Metagenomics Lesson 1

BVCN

biovcnet.github.io

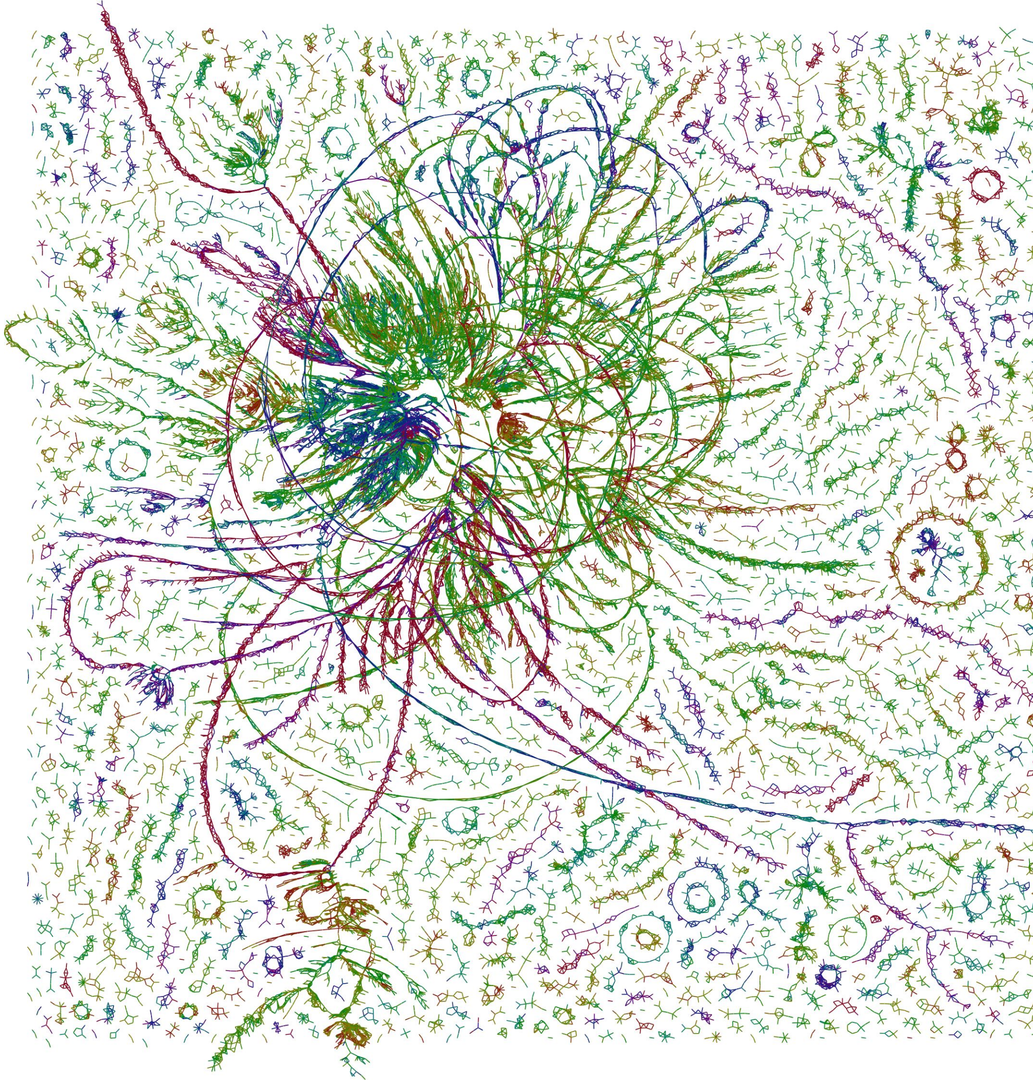
Metagenomics Lesson 1

What is Metagenomics?

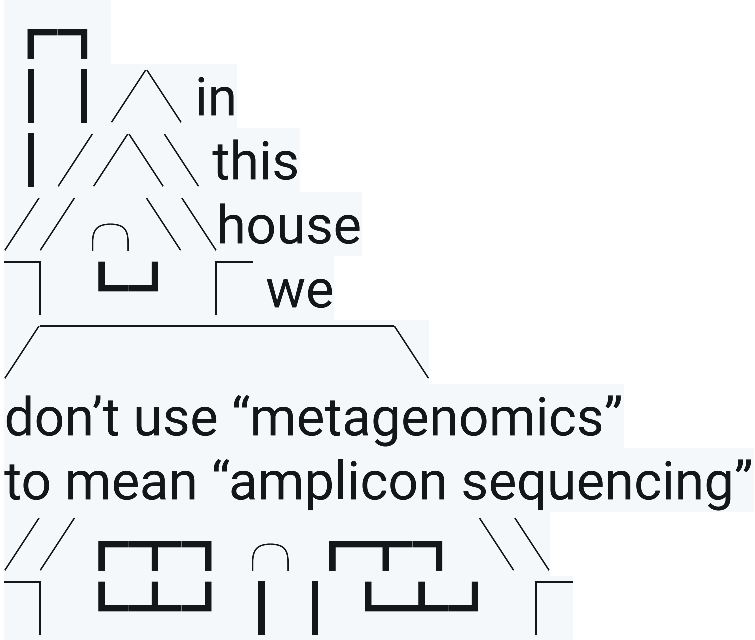
What kinds of questions can metagenomics be used to answer?

Is metagenomics right for me?

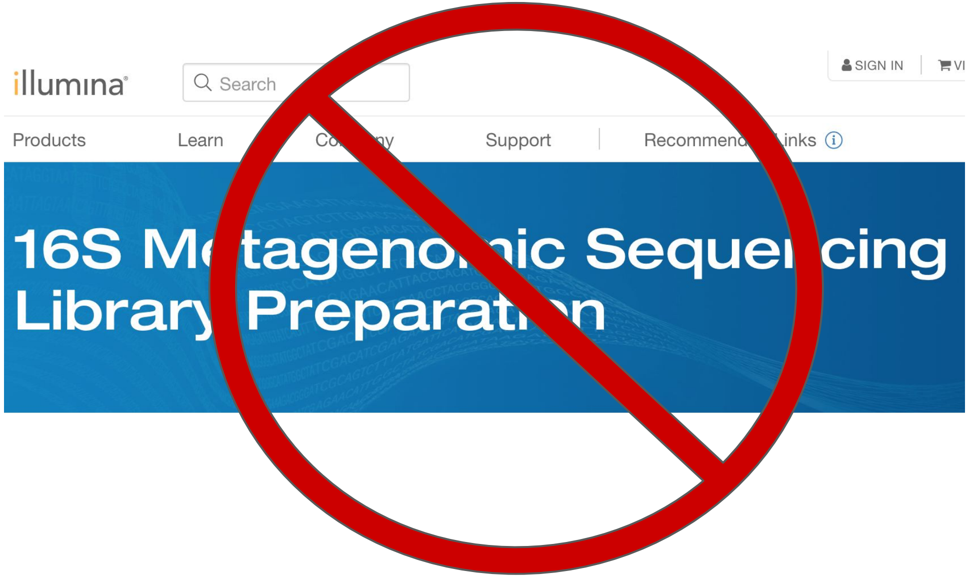
Yay! You got data! Now what?



What is Metagenomics?

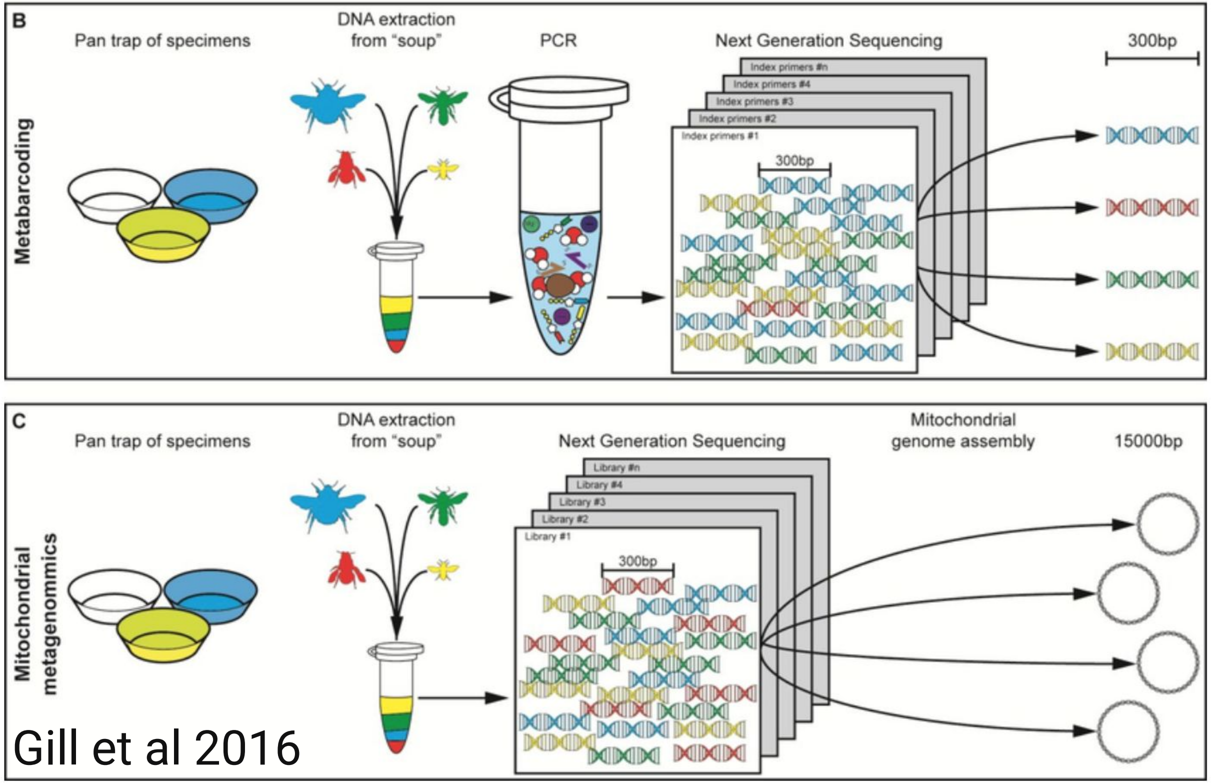


in this house we don’t use “metagenomics” to mean “amplicon sequencing”



16S Metagenomic Sequencing Library Preparation

What is Metagenomics?

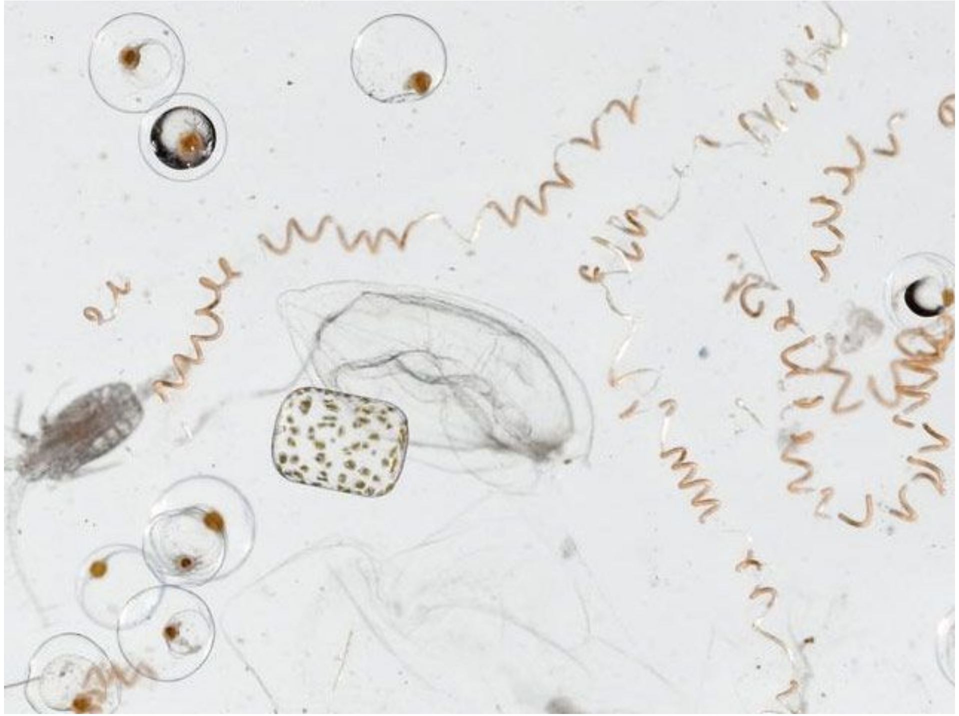


Gill et al 2016

“Amplicon sequencing” or “metabarcoding” is sequencing a specific target region from many genomes (e.g. 16S rRNA gene, *nifH* gene)

“Shotgun Metagenomics” is (incomplete) sequencing of a mixture of genomes using an untargeted approach

Incomplete because one drop of seawater contains about...



10^6 bacteria/mL * 3×10^6 bp/bacteria = 3×10^{12} bp/mL

+ 10^3 euks/mL * 3×10^8 bp/euk = 3×10^{11} bp/mL

= 3.3 Tbp/mL

= 210 MiSeq runs/mL

= 0.5 NovaSeq run/mL

0.5 Genbank/mL = 0.2 SRA / L

One scoop of soil contains about...

10^{10} bacteria/g * $4x10^{6}$ bp/bacteria = $4x10^{16}$ bp/g

+ 10^{5} euks/g * $3x10^{8}$ bp/euk = $3x10^{13}$ bp/g

= 40 Pbp/g

= 2.6M MiSeq runs/g

= 6,000 NovaSeq runs/g

6,000 Genbank/g = 2.8 SRA / g

One pinch of stool contains about...

10^{11} bacteria/g * $4\text{x}10^{6}$ bp/bacteria = $4\text{x}10^{17}$ bp/g

+ 10^{6} euks/g * $3\text{x}10^{8}$ bp/euk = $3\text{x}10^{14}$ bp/g

+ 10^{8} colonocytes/g * $3\text{x}10^{9}$ bp/cell = $3\text{x}10^{17}$ bp/g

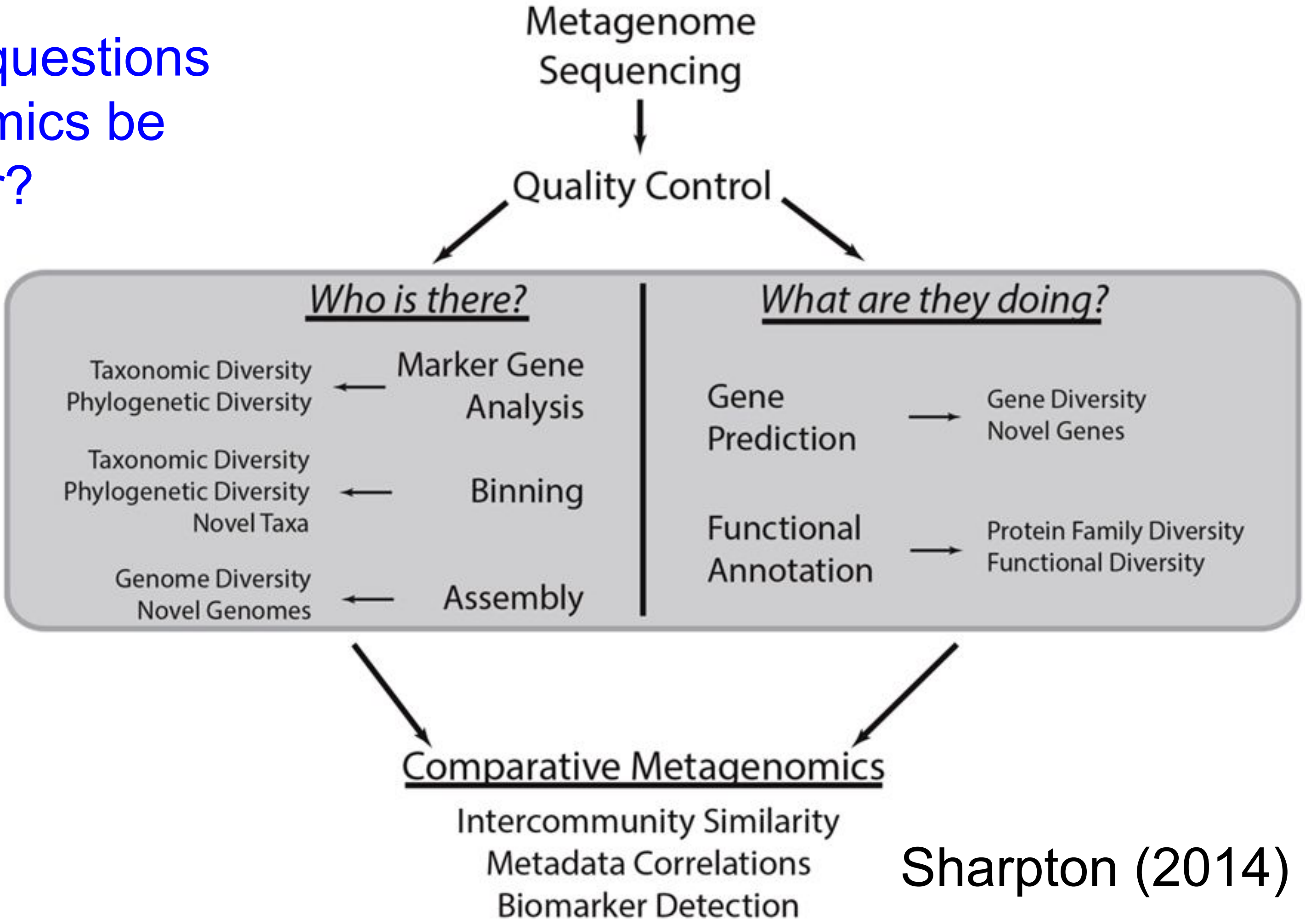
= 700 Pbp/g

= 45M MiSeq runs/g

= 100k NovaSeq runs/g

100k Genbank/g = 40 SRA / g

What kinds of questions can metagenomics be used to answer?



Metagenome Sequencing

Quality Control

Who is there?

Marker Gene Analysis → Taxonomic Diversity, Phylogenetic Diversity

Binning → Taxonomic Diversity, Phylogenetic Diversity, Novel Taxa

Assembly → Genome Diversity, Novel Genomes

What are they doing?

Gene Prediction → Gene Diversity, Novel Genes

Functional Annotation → Protein Family Diversity, Functional Diversity

Comparative Metagenomics

Intercommunity Similarity

Metadata Correlations

Biomarker Detection

Sharpton (2014)

What kinds of questions can metagenomics be used to answer?

Who is there?
(Taxonomy & Molecular Evolution)

- Is this gene present in this sample?
- How many homologs of this gene appear in this sample?
- Which genomes encode this gene?
- Is this pathogen present in this environment?
- How closely related is this uncultured strain to this cultured representative?
- How many ecotypes of this bacterium appear in this environment?
- How the h*ll many prokaryotic Phyla are there in the world??

What are they doing?
(Community Ecology & Function)

- What proteins do symbionts encode to mediate relationships with their host?
- Which genes/pathways/genomes co-occur in this environment?
- What antibiotic resistance genes does this community encode?
- How many different carbon fixation pathways exist in hydrothermal vents?
- Are there novel CRISPR-Cas systems yet to be discovered?

Is metagenomics right for me?

You might try amplicon sequencing if...

- You need to detect rare genes or species
- You're working with eukaryotes
- You have many (1000s) samples to run

You might try Quantitative PCR if...

- You only care about presence/absence
- You want to quantify how many copies of a gene/species is present in a sample

You might try isolate genomics if…

- You can isolate your organism of interest
- You're working with eukaryotes

You might try single-cell genomics if…

- Your interest is population genetics
- Your interest is in novel taxa
- Your interest is horizontal gene transfer and pangenomes

You might try meta-/transcriptomics if…

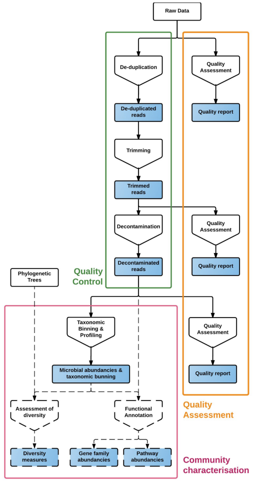
- You want levels of gene expression
- You're working with eukaryotes

Yay!

You got data!

Now what?

YAMP



ATLAS

1 **Quality Control**
- PCR duplicates removal
- Quality trimming
- Host removal
- Common contaminant removal

➢ **QC reads**

2 **Assembly**
- Error correction
- Paired-end merging
- Assembly (metaSpades/megahit)
- Post-filtering

➢ **High-quality Scaffolds**

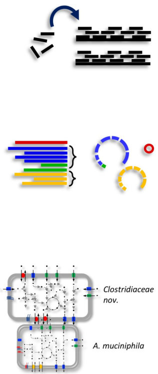
3 **Genomic Binning**
- Binning (metabat, maxbin2)
- Quality Assessment (checkM)
- Bin refining (DAS Tool)
- Dereplication (dRep)
- Quantification
- Robust taxonomic classification (CAT)

➢ **Genomes**

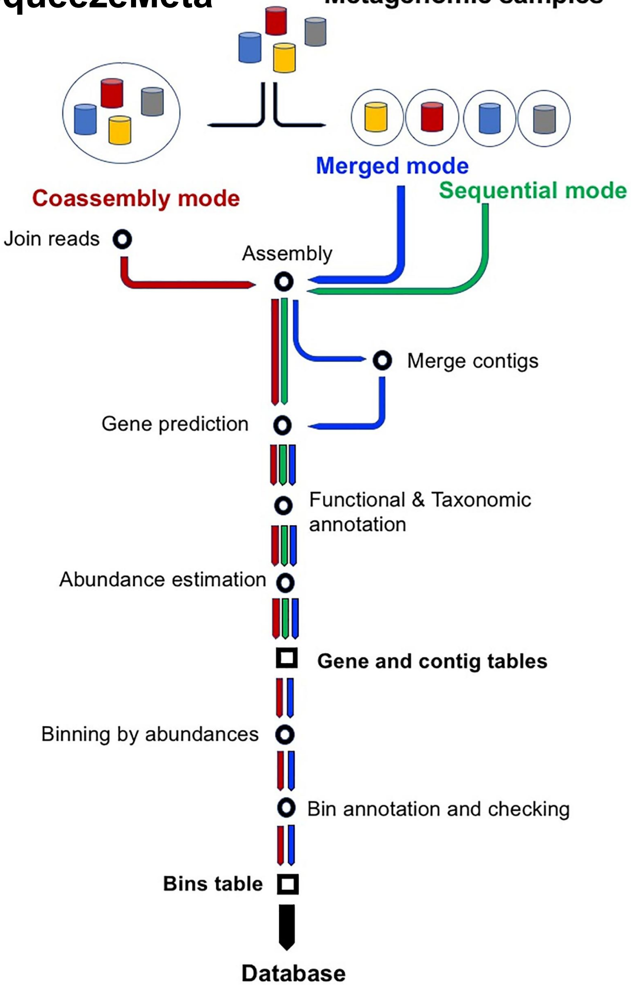
➢ **Abundances**

4 **Annotation**
- Gene prediction (prodigal)
- Cluster redundant genes (linclust/ cd-hit)
- Annotation (eggNOG)

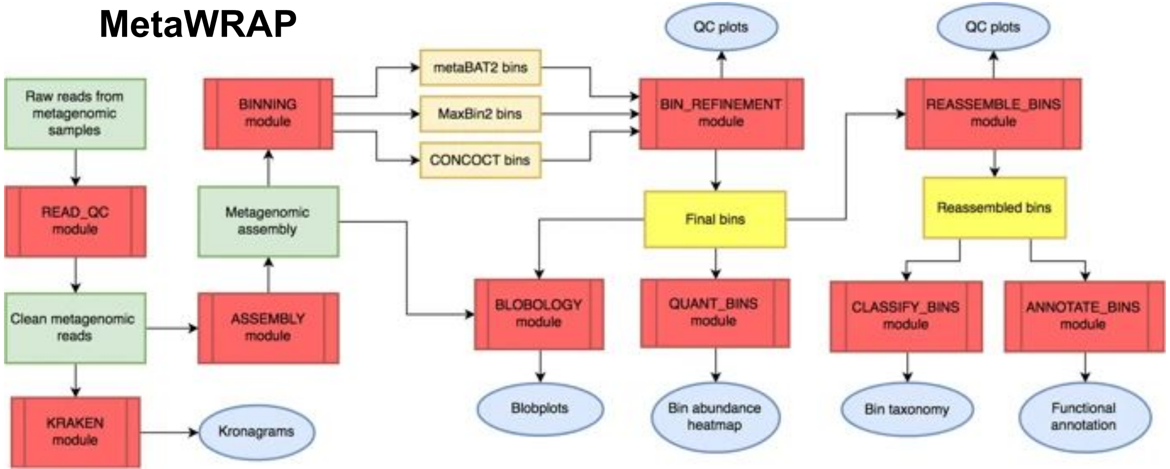
➢ **Comparable gene catalog**

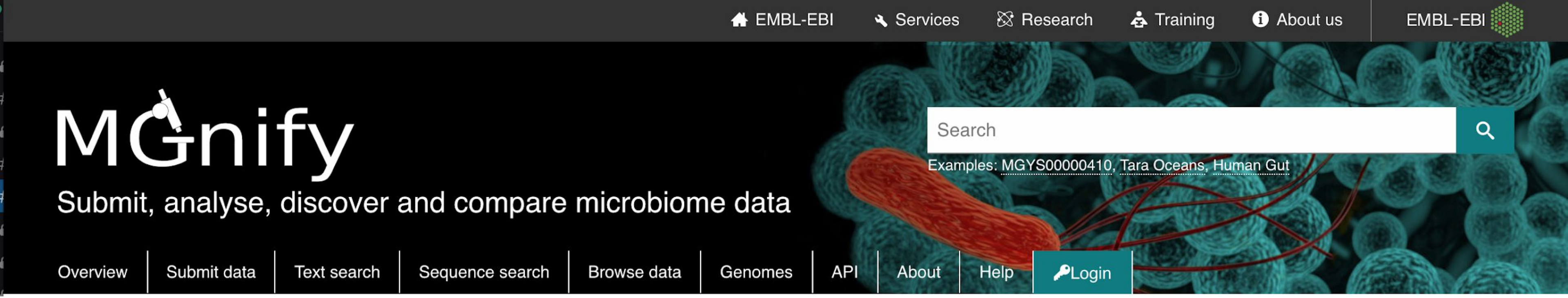


SqueezeMeta



MetaWRAP

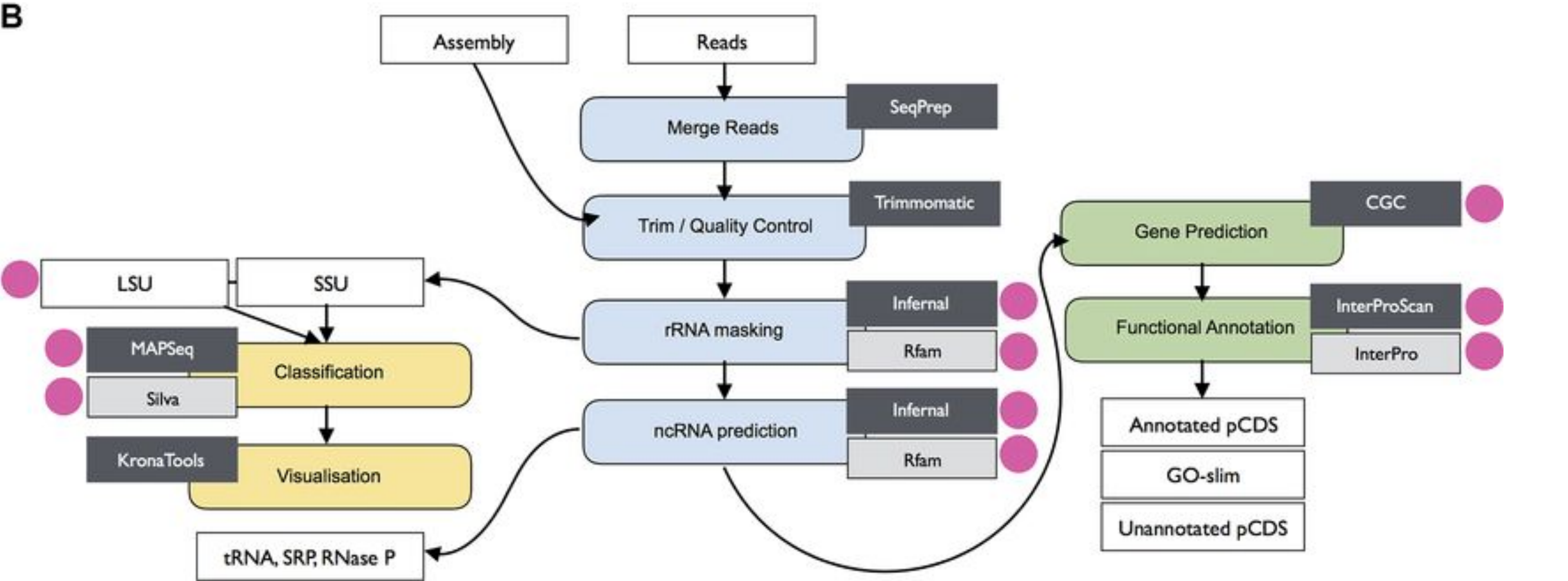




Submit, analyse, discover and compare microbiome data

Search

Examples: MGYS00000410, Tara Oceans, Human Gut



Future Metagenomics Lessons

Taxonomic Classification

Assembly

Binning

For more info go to: https://github.com/biovcnet/topic-metagenomics

Other BVCN Topics

#amplicons

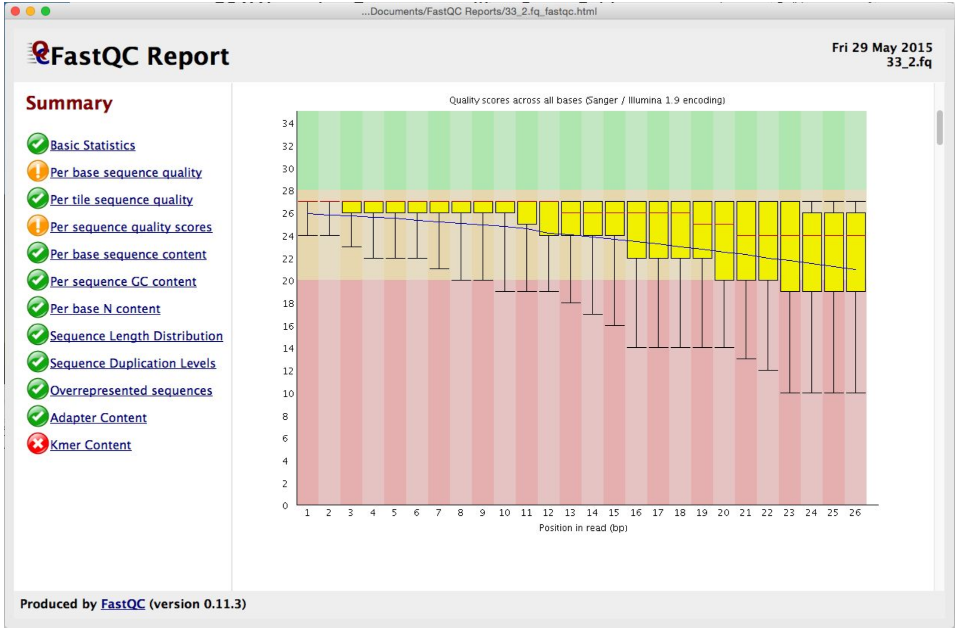
#functionalannotation

#transcriptomics

#networkscience

#population-genetics-and-comparative-genomics

Demo #1 by Alexis Marshall on Quality Control



...Documents/FastQC Reports/33_2.fq_fastqc.html

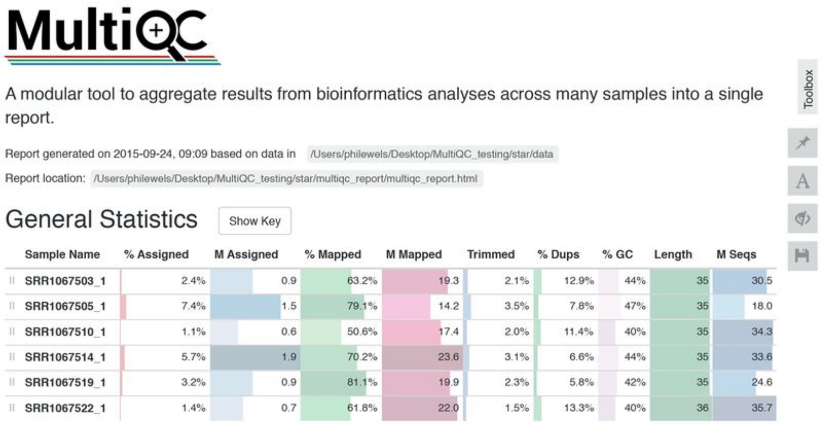
FastQC Report

Fri 29 May 2015
33_2.fq

Summary

- ✓ Basic Statistics
- ! Per base sequence quality
- ✓ Per tile sequence quality
- ! Per sequence quality scores
- ✓ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

Produced by FastQC (version 0.11.3)



MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2015-09-24, 09:09 based on data in /Users/philewels/Desktop/MultiQC_testing/star/data

Report location: /Users/philewels/Desktop/MultiQC_testing/star/multiqc_report/multiqc_report.html

General Statistics

Show Key

Sample Name	% Assigned	M Assigned	% Mapped	M Mapped	Trimmed	% Dups	% GC	Length	M Seqs
SRR1067503_1	2.4%	0.9	63.2%	19.3	2.1%	12.9%	44%	35	30.5
SRR1067505_1	7.4%	1.5	79.1%	14.2	3.5%	7.8%	47%	35	18.0
SRR1067510_1	1.1%	0.6	50.6%	17.4	2.0%	11.4%	40%	35	34.3
SRR1067514_1	5.7%	1.9	70.2%	23.6	3.1%	6.6%	44%	35	33.6
SRR1067519_1	3.2%	0.9	81.1%	19.9	2.3%	5.8%	42%	35	24.6
SRR1067522_1	1.4%	0.7	61.8%	22.0	1.5%	13.3%	40%	36	35.7

https://www.youtube.com/watch?v=7jRTyfdIXLo