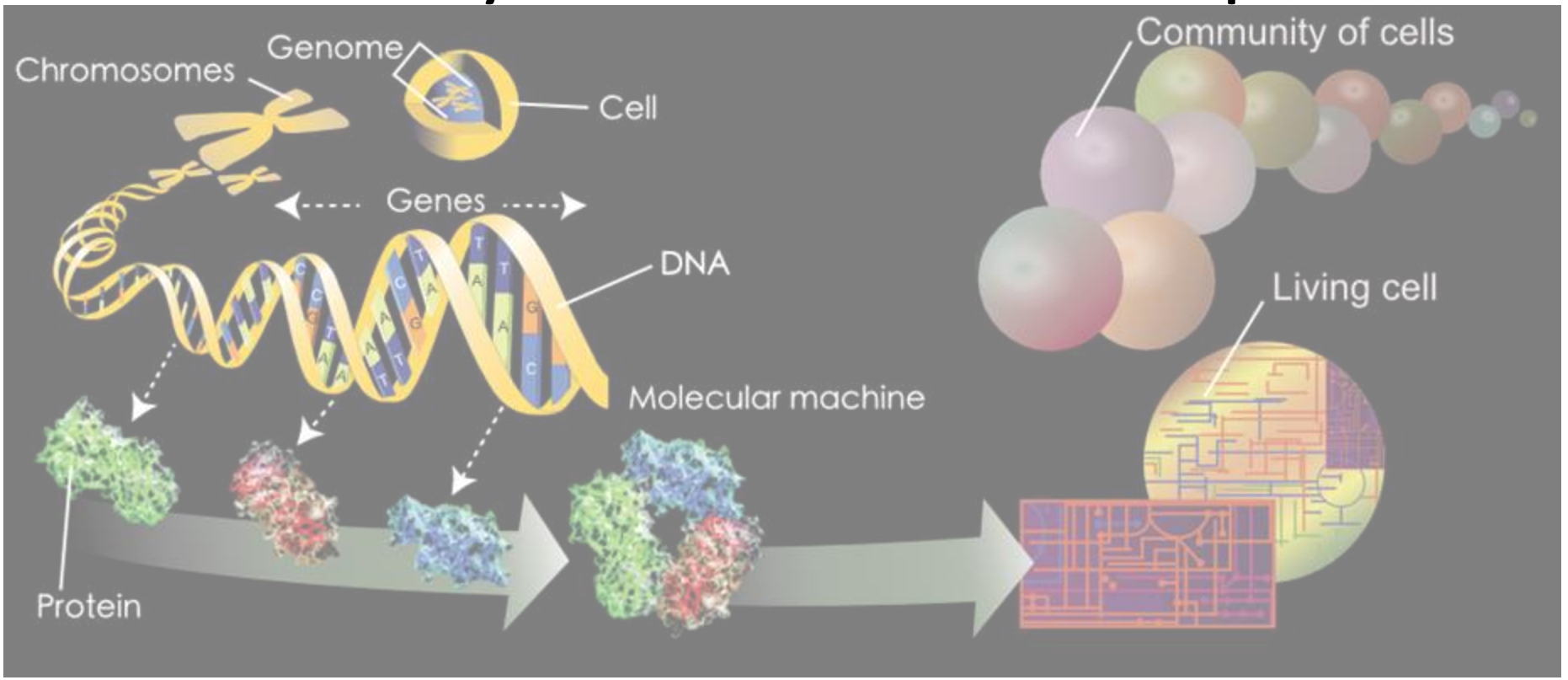
From Molecules to Ecosystems: An Introductory Tour of "OMICS" Pipelines



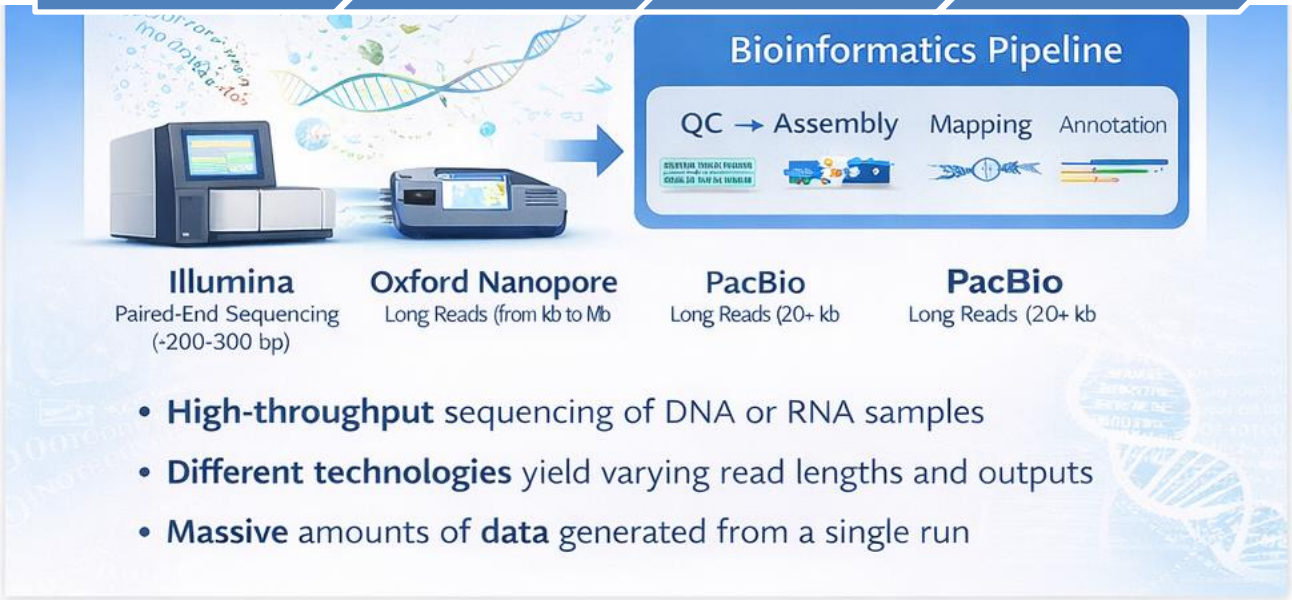
Making sense of 'big' biological data

Emma Rocke | SEEC Seminar | University of Cape Town

What do we mean by “Big Data” in Omics?

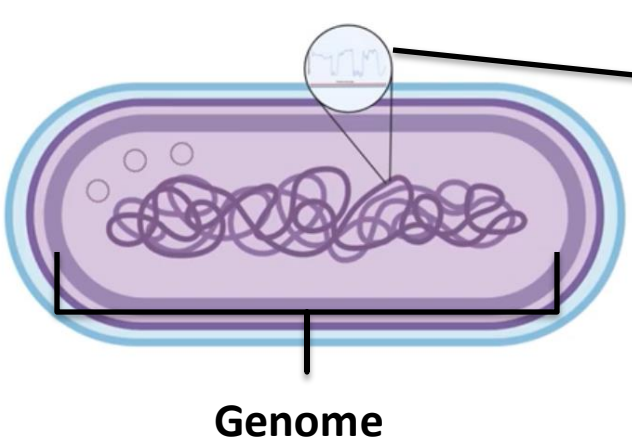
- **Volume** – billions of sequences
- **Complexity** – many genes, organisms, environments
- **Computation** – *requires pipelines to extract biological meaning*

Sample → Sequencing → Raw reads → Biological process

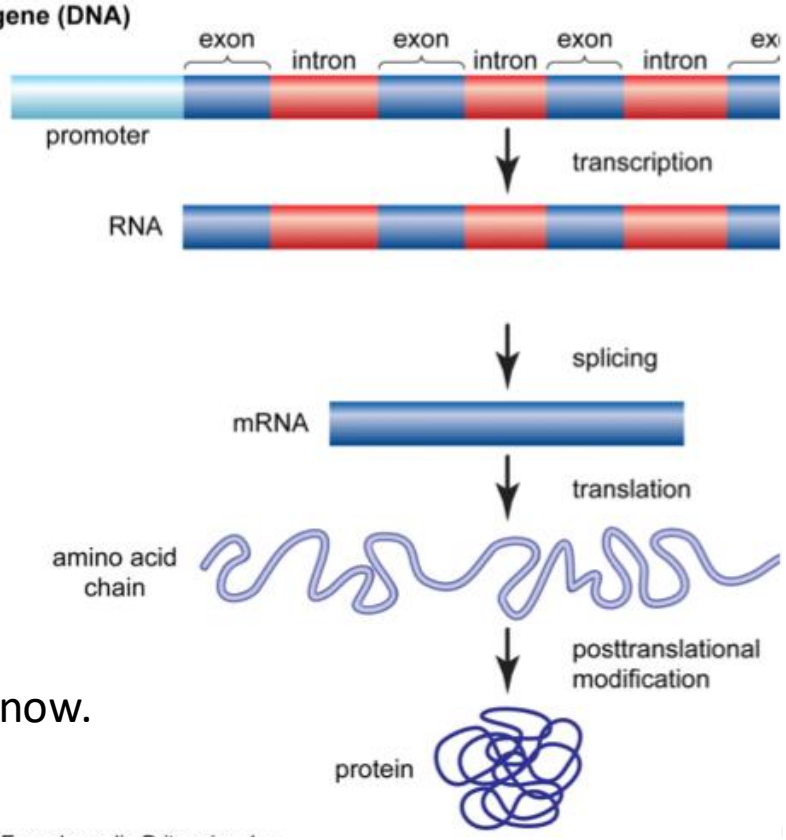


- **High-throughput** sequencing of DNA or RNA samples
- **Different technologies** yield varying read lengths and outputs
- **Massive** amounts of **data** generated from a single run

What is a **gene**, a **genome**, a **transcript**, or a **protein?**



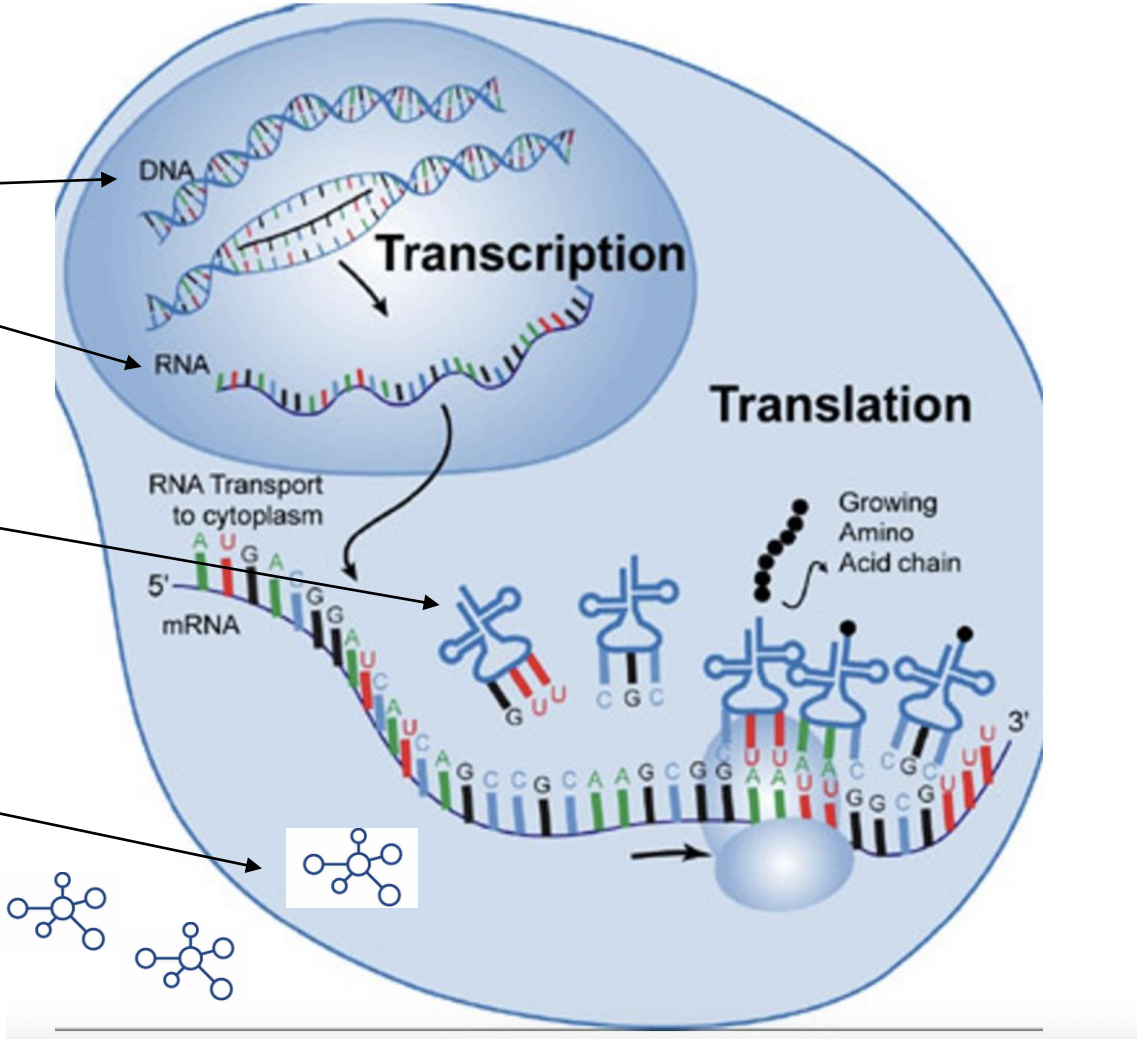
Genome



DNA=relatively stable blueprint
RNA = what's being used
Proteins/metabolites: What's happening now.

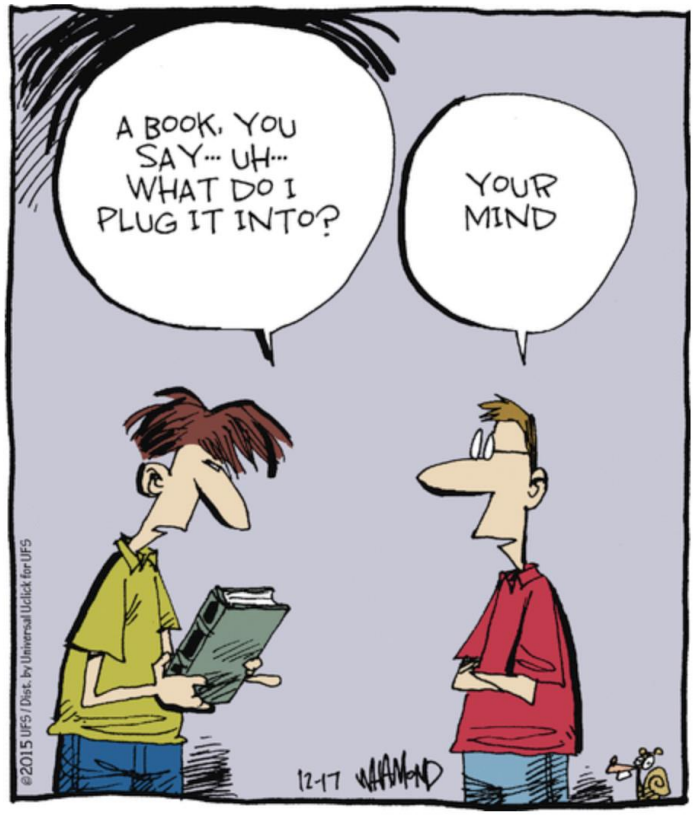
What Do We Mean by 'OMICS'?

- **Genomics** → DNA: Who's there?*
- **Transcriptomics** → RNA: What's active?
- **Proteomics** → Proteins: What's functioning?
- **Metabolomics** → Metabolites: What's produced or consumed?



*amplicon/eDNA: Who is present in an environment

Reality check!!



Omics data can feel like a black box of codes and acronyms…

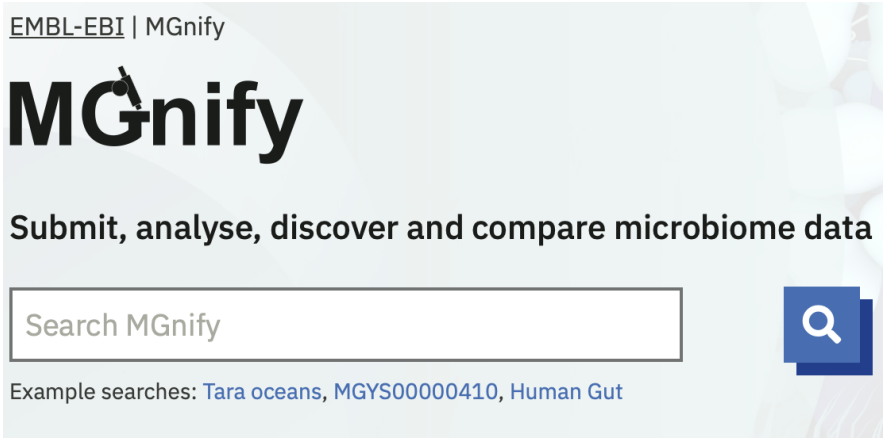
Resources – there are lots!!

DIPLOMICS

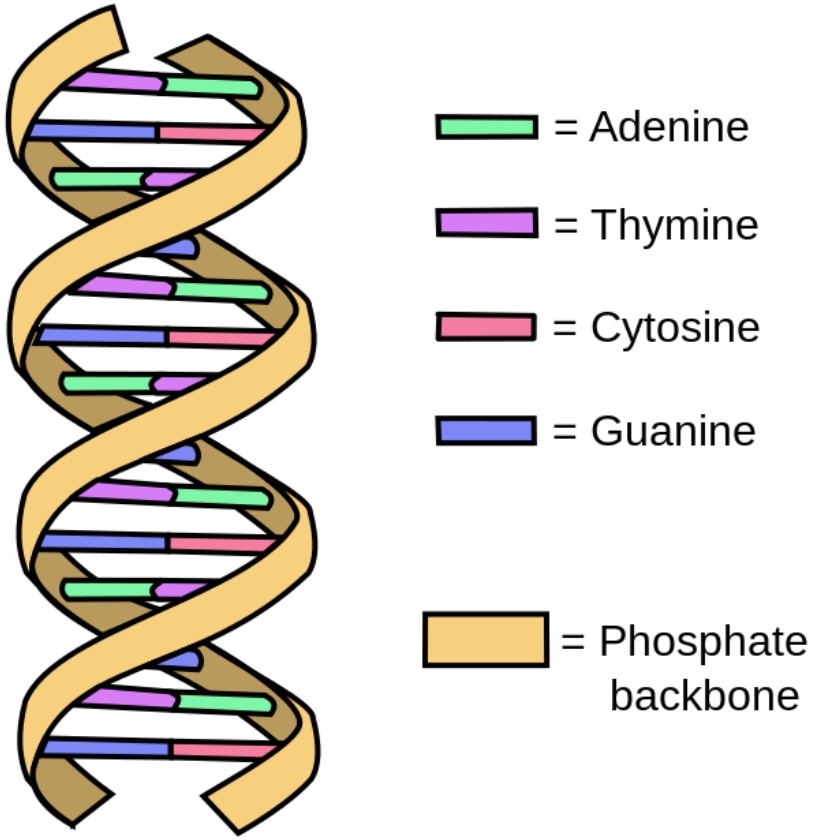
START HERE!!

https://www.diplomics.org.za/contact

Blogs:
https://merenlab.org/posts/
https://bigomics.ch/blog/



DNA



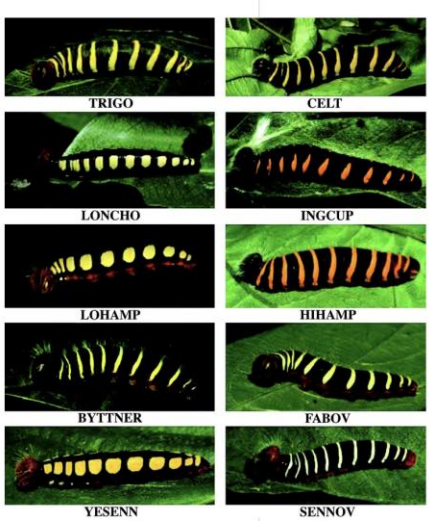
DNA

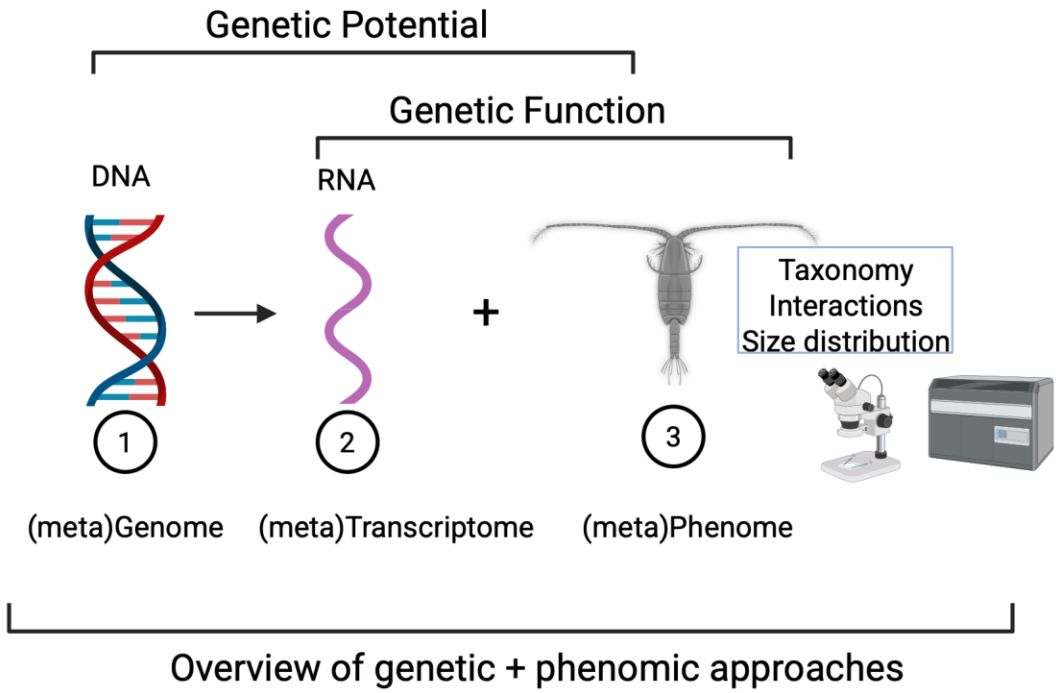
Why OMICS?

Reveals hidden biodiversity (uncultured/cryptic species)

Links molecular function to ecosystem/organism processes

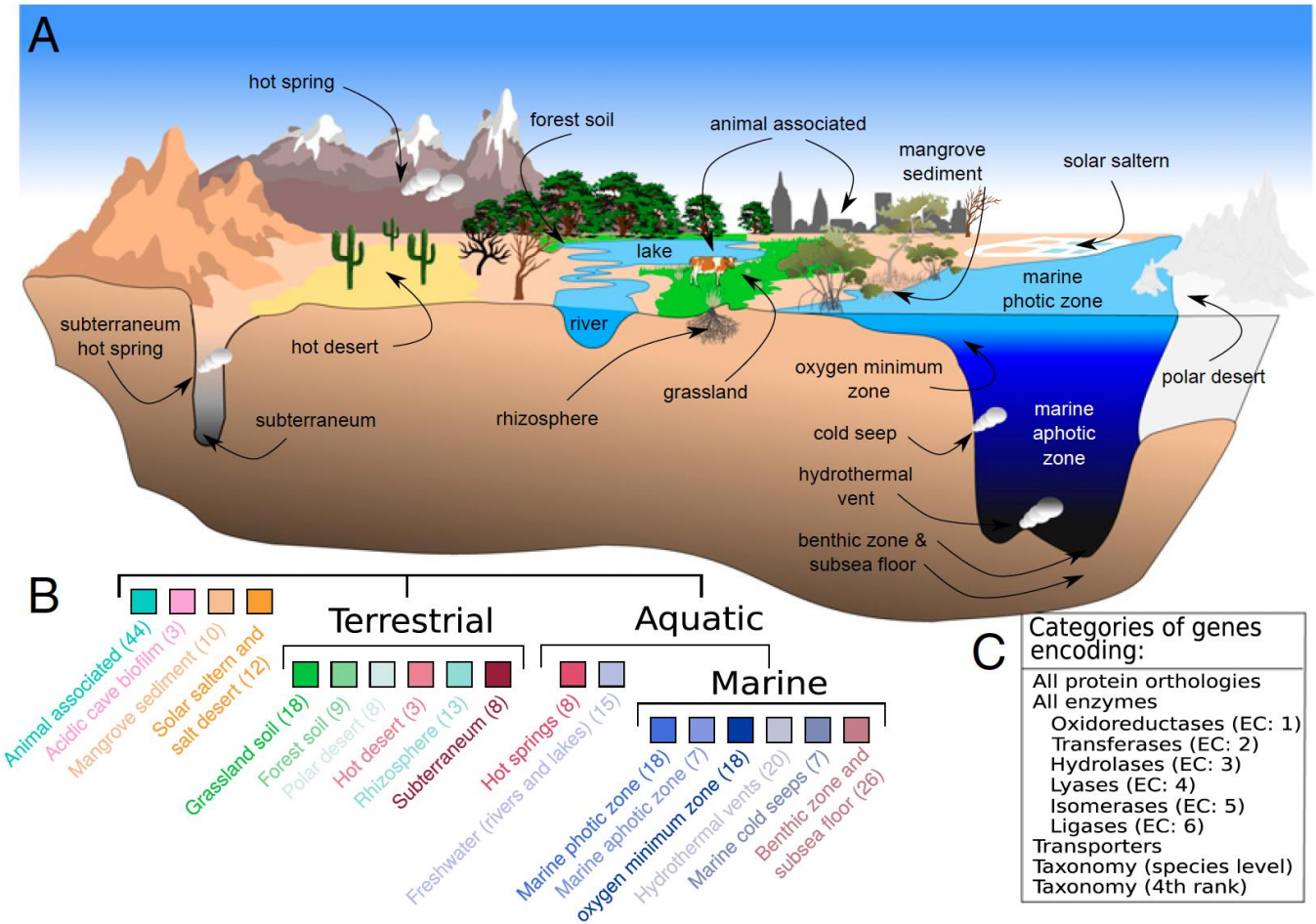
Enables early detection and monitoring (ie: AMR, stress biomarkers as a case study)





Omics turns your organisms and ecosystem into quantitative datasets you can plug into the models you already use.

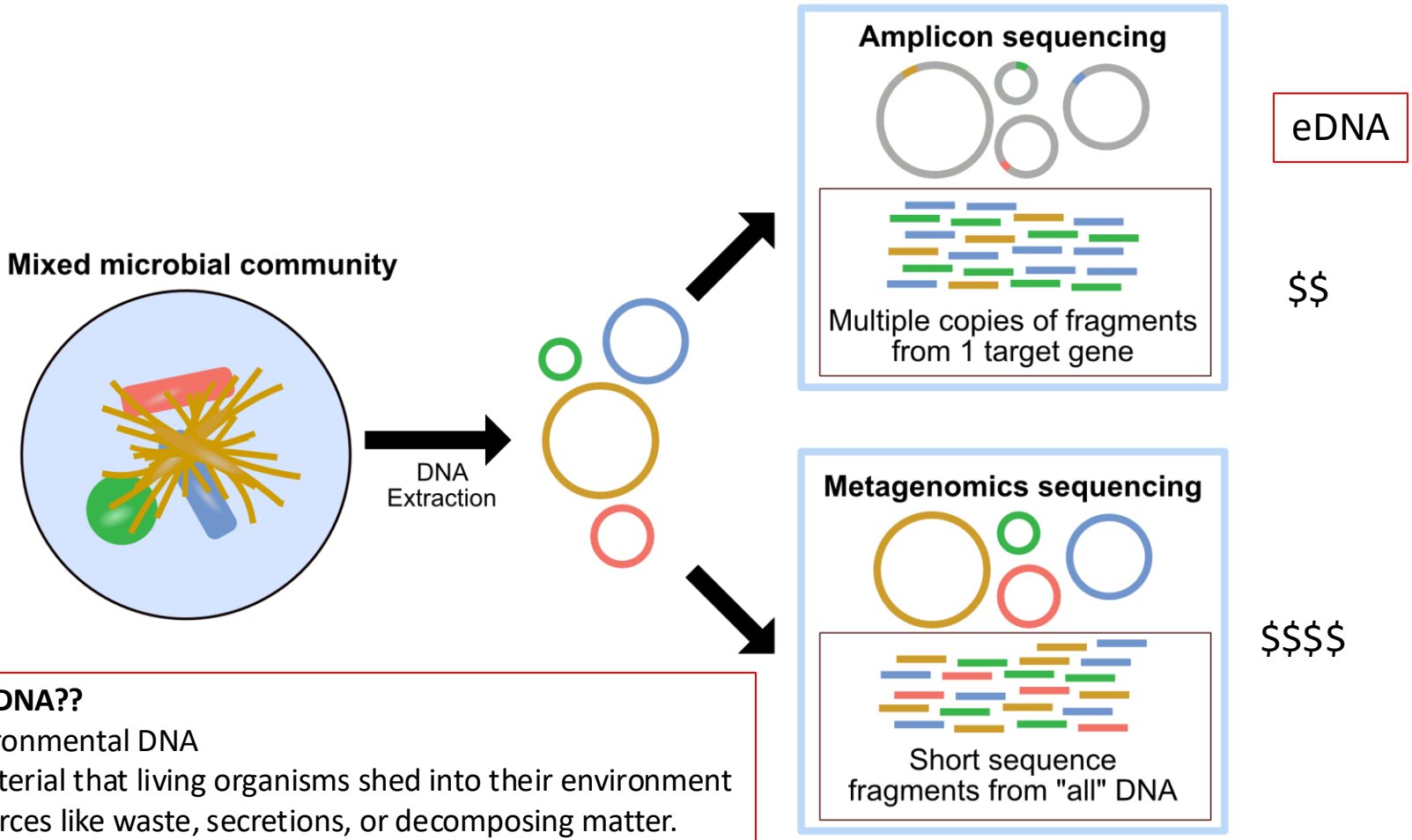
Why OMICS?



Ramirez-Flandes, PNAS, 2019

https://www.ebi.ac.uk/ena/browser/home

What data product should I use?

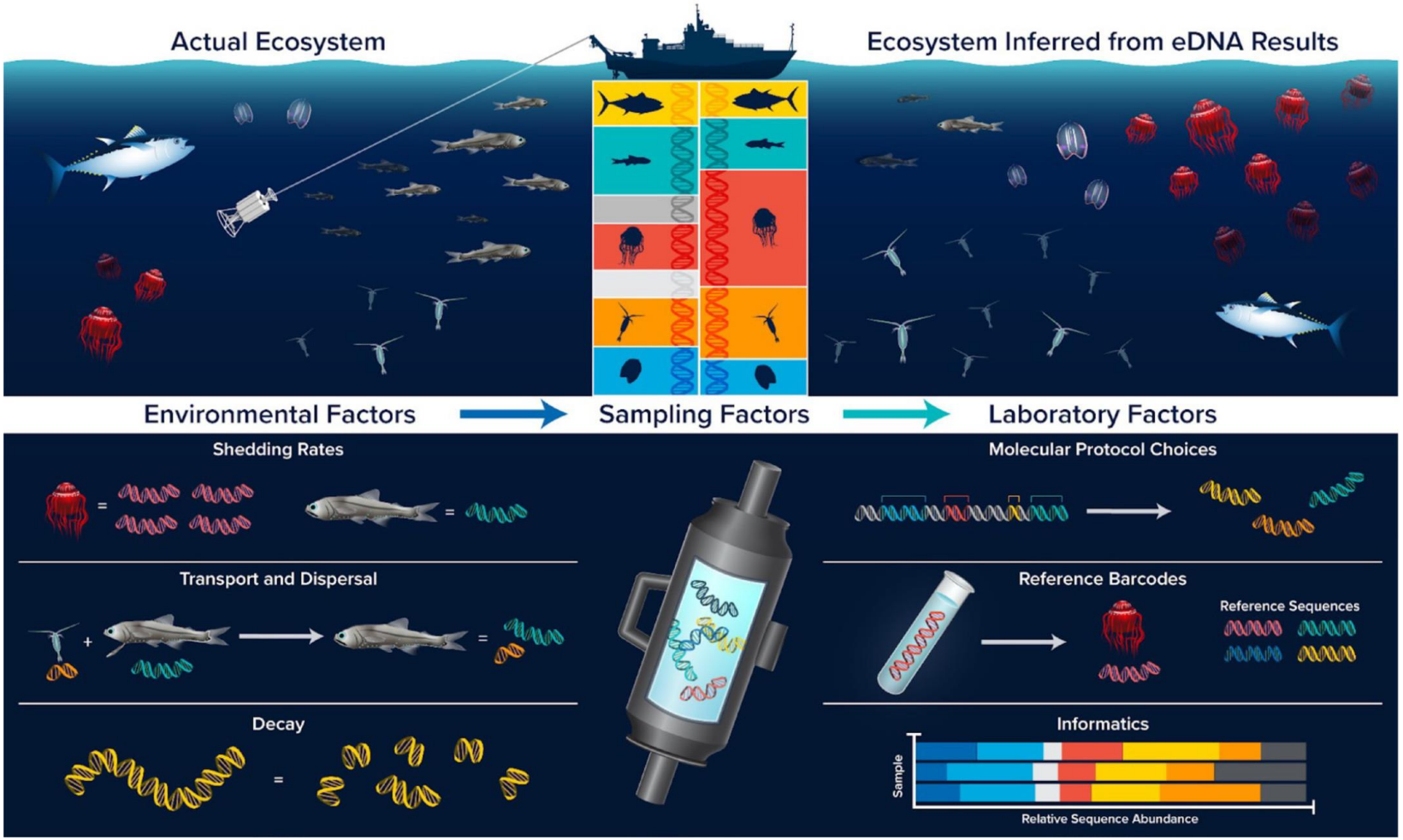


What about eDNA??

eDNA, or environmental DNA

- Genetic material that living organisms shed into their environment through sources like waste, secretions, or decomposing matter.
- Allows researchers to identify the types of species present in an area without needing to sample the organisms directly.
- Cost-effective and non-invasive tool for biodiversity monitoring, invasive species detection, and wildlife surveillance

eDNA realities



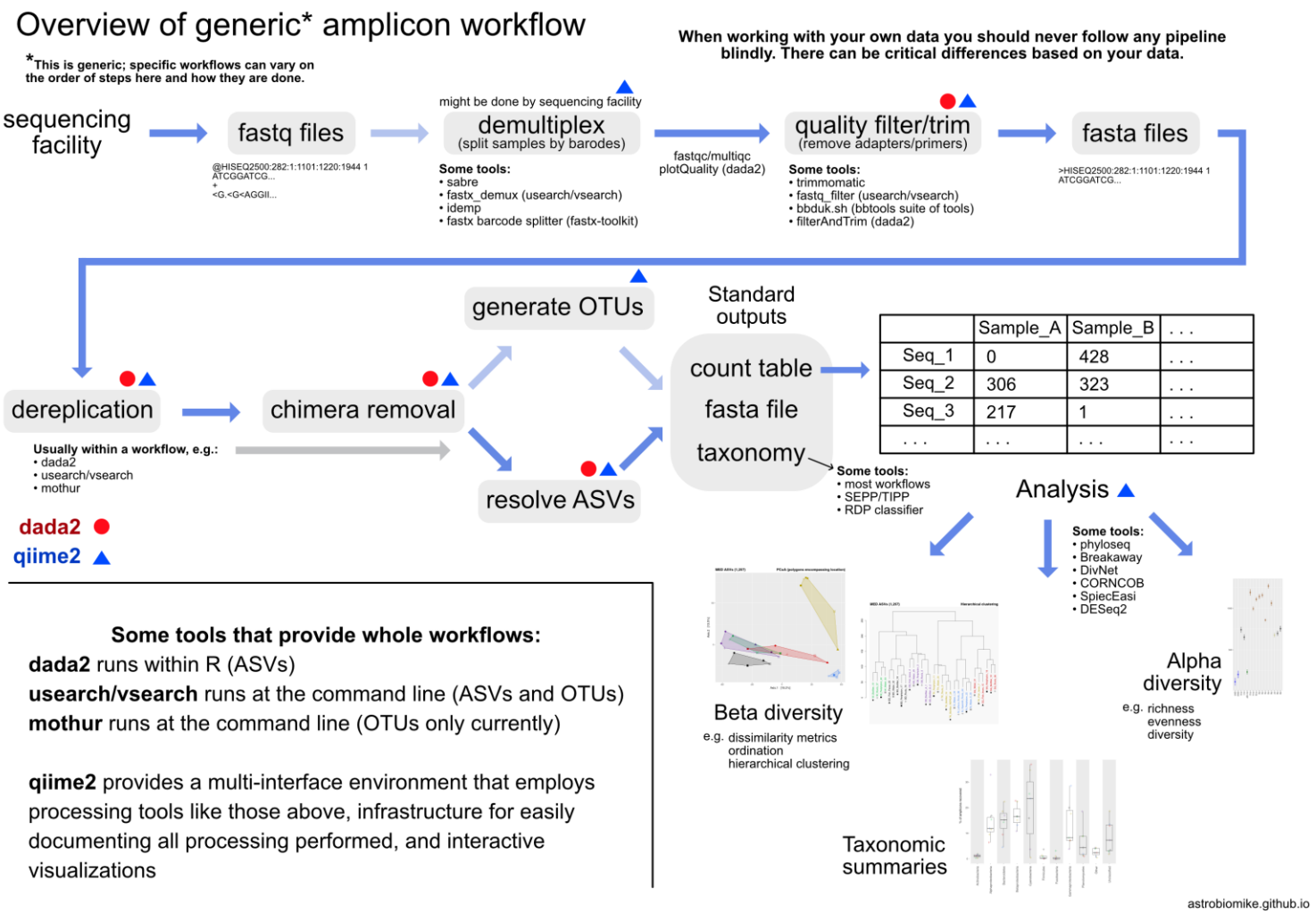
Detection limits, degredation and transport mean 'present' ='local and abundant'
+ your species ID is only as good as your chosen database and primers!

The OMICS Workflow — The Big Picture

Sampling → Data generation → **Processing** → **Annotation** → **Statistical analysis** → **Interpretation**

Same logic across ALL OMICS types

Amplicon/eDNA overview:



File types…

Fastq files = raw sequence reads from the machine
Fasta file: Your sequences!

FASTA file:

HEADER

SEQUENCE

Fasta file extensions:

Extension	Meaning	Notes
fasta, fa[9]	generic FASTA	Any generic fasta file. See below for other common FASTA file extensions
fna	FASTA nucleic acid	Used generically to specify nucleic acids.
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome.
faa	FASTA amino acid	Contains amino acid sequences. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

Ampliseq: one stop shop:

https://nf-co.re/ampliseq/2.6.1/

```
nextflow run nf-core/ampliseq \
   -profile singularity \
   --input "samplesheet.tsv" \
   --FW_primer GTGYCAGCMGCCGCGGTAA \
   --RV_primer GGACTACNVGGGTWTCTAAT \
   --metadata "data/Metadata.tsv" \
   --outdir "./results"
```

How many genes regions in your sample?

Spreadsheet indicating file path to your sequences

Primer set that was used

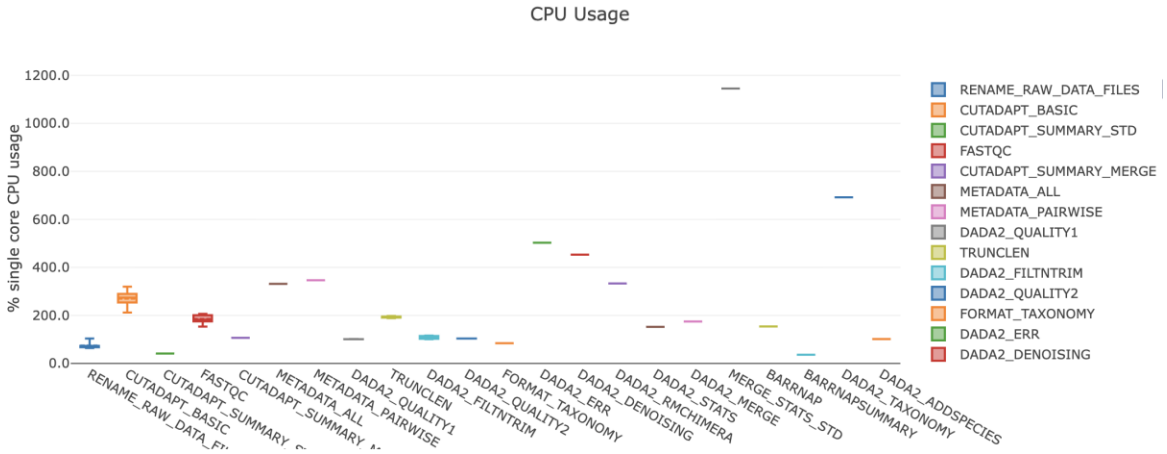
Standardised, documented pipeline = less time reinventing wheels (github is a **great** resource here)

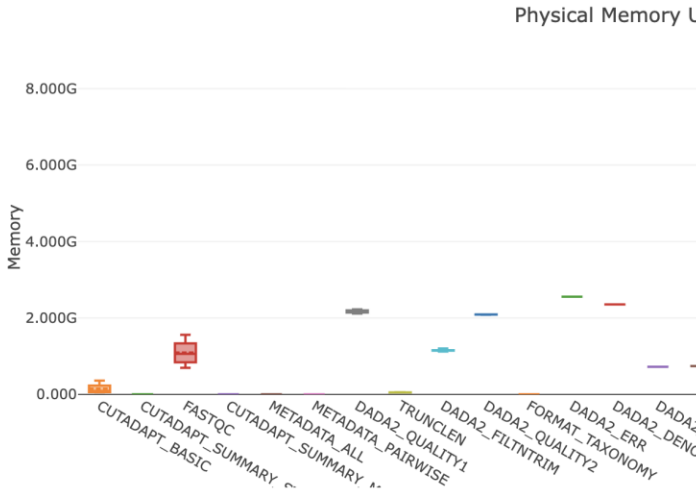
Produces summary reports, diversity metrics, taxonomic tables ready for R

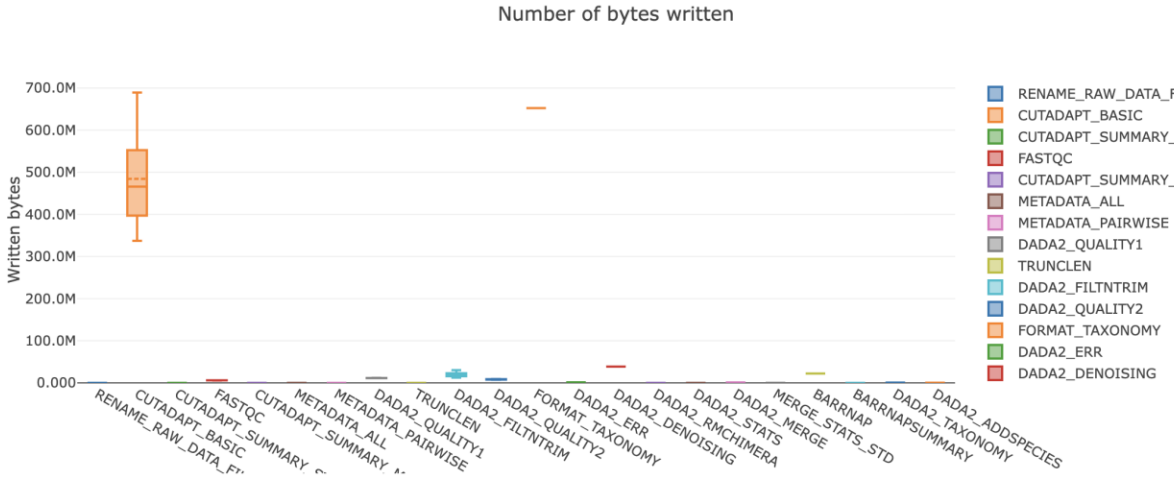
You will need: Raw sequence files, primer info and a sample metadata sheet

R packages:
Phyloseq
Microeco

QC Summary reports: important to check

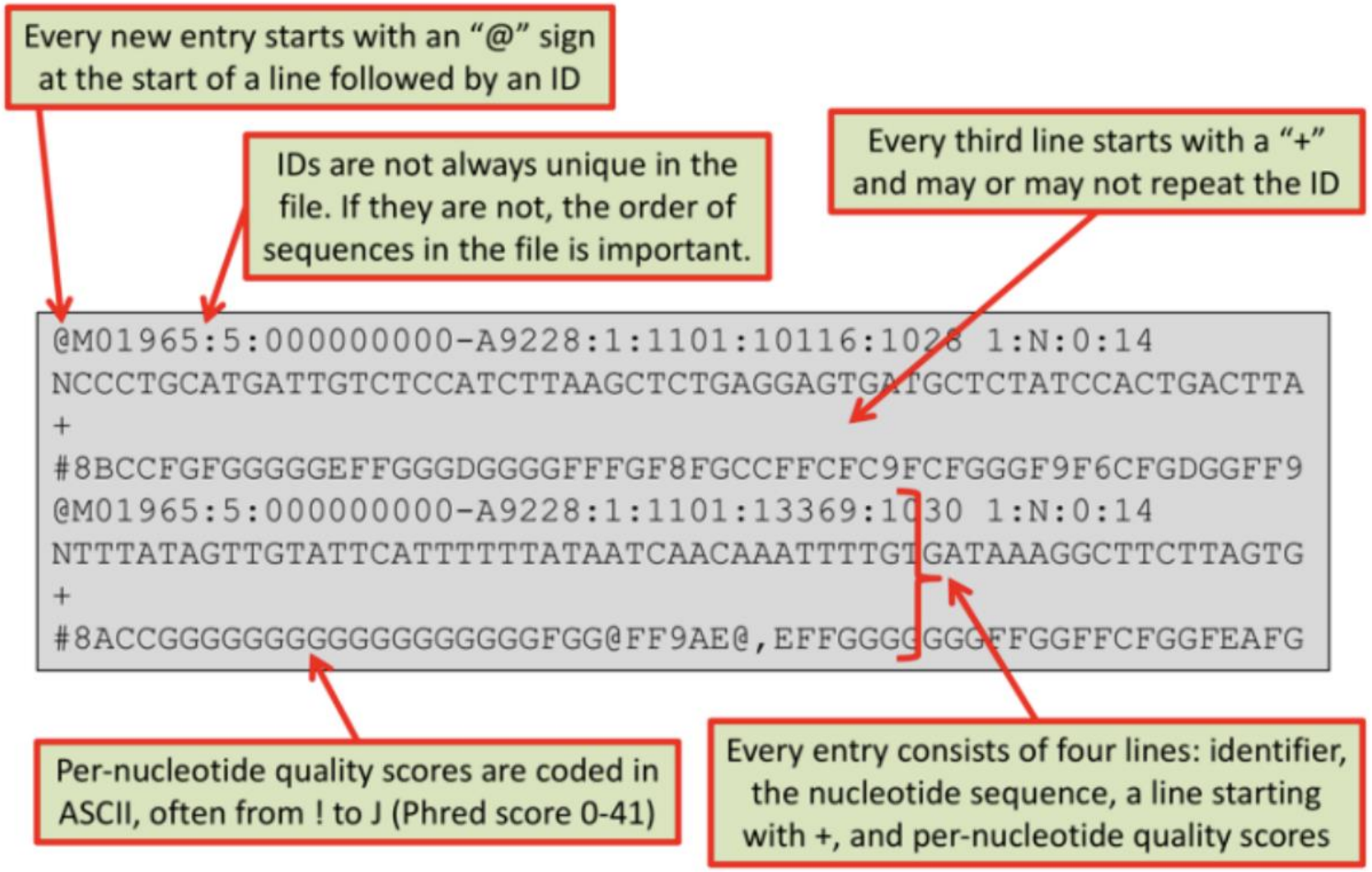




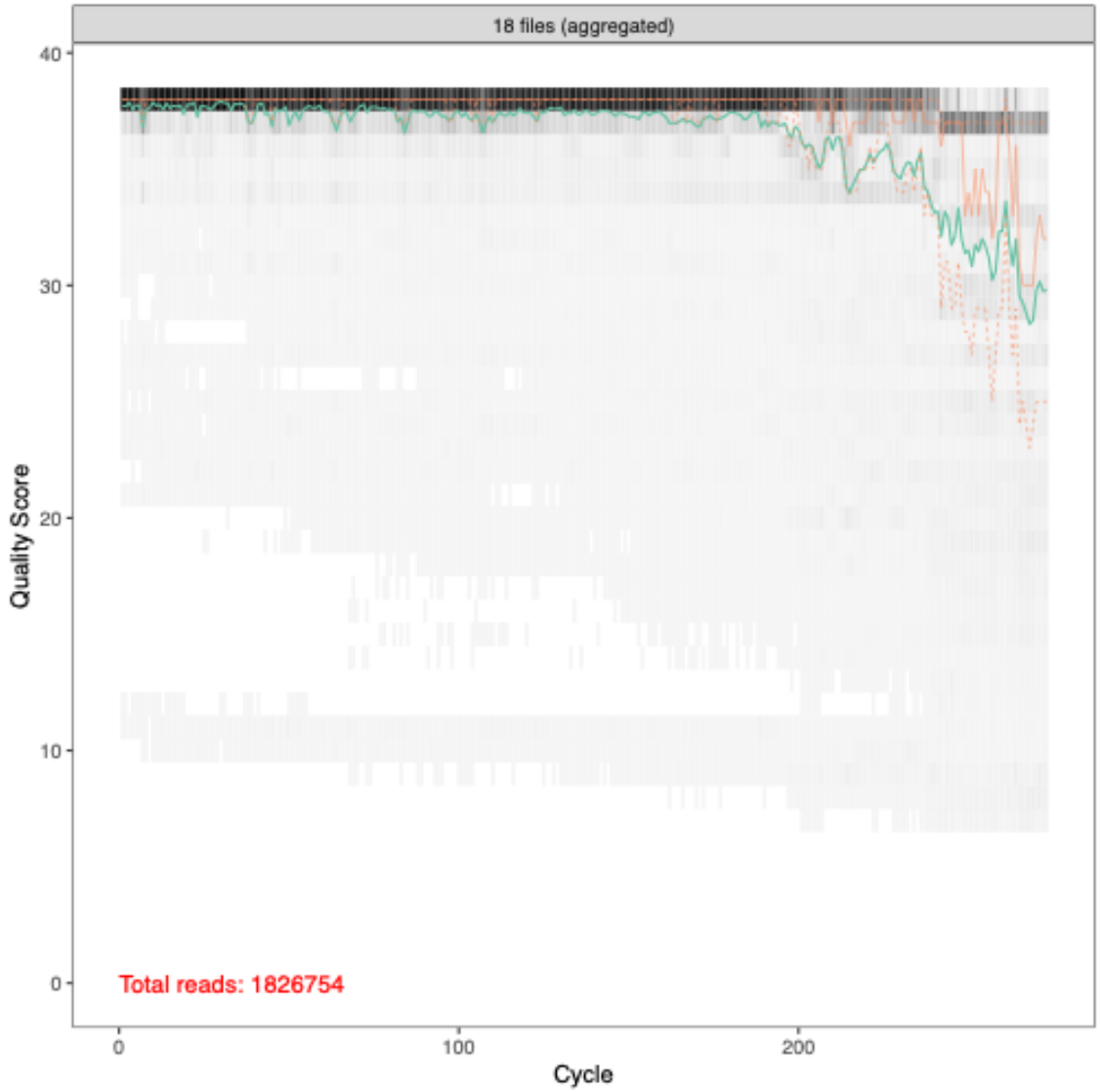


1. Quality filtering: QC

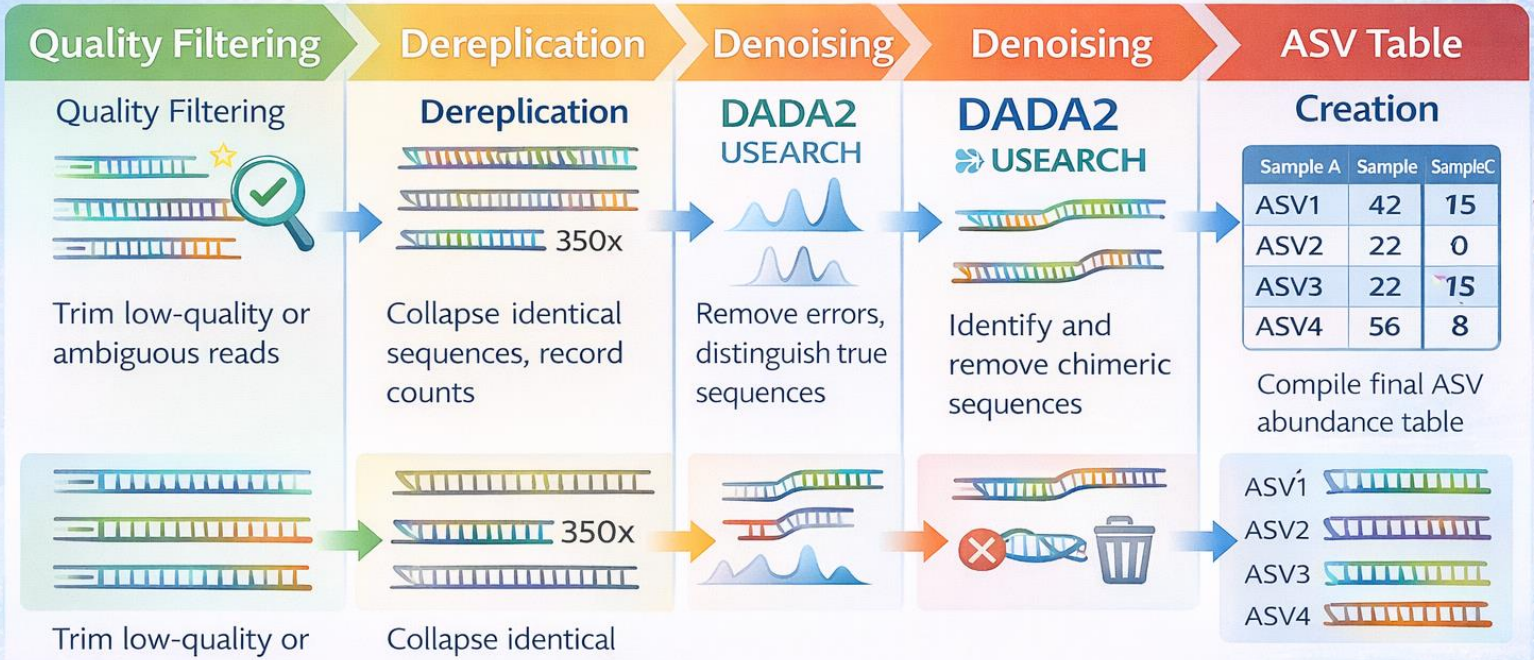
FASTQ file---FASTA file:



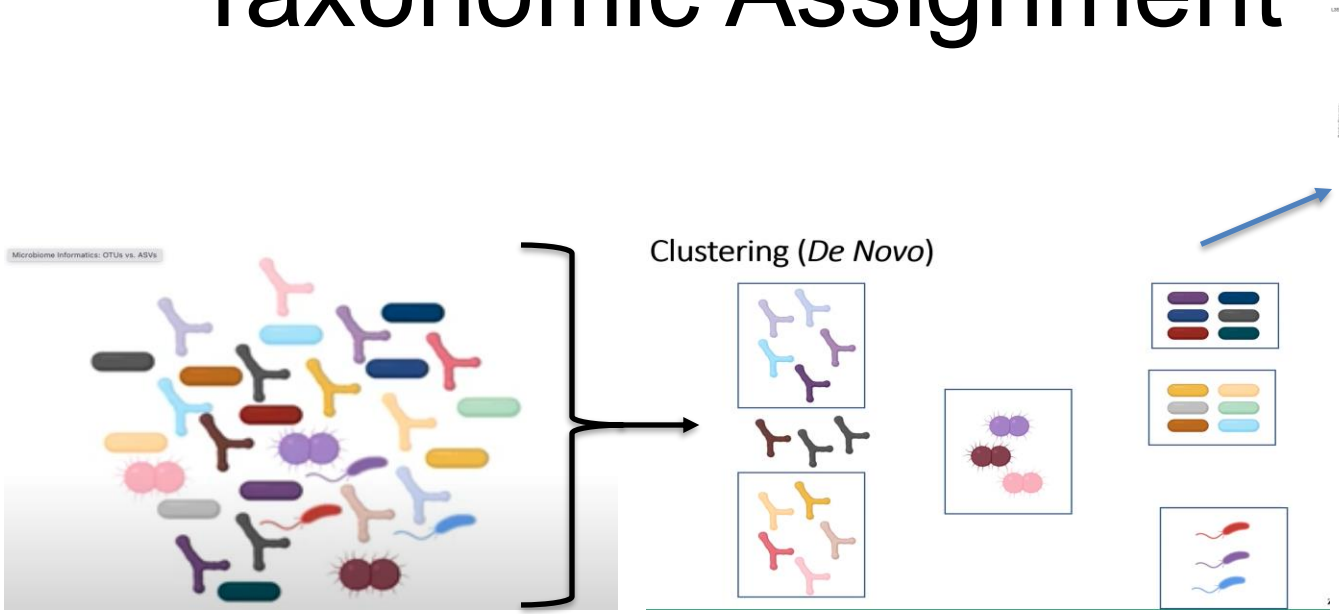
Preprocessing – Quality scores



Steps in ASV Generation



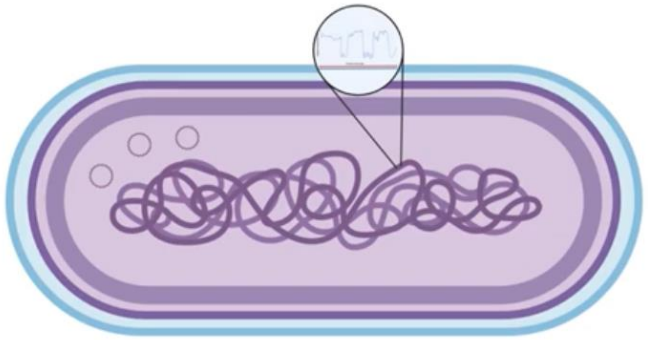
Under the hood: ASV's and Taxonomic Assignment



ASV= Amplicon sequence variant (statistical inference of true biological sequences)

A few warnings…

Amplicon sequence variant



A Quick Note

16S Sequencing Challenges

Targeted sequence with a few bases differentiating species

- Sequencing is imperfect
  - Illumina usually makes some base call errors
  - Nanopore makes more
  - Errors are not necessarily evenly-distributed
- We do not want errors to be confused with real diversity/new species

FOR OUR PURPOSES:

= 1 sequence and one species

REAL BACTERIA

One species can have multiple, different copies of a gene

Two similar species may share an identical sequence

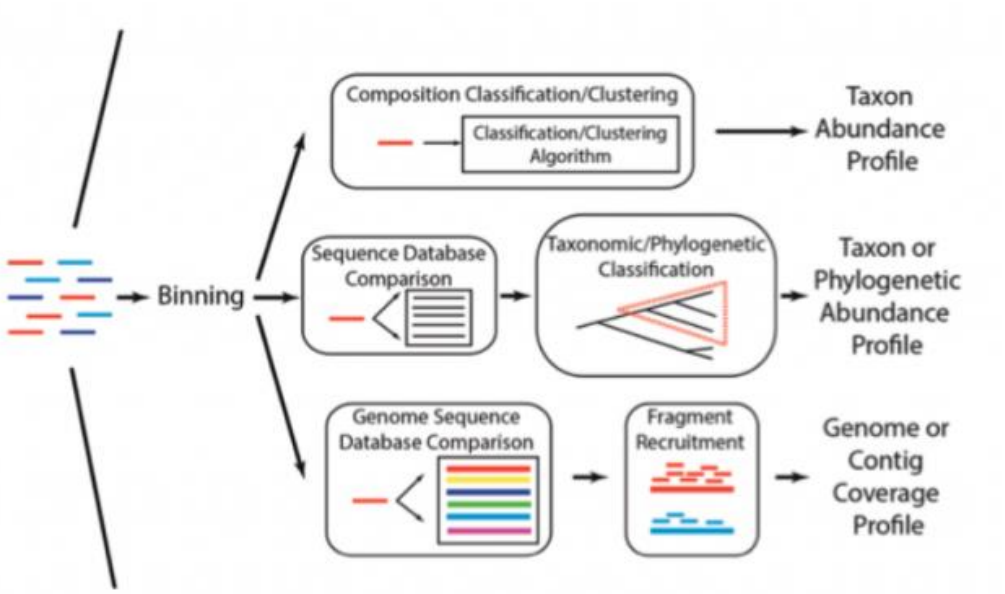
ASV inference

ASV_table

ASV_ID	E10a	E10b	E11a	E11b	E13a	E13b	E14b	E14c	E4a	E4astar	E4b	E4bstar	E5a	E5b
cae609ea893f535d61720405d255b8ad	4285	9764	6399	5404	1506	1307	26962	35226	10672	5626	5455	3756	18177	17938
9f2865c7bbbce298473d4ae532a335e6	18204	15465	107	27	7611	8863	17	23	122	23631	22379	27871	131	59
0002c03993330198b02457afef4d00e5	170	150	5446	5811	843	567	310	283	6422	559	227	108	10923	3508
d11d8a29479c2d36943b2b68362221f4	4088	1621	1919	2420	6356	5455	122	147	580	1675	5628	1376	660	260
ffb5ad37299ba5fbe1946a5639f722cd	0	0	29	22	195	13	24000	8291	181	0	0	0	246	3357
eca00533a595089a33766b2f14a1961c	112	46	4382	3259	199	195	468	417	3492	272	152	161	5069	3681
5b7ab8aab9e49afadb578146ef26e5a7	33	0	3975	2728	134	0	3500	2506	1663	0	39	0	2518	2335
9fe4b245dc54657f85440234d48a02ed	1919	1321	125	0	2540	3304	0	0	34	3140	2352	1622	61	0
0df0db8002857b2a9bfe61fd662710da	7	0	931	445	0	0	17	31	4132	12	4	0	6162	4099
e3ea1a931ec40e4c21752dac82031b8d	4153	1523	0	0	598	289	0	6	0	680	5578	592	0	0

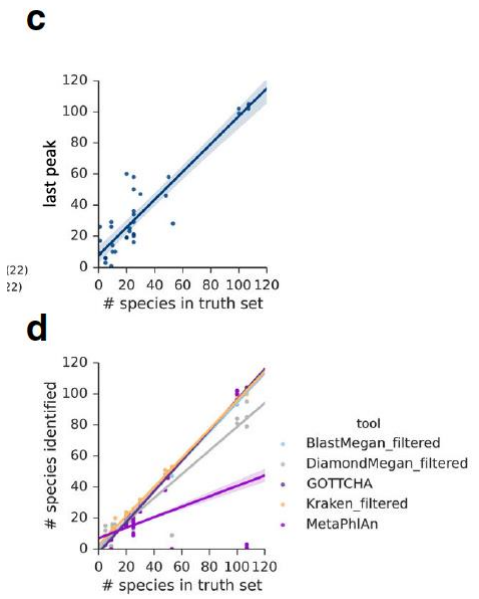
Note: Normalisation. Relative abundance/ Hellinger… have a look at your dataset !

Taxonomy/gene calling



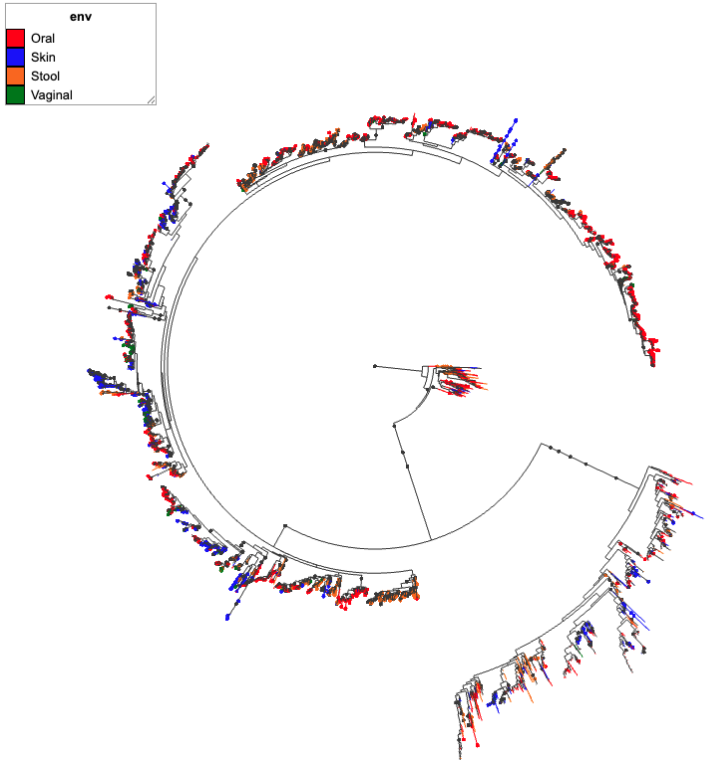
Comprehensive benchmarking and ensemble approaches for metagenomic classifiers

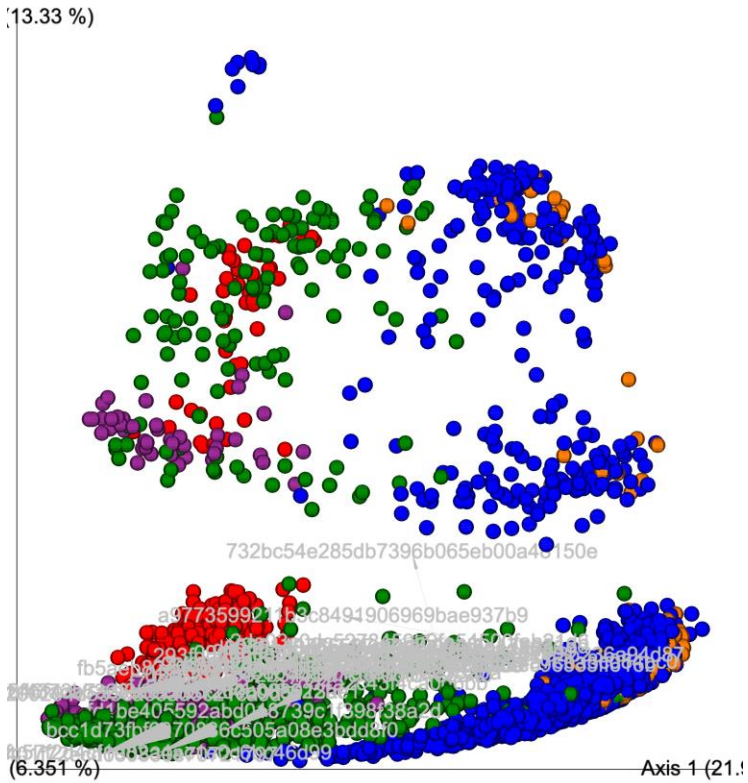
Alexa B. R. McIntyre[1,2,3], Rachid Ounit[4], Ebrahim Afshinnekoo[2,3,5], Robert J. Prill[6], Elizabeth Hénaff[2,3], Noah Alexander[2,3], Samuel S. Minot[7], David Danko[1,2,3], Jonathan Foox[2,3], Sofia Ahsanuddin[2,3], Scott Tighe[8], Nur A. Hasan[9,10], Poorani Subramanian[9], Kelly Moffat[9], Shawn Levy[11], Stefano Lonardi[4], Nick Greenfield[7], Rita R. Colwell[9,12], Gail L. Rosen[13*] and Christopher E. Mason[2,3,14*]

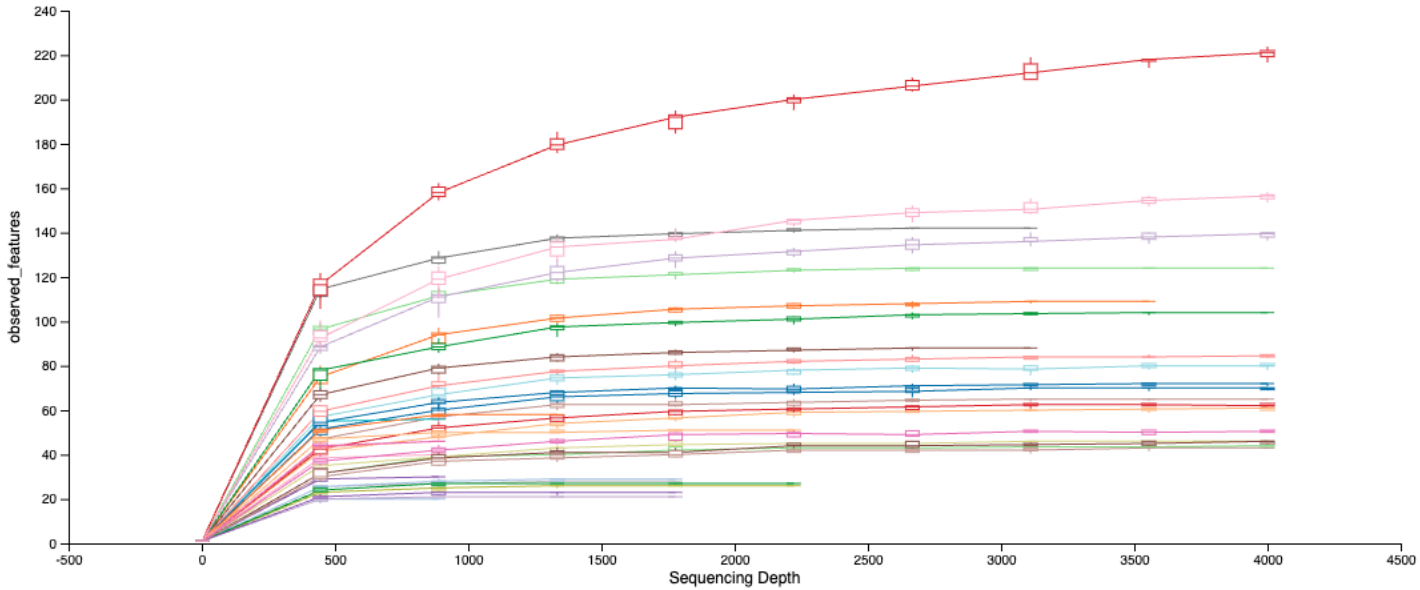


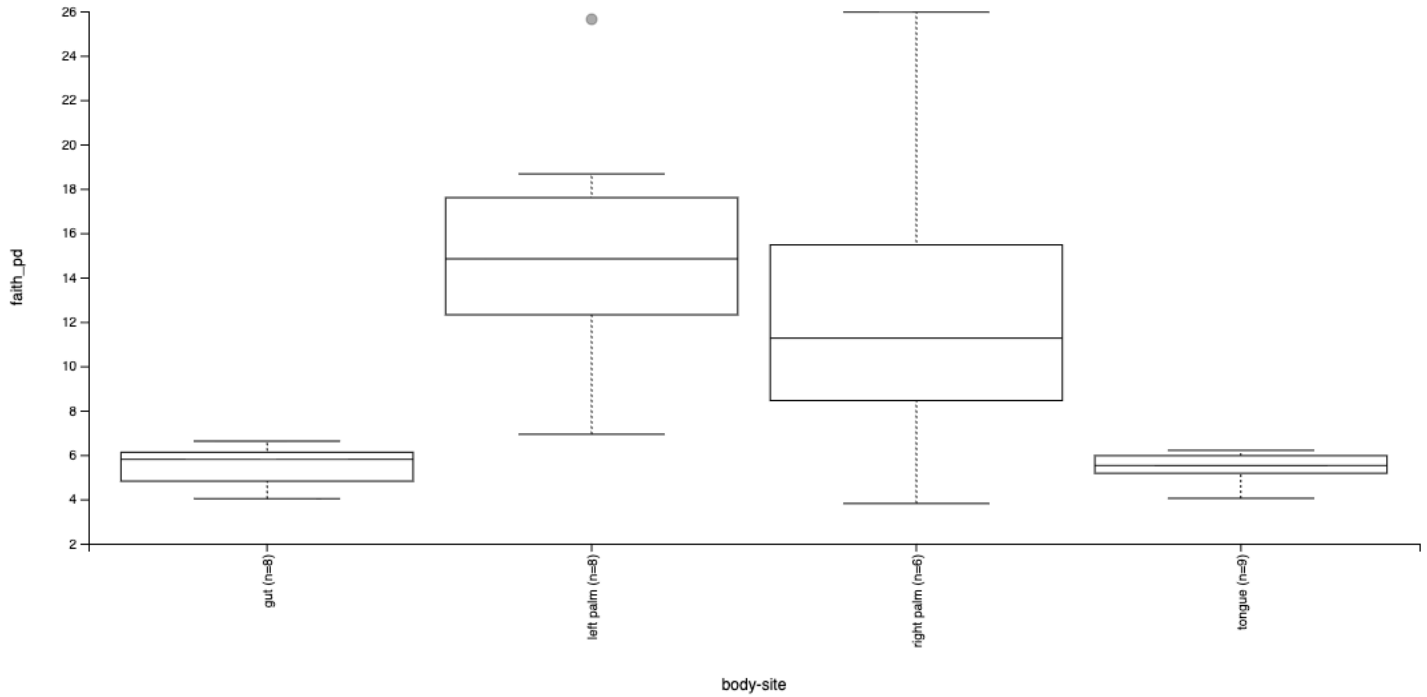
A mock community is key here

What more can we do?



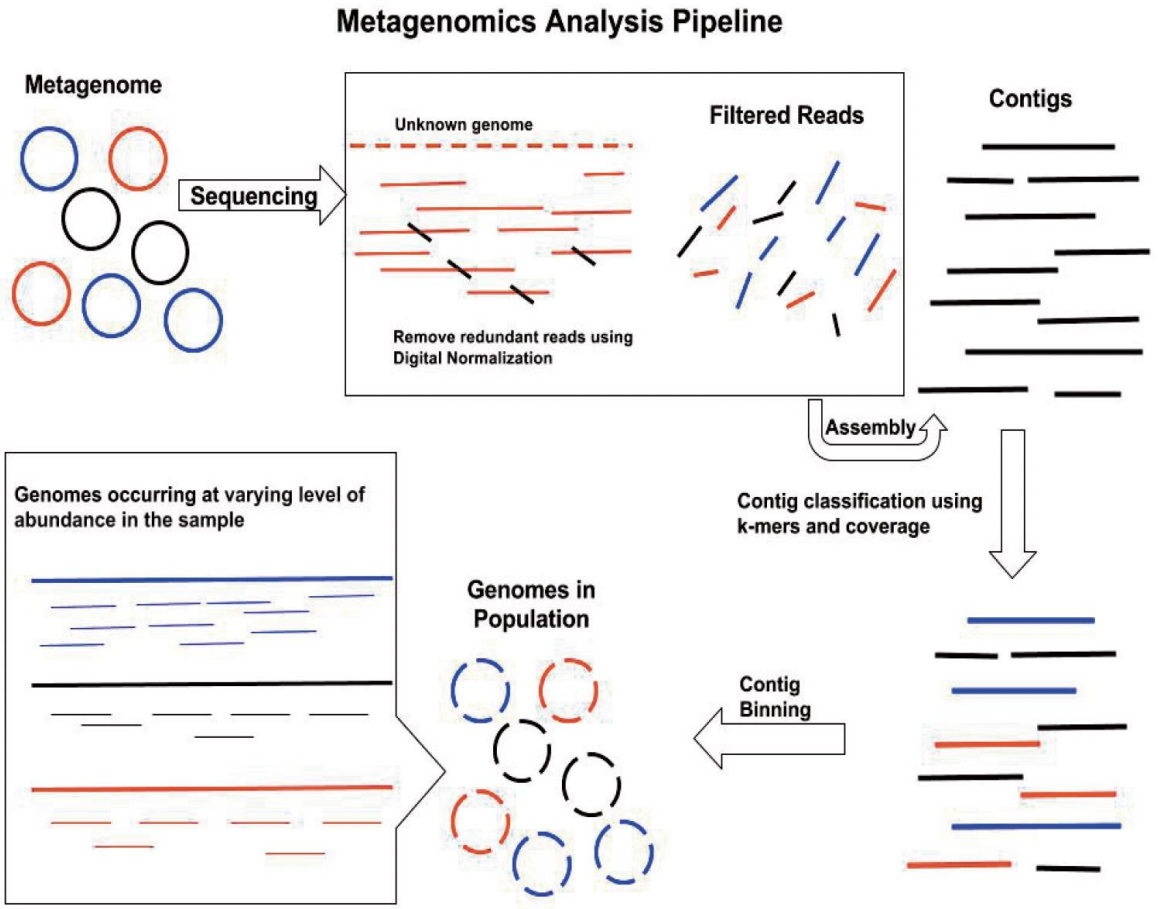



observed_features
Sequencing Depth
0
20
40
60
80
100
120
140
160
180
200
220
240
-500
0
500
1000
1500
2000
2500
3000
3500
4000
4500


faith_pd
body-site
2
4
6
8
10
12
14
16
18
20
22
24
26
gut (n=8)
left palm (n=8)
right palm (n=6)
tongue (n=6)

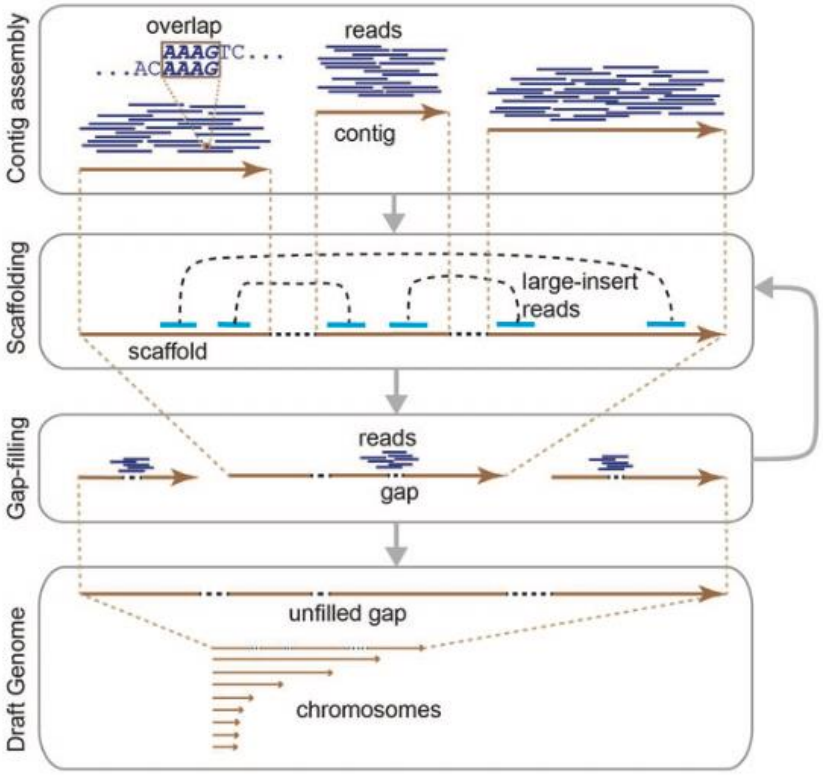
Genomics & Metagenomics

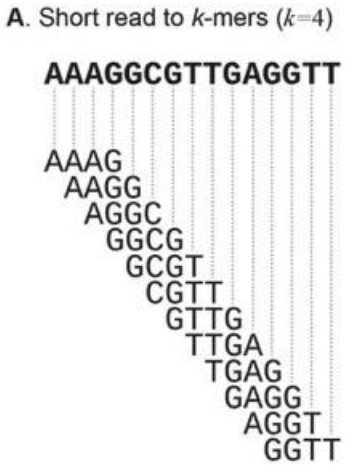
- Sequencing DNA from isolates or communities
- Produces FASTQ files → **assembly** → annotation → gene counts
- Tools: FastQC, MEGAHIT, MetaSpades MetaBAT, eggNOG-mapper

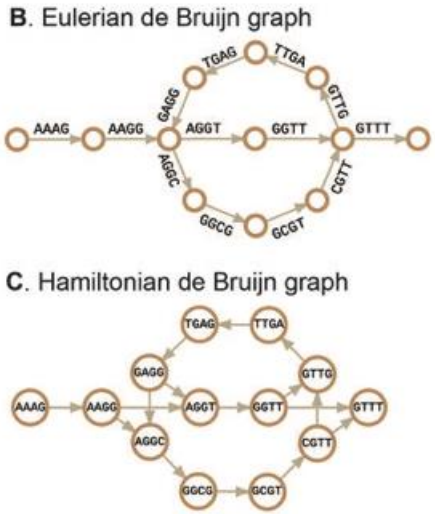


ASSEMBLY….

‘Denovo’ versus ‘Mapping’







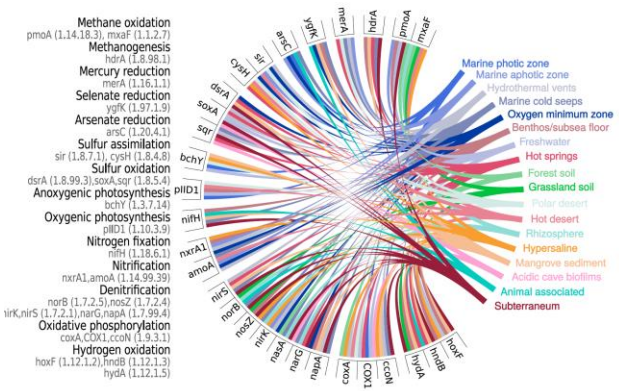
Sohn and Nam, Briefing in Biofinformatics, 2016

Transcriptomics (Who's Active?)

- RNA → cDNA → sequencing
- Key step: differential expression across conditions (careful!!how do we normalize??: https://bigomics.ch/blog/why-how-normalize-rna-seq-data/)
- Analogy: same instrument (genome), different songs (transcripts) will tell a different story

Statistical Thinking in OMICS

- High-dimensional abundance tables
- PCA, NMDS, clustering, regression, random forests
- Ecological parallels: taxa ⇄ genes; community ⇄ pathways



End product:

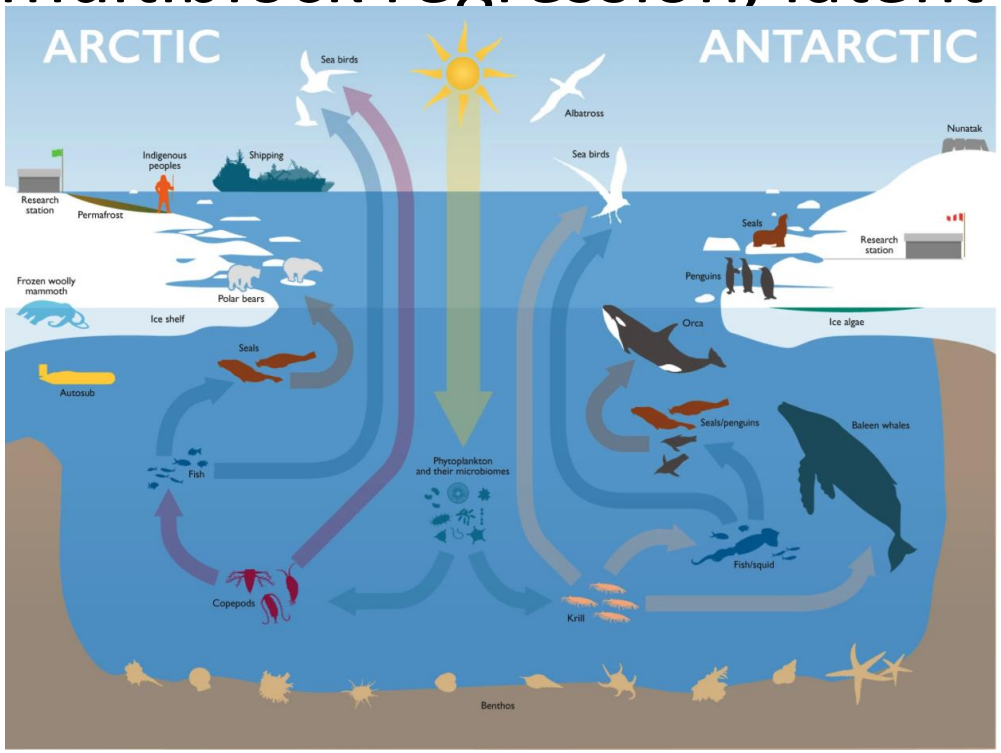
Omics gives you big multivariate matrices; your PCA/NMDA/regression/… skills still apply here.

Typical Analytical Questions

- Diversity: how rich and even is the community?
- Function: which pathways dominate where?
- Differential features: which respond to stress?
- Linkages: environment → function → ecosystem process

Integrating Multiple ‘OMES’

- Multi-omics connects layers for causality
- Methods: MixOmics, MOFA, DIABLO
- View: multiblock regression, latent variables



Clark et al, Nature Communications 2023

Common Pitfalls

- **Compositionality** & false correlations
- **Batch effects** / unequal coverage
- **Unknown genes**
- Over-interpretation of small effects

These are design and analysis issues – not reasons to avoid omics!

Computational & Ethical Landscape

- HPC, cloud pipelines, reproducibility (Nextflow, Snakemake, Docker)
- FAIR + CARE data principles:
- FAIR + CARE principles help ensure data are re-usable and that communities benefit, not only high-income labs.

Future Horizons

- Portable sequencing (Nanopore/Pacbio field kits)
- AI/ML for annotation & pattern discovery
- Predictive eco-models linking OMICS → fluxes → climate

Summary

- OMICS = quantitative molecular ecology
- Pipelines bridge raw data to ecological insight
- Integration & reproducibility are the next frontier

Acknowledgements & Further Resources

MARiS | SEEC | DIPLOMICS

Key references:

Knight et al. 2018 – Nature Biotech (Microbiome analysis)

Misra et al. 2022 – Nature Methods (Multi-omics)

Love et al. 2014 – DESeq2; Rohart et al. 2017 – mixOmics

Pipelines:

https://www.geneious.com

https://mothur.org

https://qiime2.org

https://nf-co.re/ampliseq/2.6.1/

DATABASES

https://www.ncbi.nlm.nih.gov

https://www.ddbj.nig.ac.jp/index-e.html

BOLDSYSTEMS

Boldsystems.org