Ocean Omics: Metagenome Data Processing

SIOB 278
Week 1: Introduction
April 04, 2017

DATE	TOPICS / ACTIVITIES
Apr 05	**Introduction to metagenomics, research questions, and experimental design** *AWS access; Command line operations; Local software needs; Selection of independent data sets*
Apr 12	**Data types, expectations, and quality control** *Sequencing technologies; File formats; Data quality control processing (FastQC, Trimmomatic)*
Apr 19	**Assembly of metagenome sequence data** *Assembly processing (IDBA-UD, MetaSPAdes); Assembly statistics & validation (Quast)*
Apr 26	**Check point 1: Research questions, discussion, and papers**
May 03	**Gene prediction and functional annotation of assembled contigs/scaffolds** *Annotation pipeline (Prokka); Web-based analysis portal (JGI-IMG); Annotation viewing (Artemis)*
May 10	**Binning of assembled sequences I: DNA compositional metrics** *Nucleotide frequency analysis (%GC); Coverage (mapping); Analysis & visualization (Anvi'o)*
May 17	**Binning of assembled sequences II: taxonomy assignments** *Homology searches (BLAST, DIAMOND, DarkHorse); Visualization of binned data (Anvi'o)*
May 24	**Check point 2: Research questions, discussion, & papers**
May 31	**Ad-hoc analyses: population genome assembly, gene & pathway discovery**
Jun 07	**Presentation of independent research projects**

Extent of Microbial Diversity

Environment	Average No. cells per volume	Total No. cells ($\times 10^{28}$)	Estimated No. taxa
Aquatic habitats			
Freshwater	700,000	2	200,000
Ocean:	**100,000**	**6**	**2,000,000 (20,000 Archaea)**
< 200 m	800,000	2	
> 200 m	100,000	4	
Marine subsurface			
0-10 cm	280,000,000	25	
>10 cm	14,000,000	330	
Terrestrial habitats			
Soil	800,000,000	26	4,000,000
Subsurface (> 10 m)	800,000	25-250	
Host-associated habitats			
Human gut	320,000,000,000	0.000039	1,000

~ 5×10^{30} microbial cells globally representing ~10^7 species

*Data from Whitman WB *et al*. *PNAS* (1998), Karner MB *et al*. *Nature* (2001), Curtis TP *et al*. *PNAS* (2002), Curtis TP *et al*. *Phil Trans B* (2006)

ENVIRONMENTAL MICROBIAL BIOLOGY

outstanding questions relating microbial processes and activities in the environment

☑Physiology & metabolism

Genetic & metabolic potential of uncultivated organisms

Genotype : phenotype linkage

Activity in the environment

Genes of unknown function

Cultivating the uncultivated

☑Autecological phenomena

Environmental adaptation

Biogeographical patterning

Diversity & spatio-temporal dynamics

Environmental control of expression

☑Evolutionary processes

Modes and tempo of species evolution / selection

Species definition

Horizontal gene transfer dynamics

Evolution of organelles...

☑ Discovery

Novel genes, metabolisms, physiologie

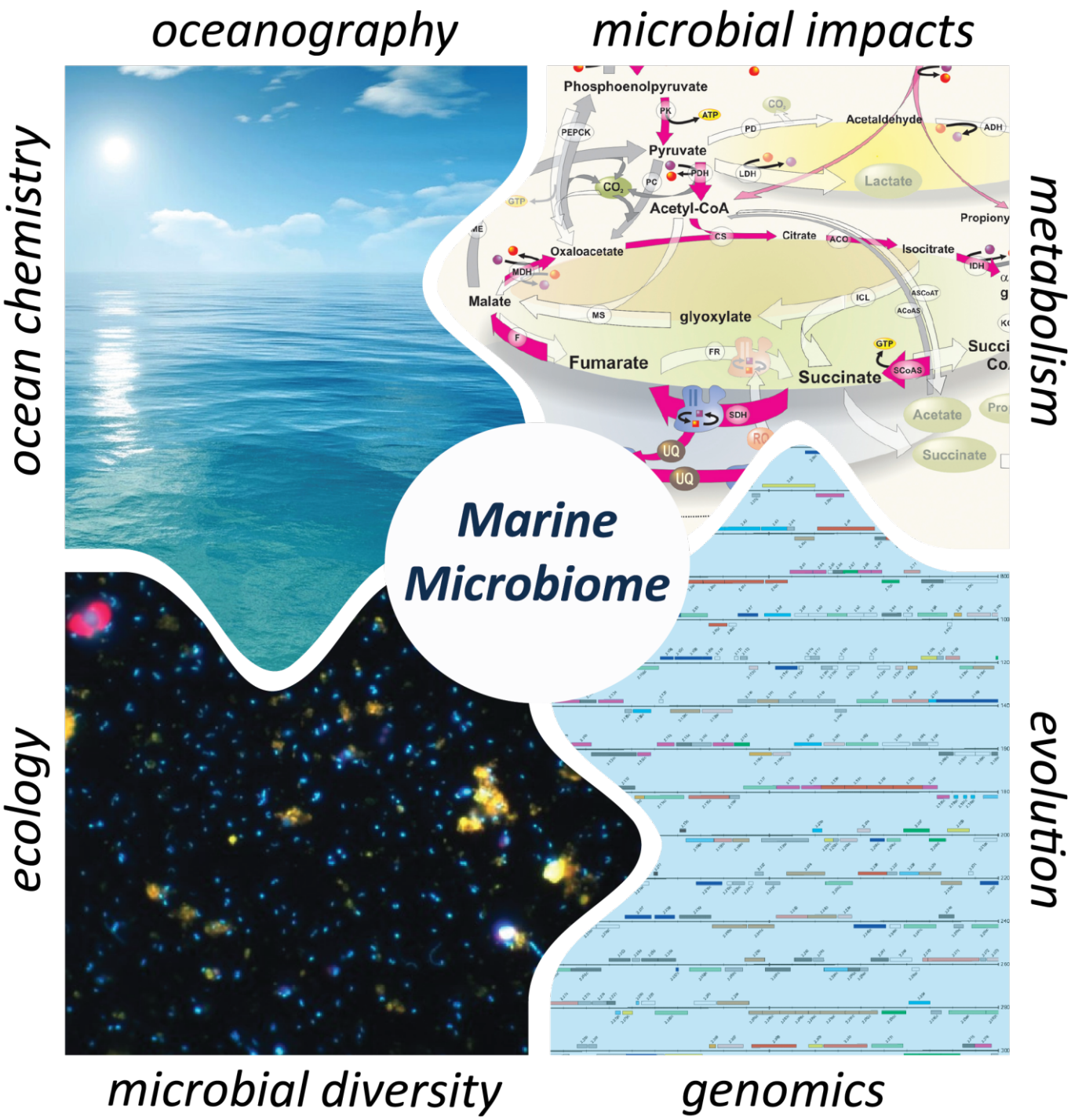
Novel diversity

Biotechnology / biomedicine products

Human health

Genomic approaches to characterize environmental microbial biology

- ❑ 16S rRNA gene tag sequencing
- ☑ Metagenomics
- ❑ Metatranscriptomics
- ❑ Metaproteomics
- ❑ Metabolomics
- ❑ Isolate sequencing


oceanography
microbial impacts
ocean chemistry
metabolism
Phosphoenolpyruvate
PK
ATP
PEPCK
PD
Acetaldehyde
ADH
Pyruvate
PDH
LDH
Lactate
PC
Acetyl-CoA
Propiony
CS
Citrate
ACO
Isocitrate
IDH
Oxaloacetate
MDH
Malate
MS
glyoxylate
ICL
ASCoAT
ACoAS
F
FR
Fumarate
Succinate
SCoAS
Succi
Co
Acetate
Succinate
SDH
UQ
Marine
Microbiome
ecology
evolution
microbial diversity
genomics

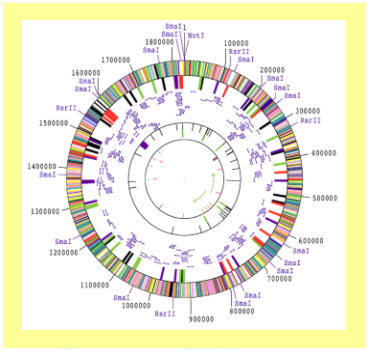
…the metagenomic approach…

in situ

- Field collection
- Geochemistry
- Sample analysis
- Sample preservation

ex vivo

- DNA extraction
- PCR amplification
- Library construction
- RNA extraction
- cDNA synthesis
- DNA sequencing
- Protein extraction
- LC/MS-MS analysis



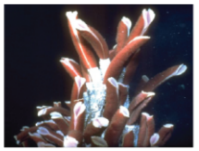
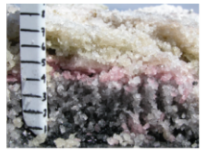
in silico

- Sequence processing
- Quality control filtering
- Sequence assembly
- Phylogenetic binning
- Functional annotation
- Comparative genomics
- Metabolism/physiology
- mRNA/protein expression

in situ



Environmental Sample

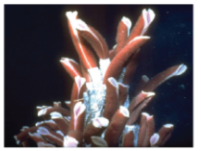
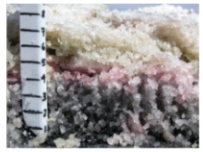




ex vivo

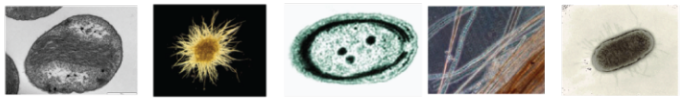
in silico

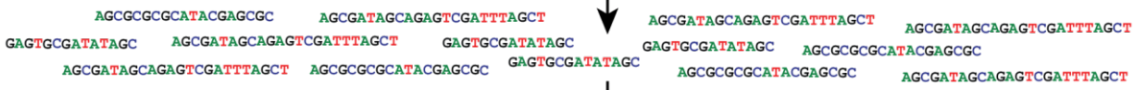


Environmental Sample

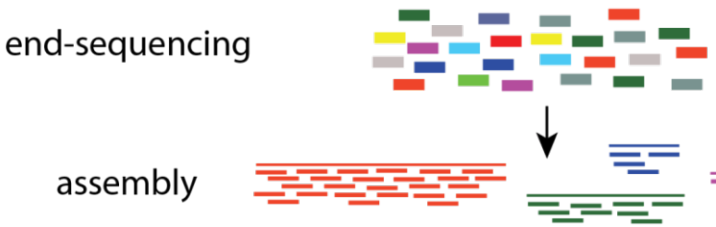




[Geochemistry / Hydrology / Minerology / Temperature / Pressure / Location / Time]





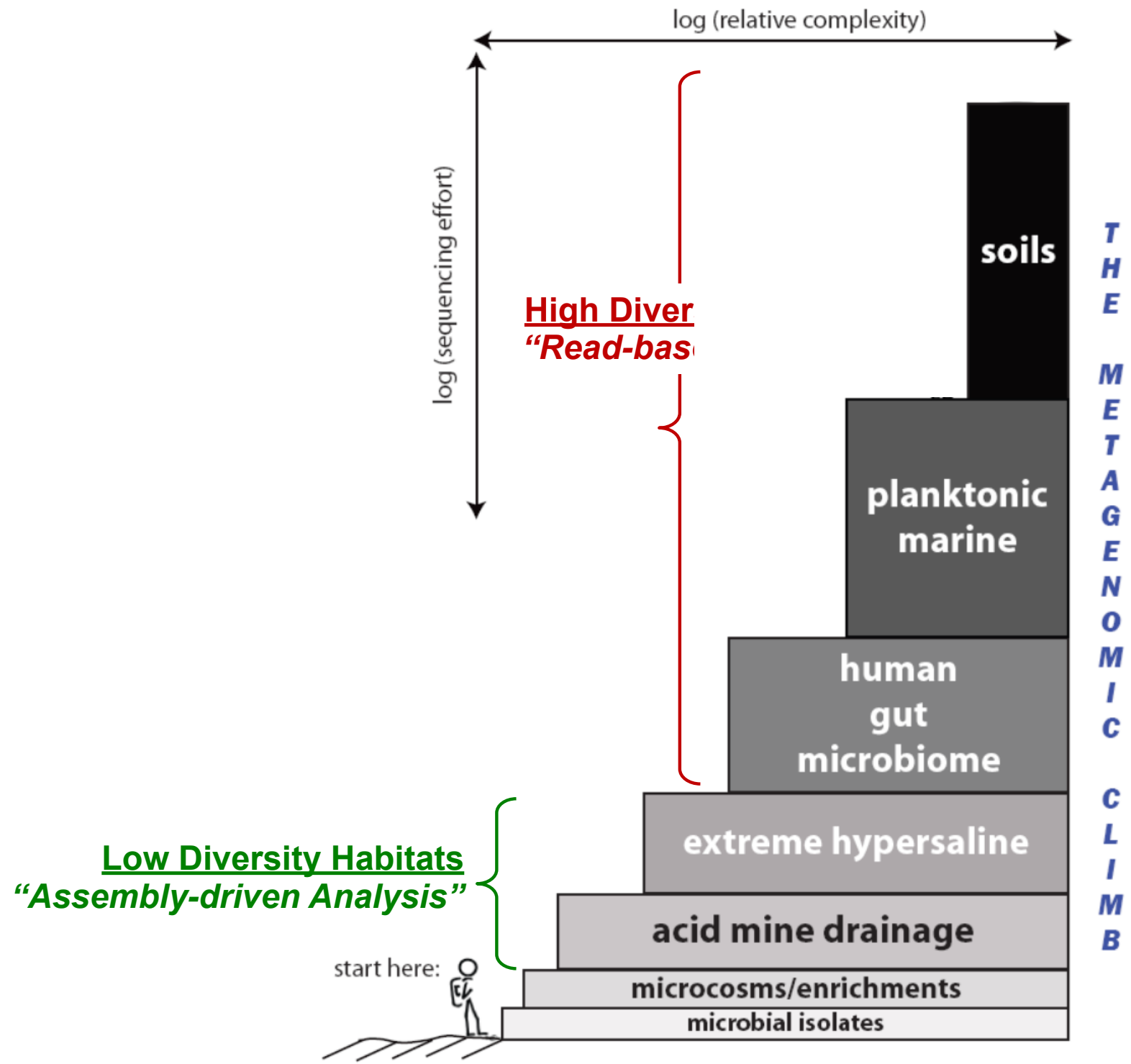
whole genome shotgun sequencing



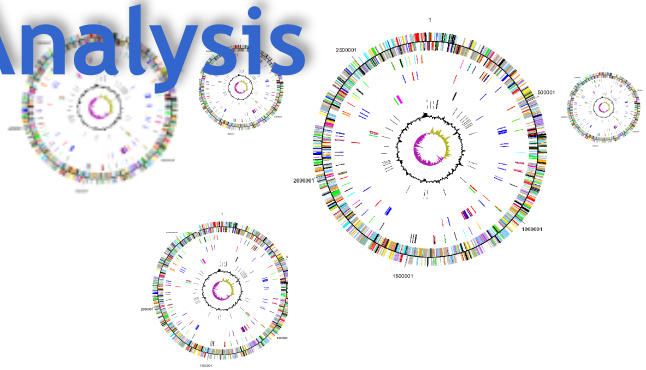
“Read-based” analysis

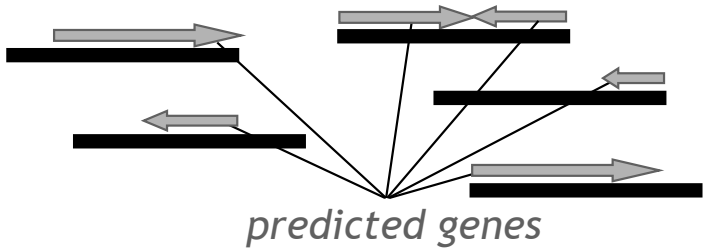
“Assembly-driven” analysis

Metagenomic Adventures


log (relative complexity)
log (sequencing effort)
High Diver
"Read-bas
soils
planktonic marine
human gut microbiome
extreme hypersaline
acid mine drainage
microcosms/enrichments
microbial isolates
start here:
THE METAGENOMIC CLIMB
Low Diversity Habitats
"Assembly-driven Analysis"

"Assembly-driven" vs "Read-based" Analysis





- ☑ Improved phylogenetic binning of DNA fragments
- ☑ Insight into operon structure & genomic islands
- ☑ Improved metabolic reconstruction for organisms
- ☑ Provides reference sequences for transcriptomic, proteomic, & read-based studies
- ☑ Rapid prediction of genes & functional annotation
- ☑ Provides a census of taxa & metabolic processes
- ☑ Allows comparative analysis within & between samp
- ☑ Broad view of community structure & function

The level of environmental sequencing must be commensurate with the expected diversity of the sample

undersampling complicates interpretation due to incomplete data sets

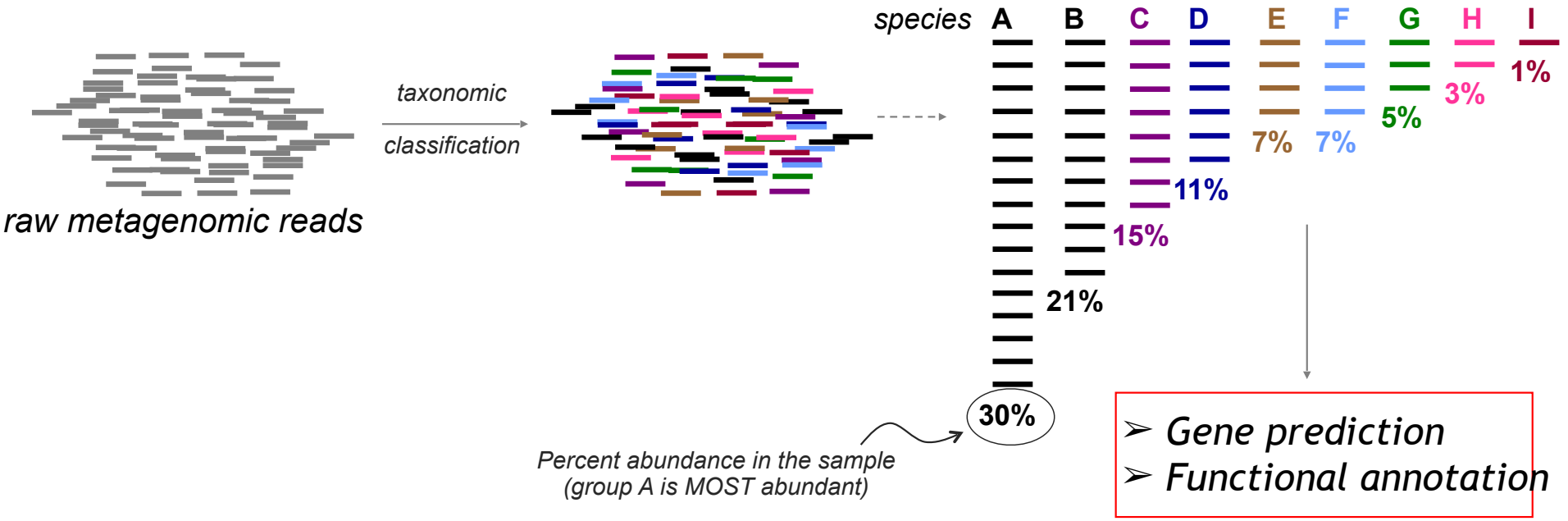
"Assembly-driven" vs "Read-based" Analysis

- **Read-based analyses** look at the genes encoded on individual metagenomic sequencing rea
- **Assembly-driven analyses** seek to assemble complete genomes from metagenomic reads

---

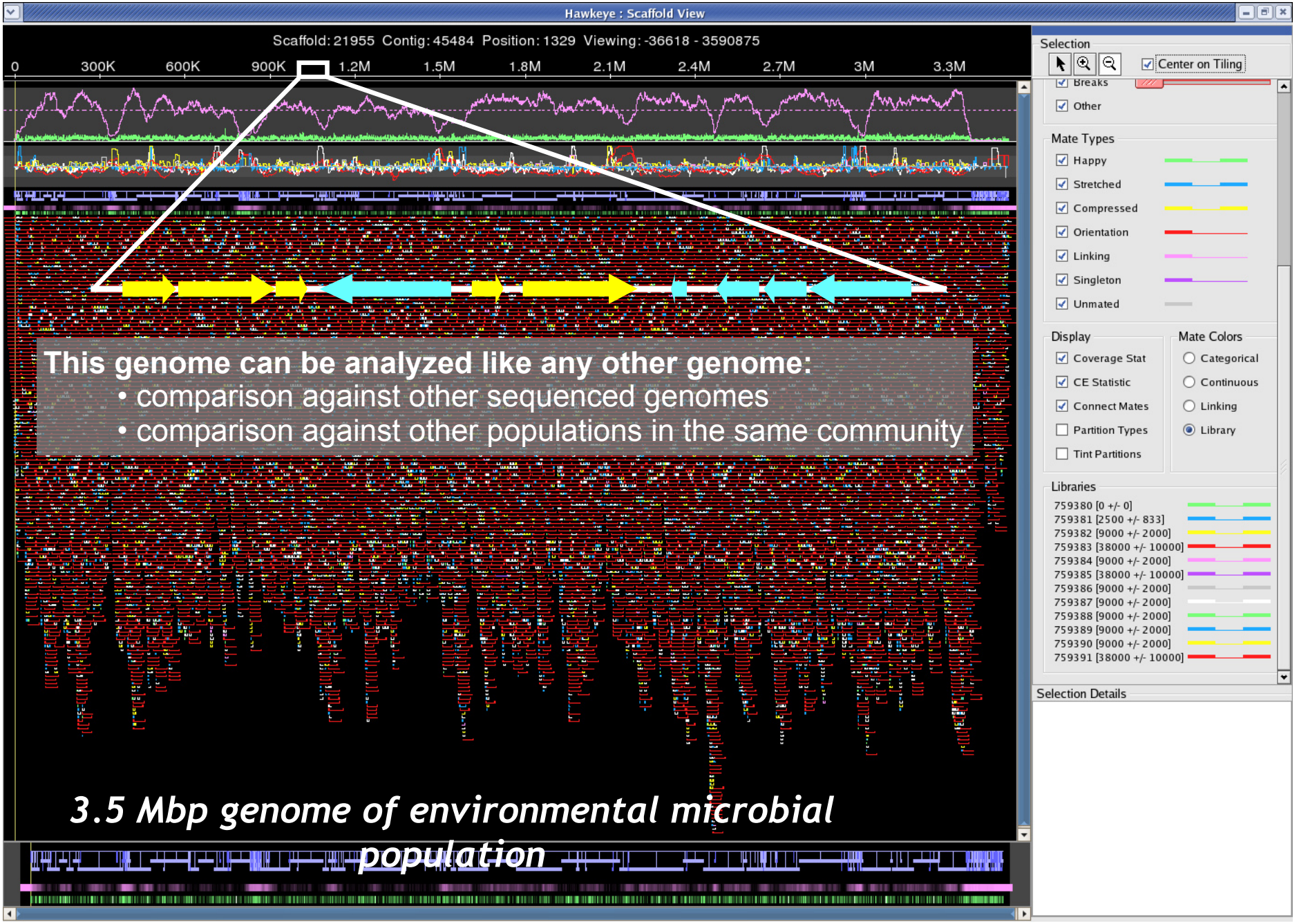
Having only a few pieces of the puzzle complicates interpretation but there are still bits of information in these small pieces (metagenomic reads)

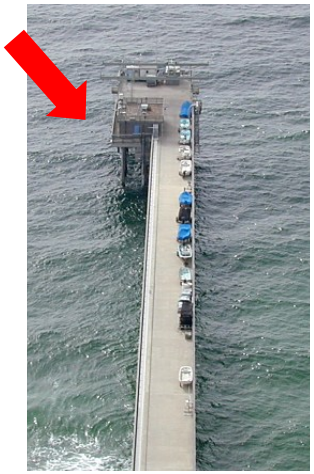
“Read-based” Sequence Analysis



No assembly required to provide an inventory of the taxonomic diversity and functional capabilities within the community sampled

“Assembly-driven” Analysis





Environment

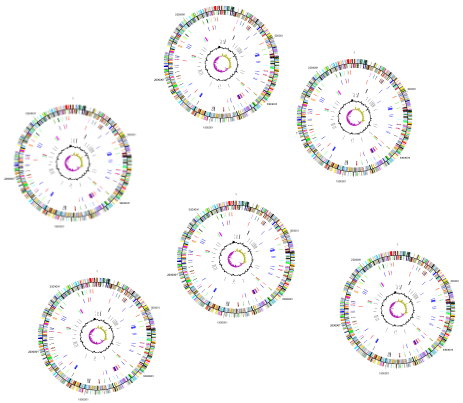
Seawater

Harvest cells
&
Extract DNA

Make genomic
sequencing library

Sequence
Sequence
Sequence

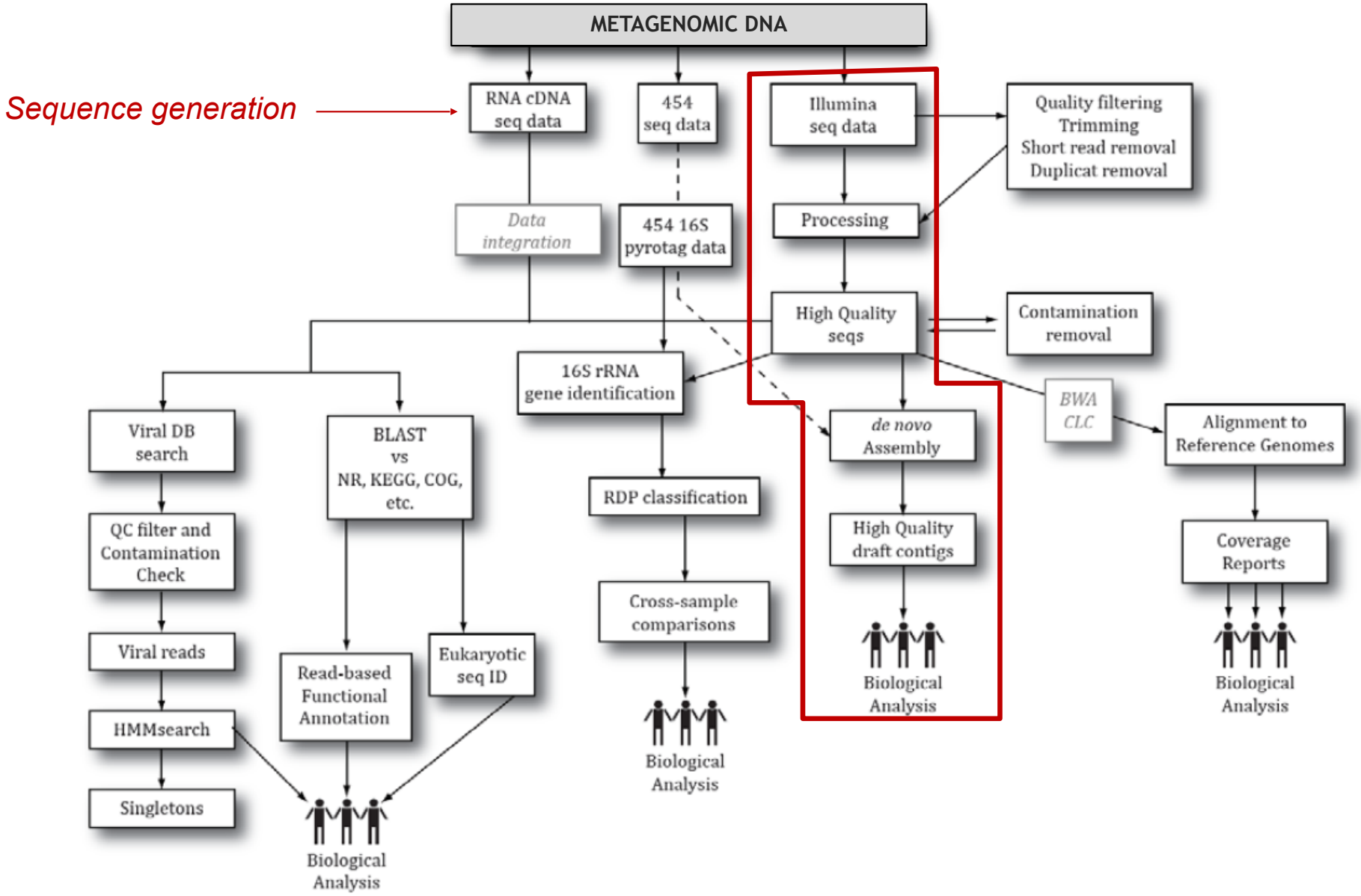
Assemble
&
Analyze



Genomes from my environment

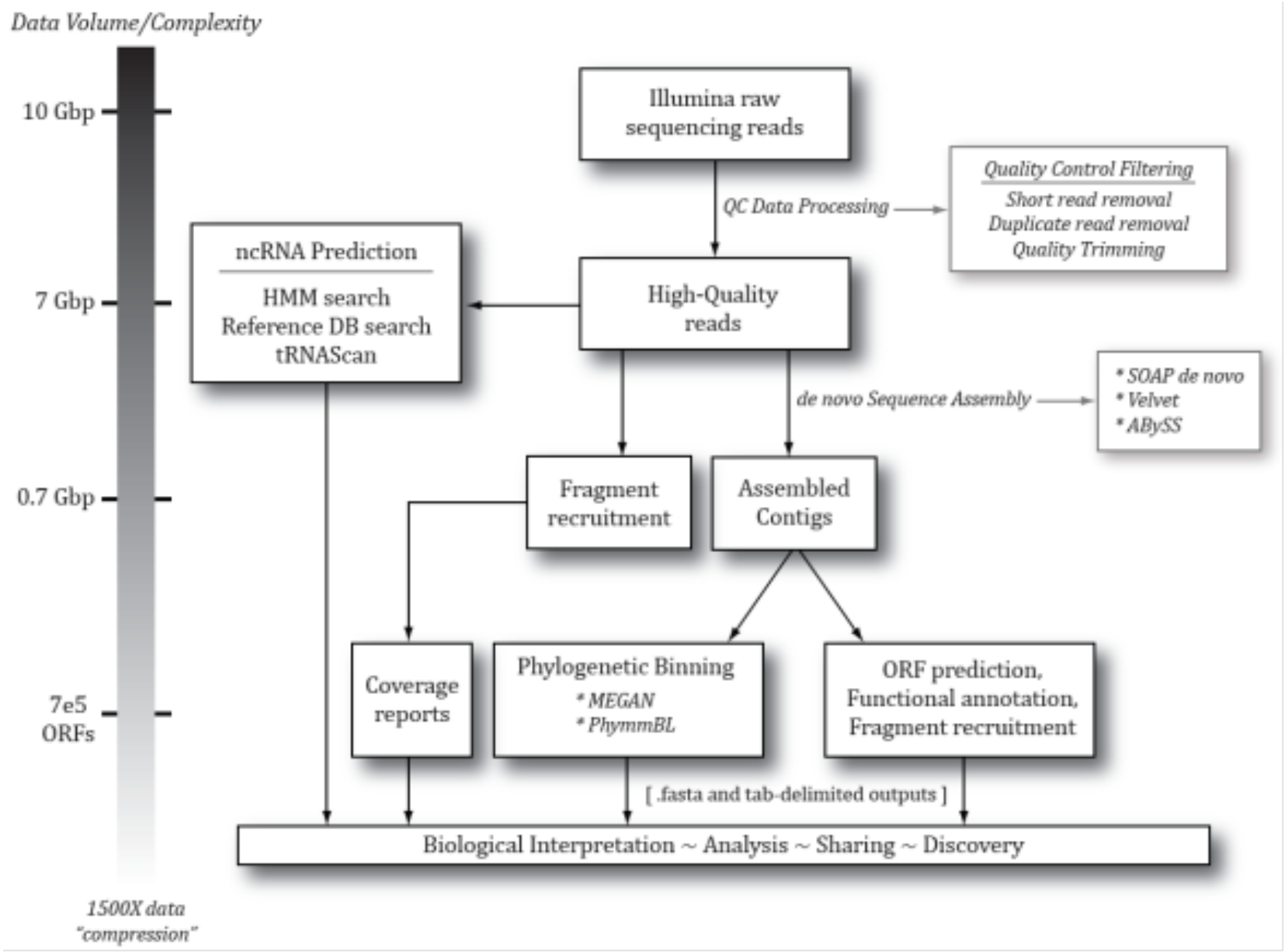
Schematic of simple metagenomic workflow

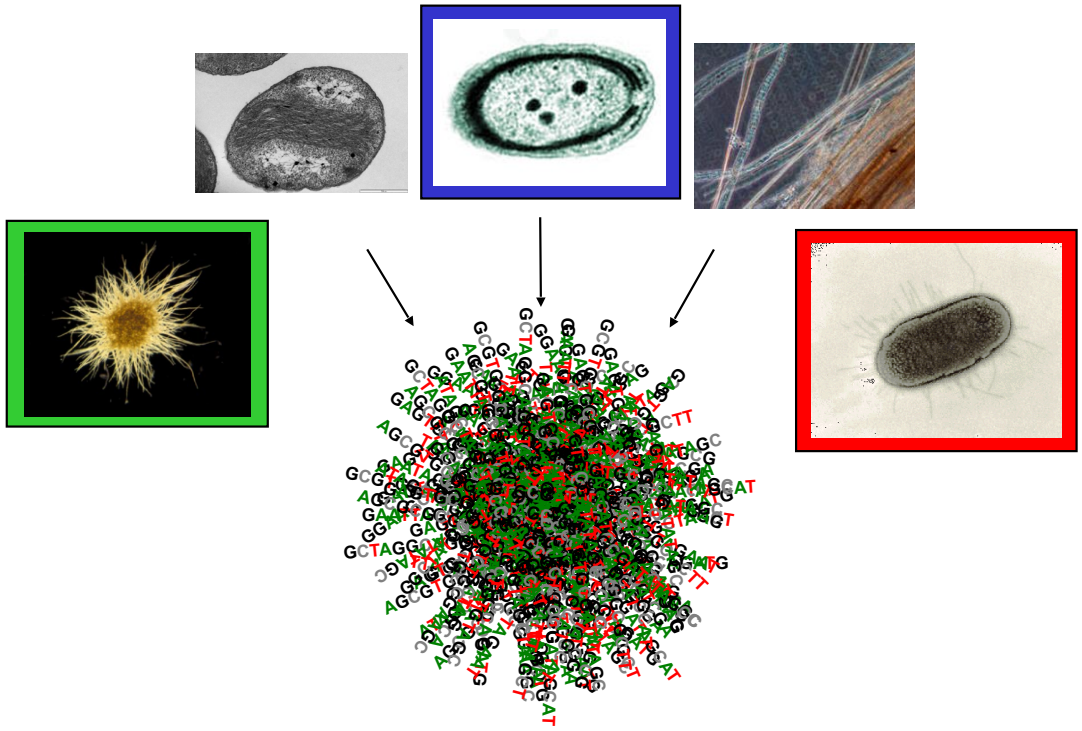
Analysis scheme for processing metagenomic data sets



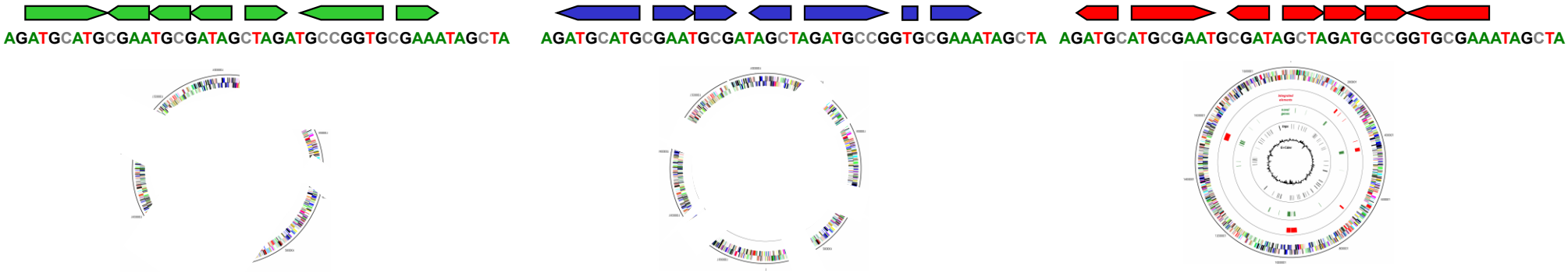
Flowchart from Weinstock G & Allen EE, NGS workshop report (Mar 2011)

Metagenome *de novo* assembly workflow

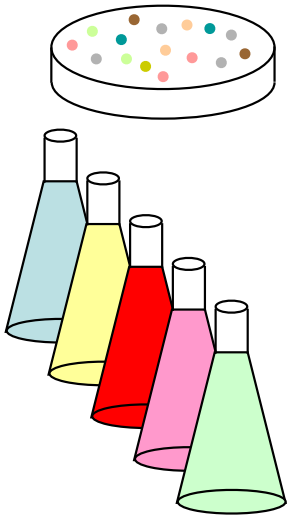


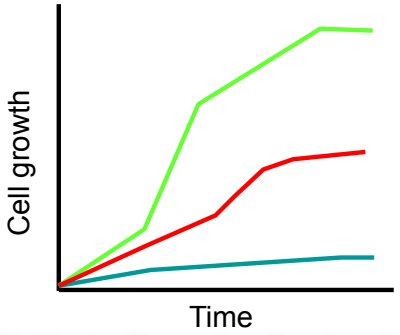


Metagenomic Assembly, Phylogenetic Binning, Functional Annotation



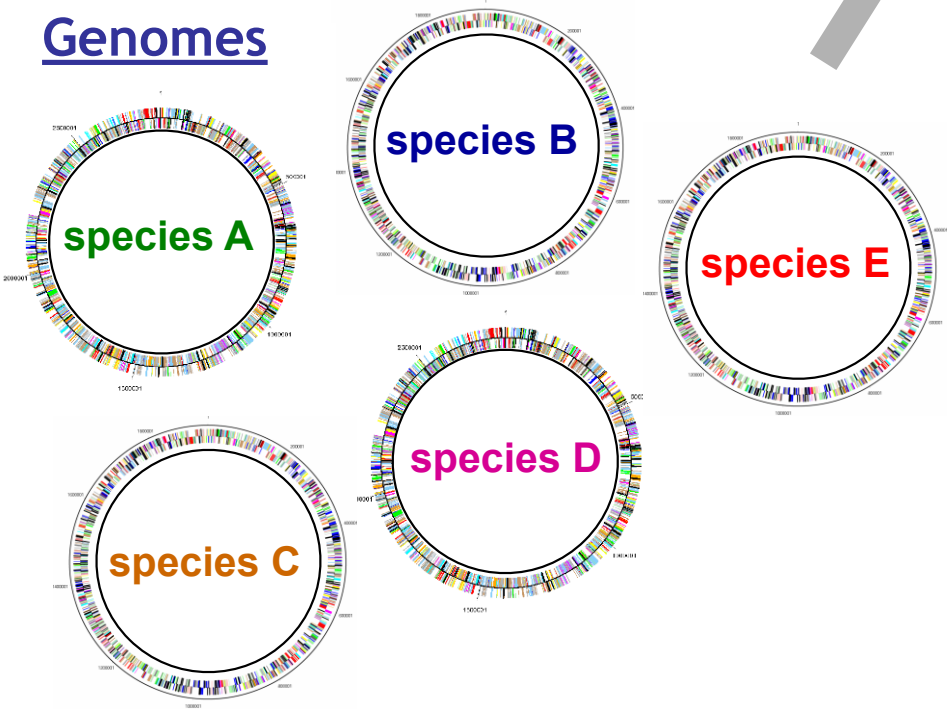
Experimentation



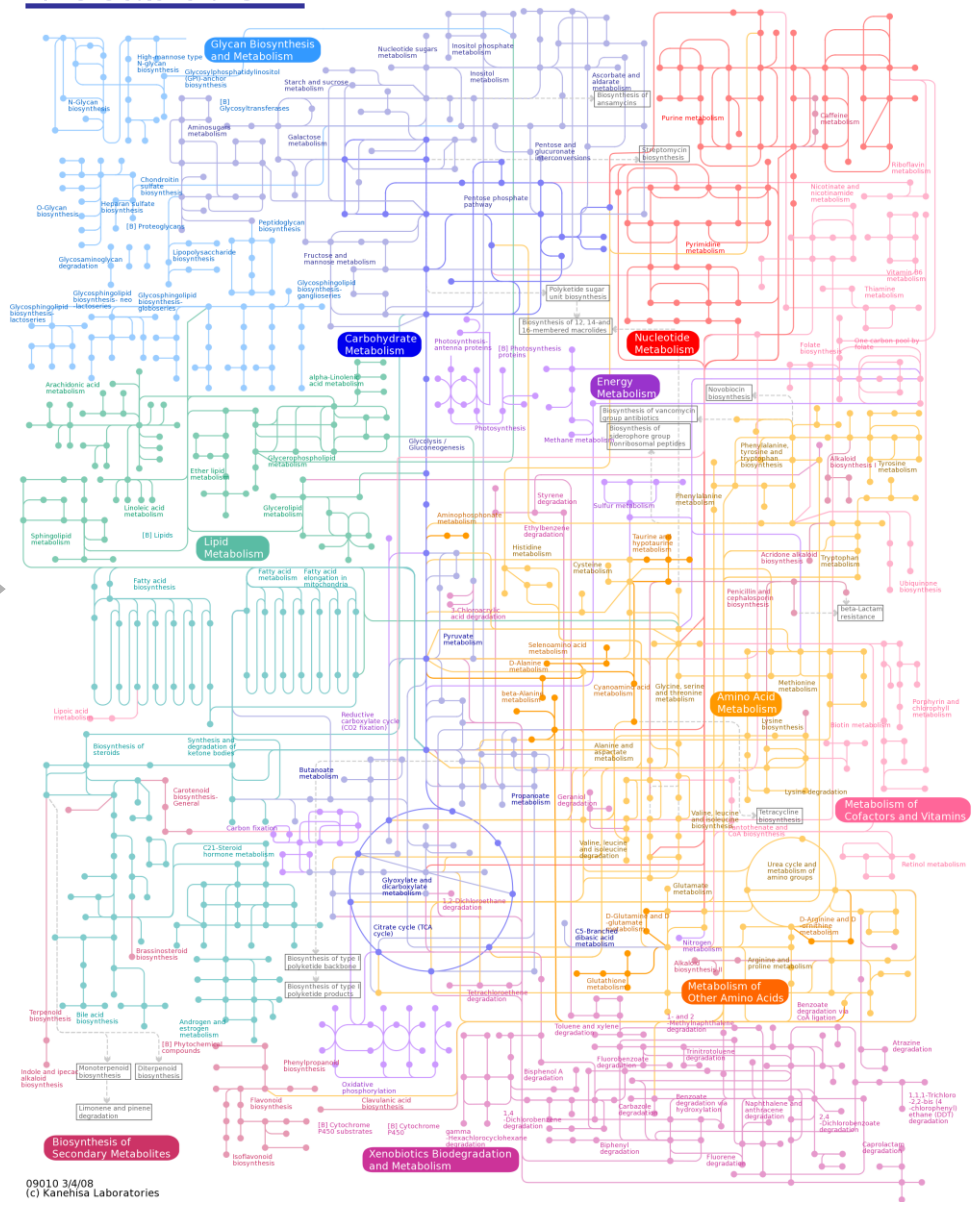


Nutrients, Temp, *hv*, O_2, chemicals, etc.

Genomes



Metabolism



…how an organism makes a living…