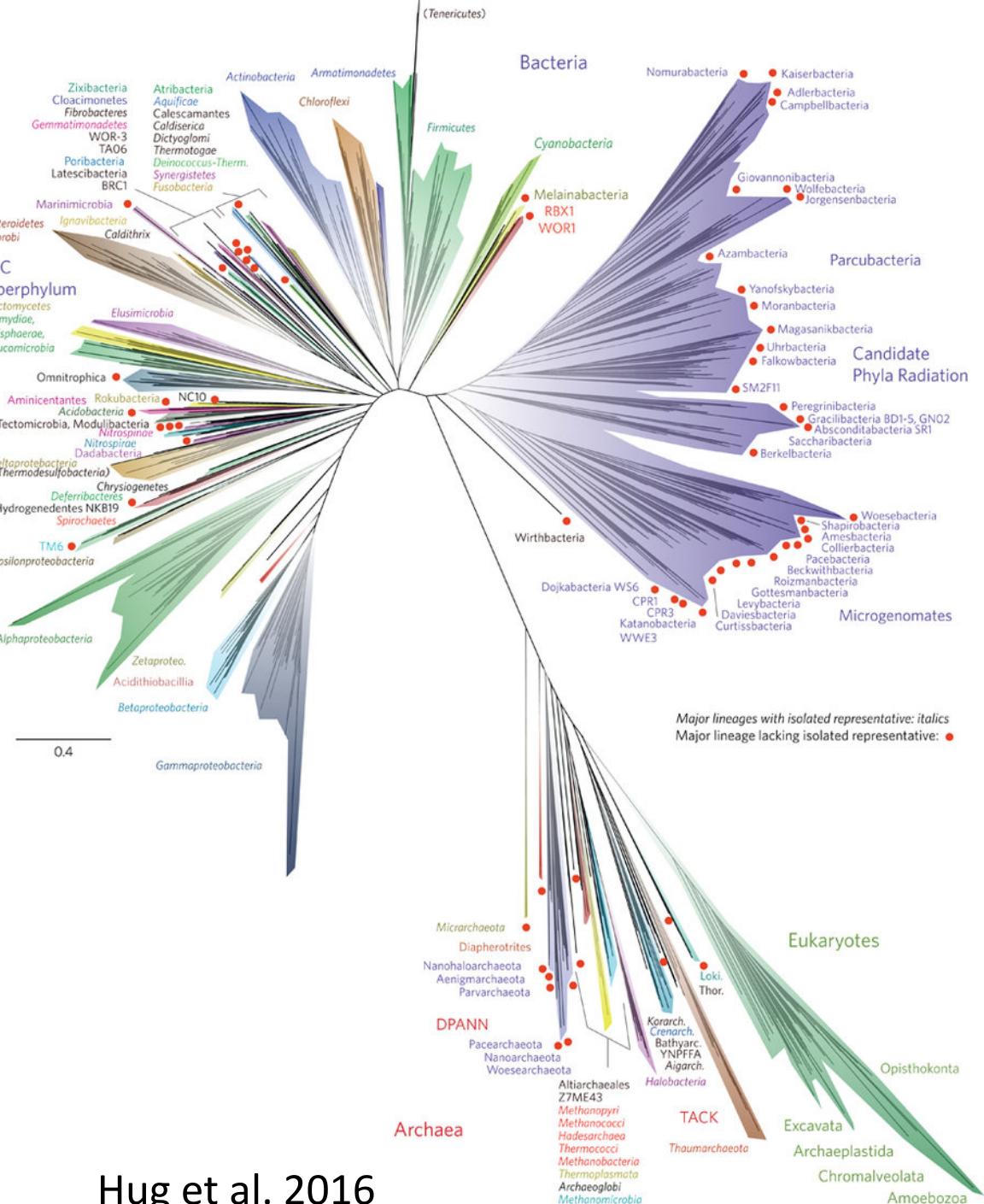
CONTIGS



Same
community
without
cultivation of
the members

Oomics

- Metagenomics
 - Metatranscriptomics
 - Metaproteomics
 - Meta-metabolomics

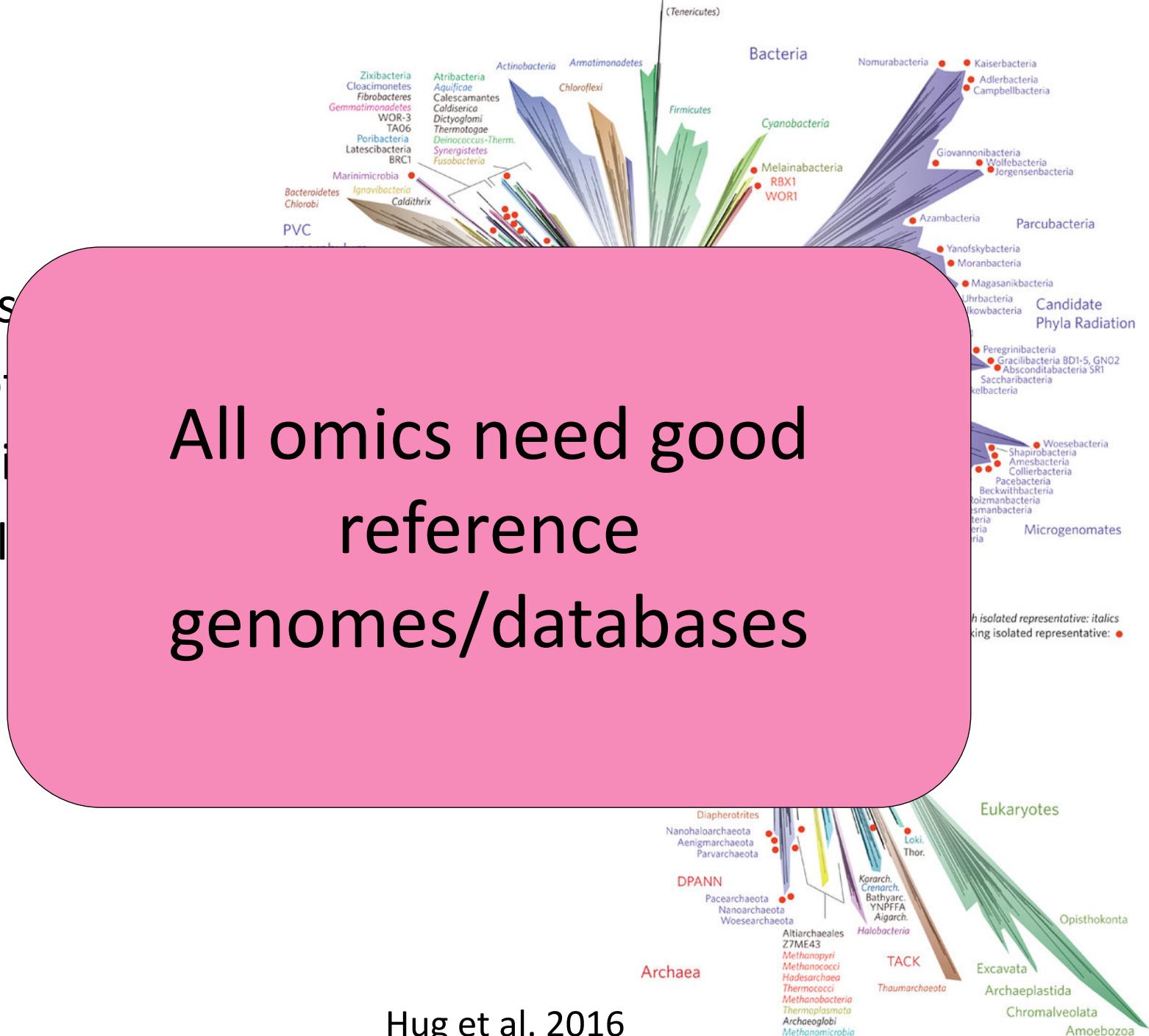


Omics

- Metagenomics
- Metatranscriptomics
- Metaproteomics
- Meta-metabolomics

All omics need good reference genomes/databases

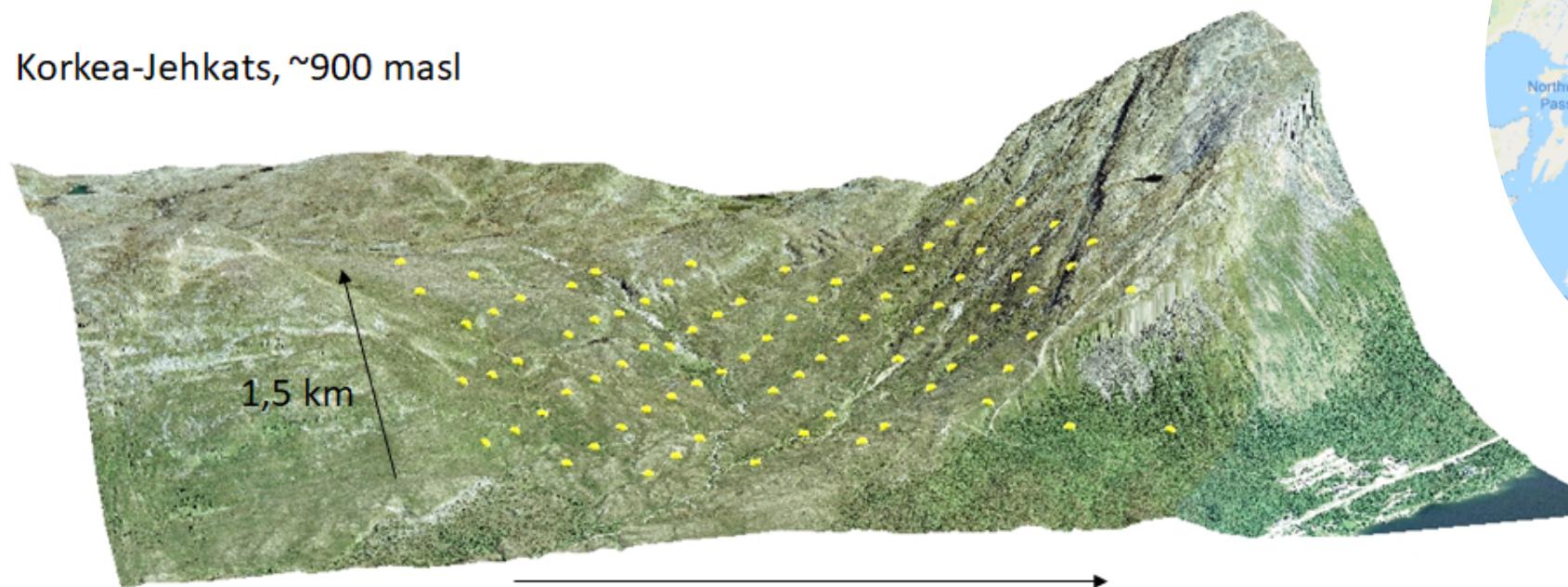
Hug et al. 2016



Dataset for this course

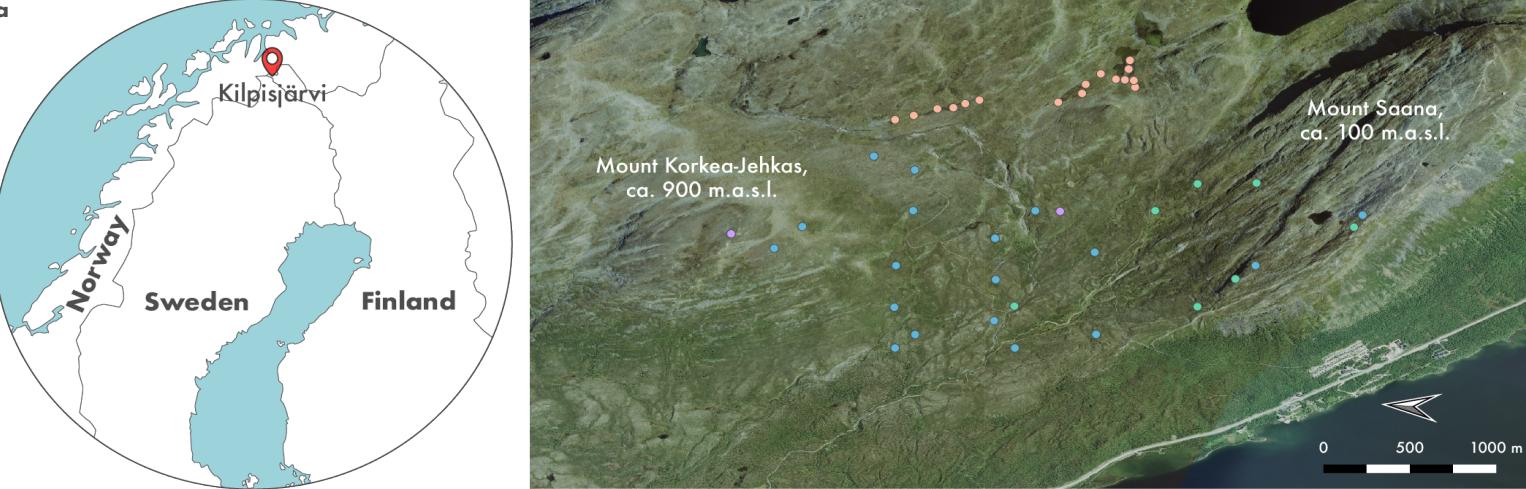
Kilpisjärvi
($69^{\circ}03'N$; $20^{\circ}50'E$)

Korkea-Jehkats, ~900 masl

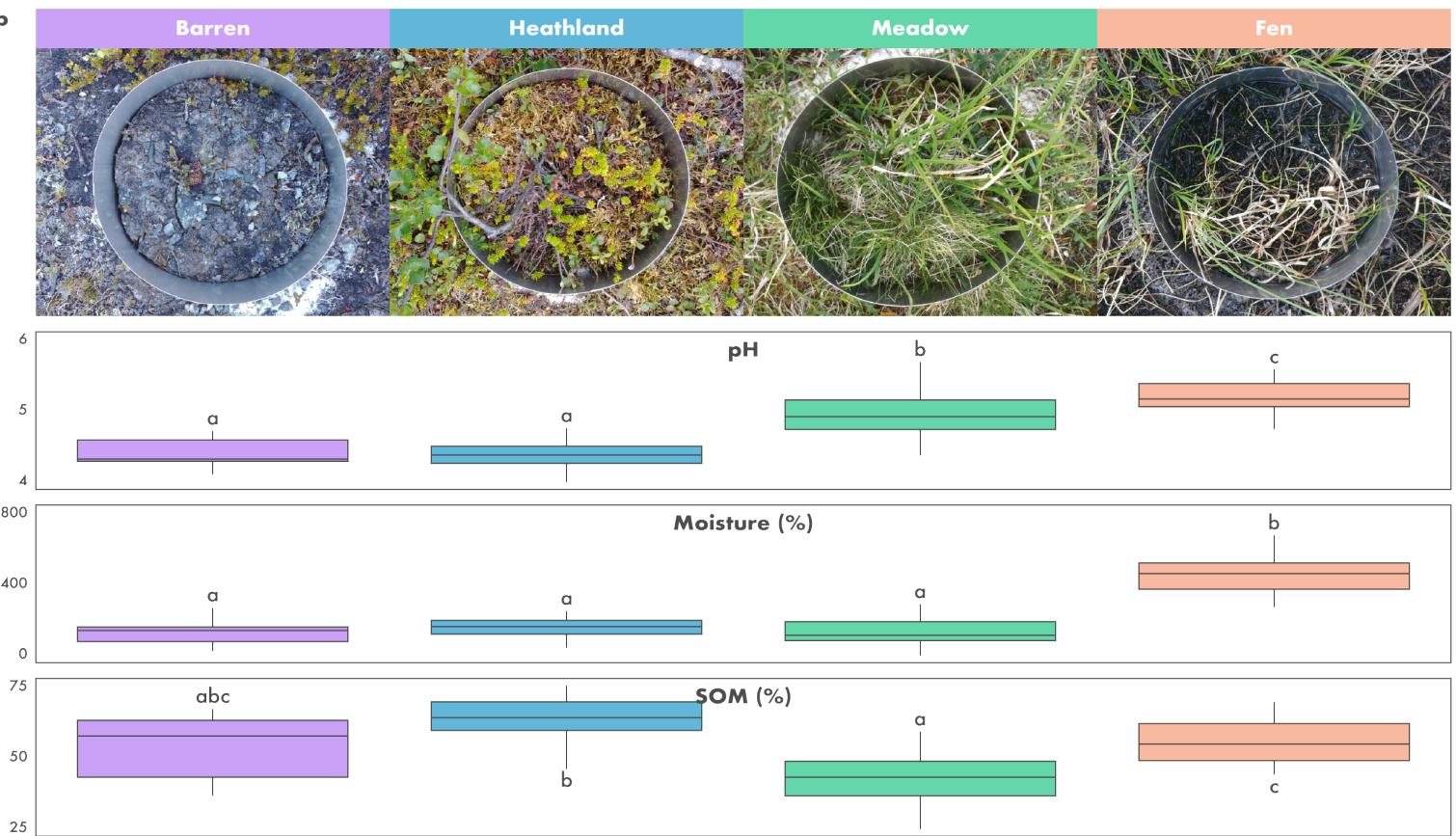


Mount Saana, ~1000 masl





- 4 samples
 - Sample01/Sample04: heathland soils
 - Sample02/Sample03: fen soils
- Sequencing:
 - Illumina NovaSeq: all samples
 - Nanopore MinION: Sample03 and Sample04



Schedule

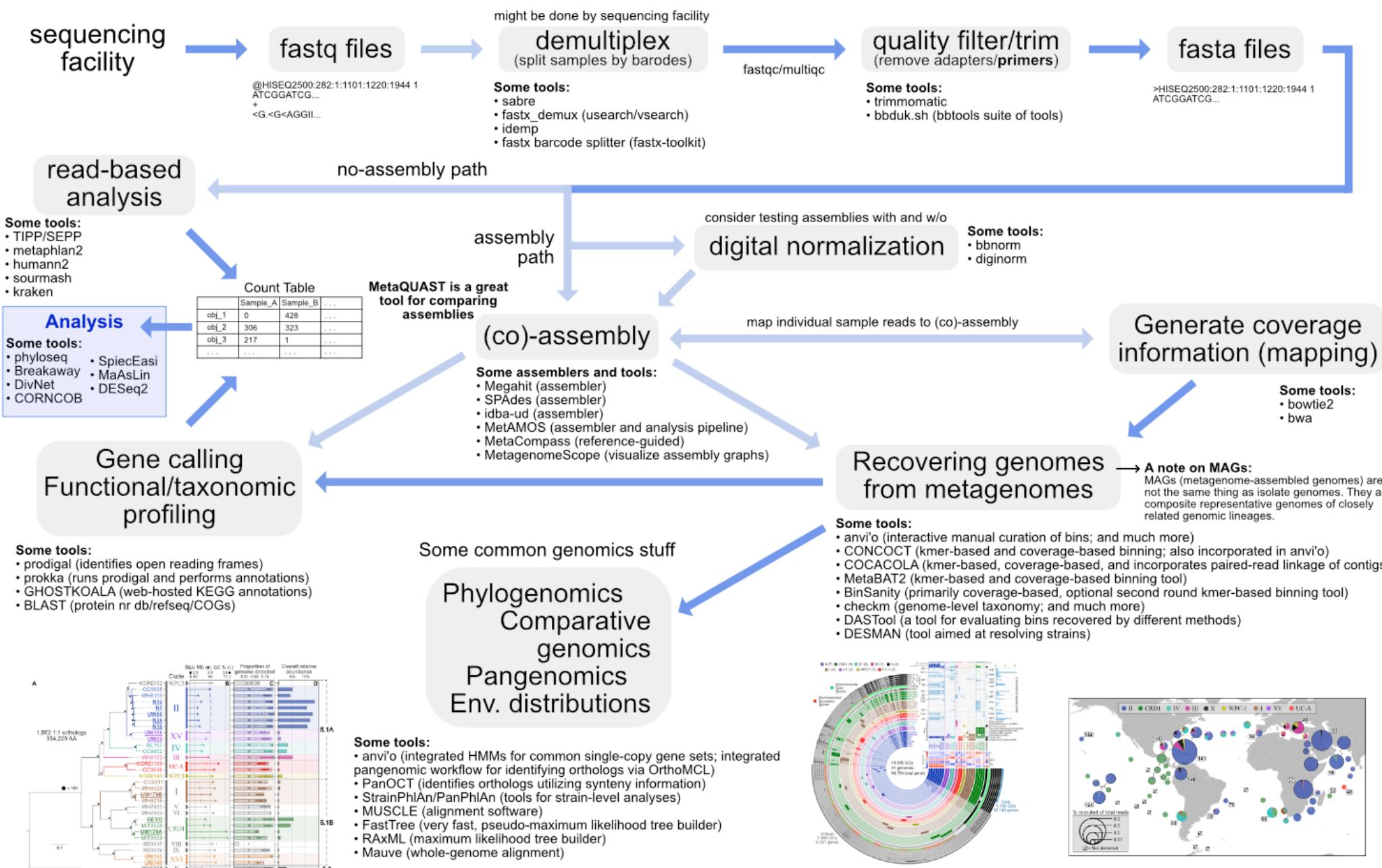
| Time | Monday | Tuesday | Wednesday | Thursday | Friday |
|--------------------|-------------------------------|---------|--|------------------|----------------|
| Morning | Intro CSC QC + trimming | MEGAN | Assembly Genome resolved metagenomics | Anvi'o interface | MAG annotation |
| Lunch break | | | | | |
| Afternoon | Read-based analysis | METAXA | Files for Anvi'o | Anvi'o | MAG analysis |
| | | | | | |

Biocenter 1 room 1014

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.



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