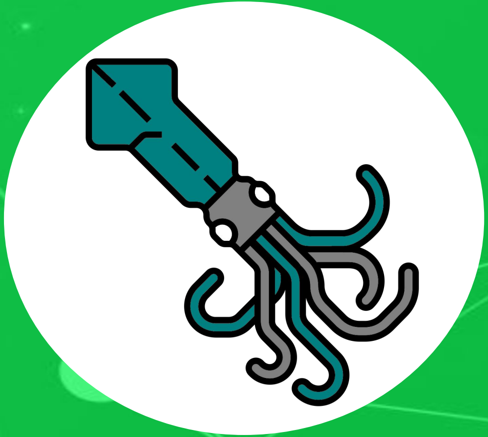
Centre for Genomic Research

UKRI Natural Environment Research Council

NEOF NERC ENVIRONMENTAL OMICS FACILITY

Intro to Shotgun Metagenomics



UNIVERSITY OF LIVERPOOL

The University Of Sheffield.

UNIVERSITY OF LIVERPOOL

UKRI Natural Environment Research Council

Centre for Genomic Research

NEOF NERC ENVIRONMENTAL OMICS FACILITY

The University Of Sheffield.

Websites
NEOF: https://neof.org.uk/
NERC: https://nerc.ukri.org/
CGR:
https://www.liverpool.ac.uk/genomic-research/

Twitter
NEOF: @NERC_EOF
NERC: @NERCscience
CGR: @CGR_UoL

Format & Schedule

This intro

Bookdown

Theory

Practice

Exercises

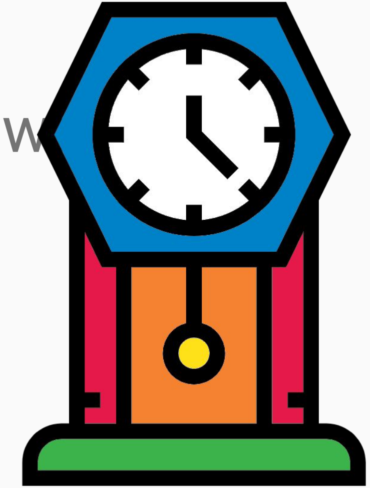
MCQs

Work at your own pace

We are here to help

Time with breaks in betw

- 10:00-11:15
- 11:30-12:30
- 13:30-14:45
- 15:00-16:00



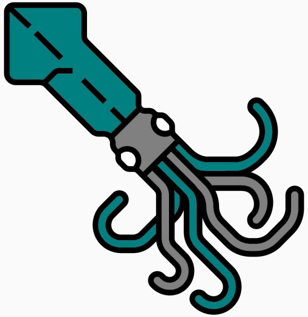
Days

Day 1

QC

Read approach

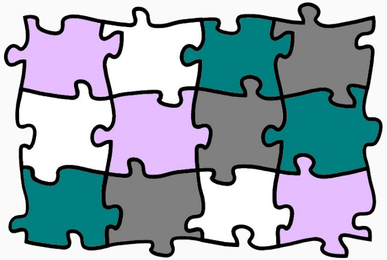
- Kraken2 & Bracken
- HUMAnN



Day 2

Assembly Approach

- Megahit
- MetaBAT2



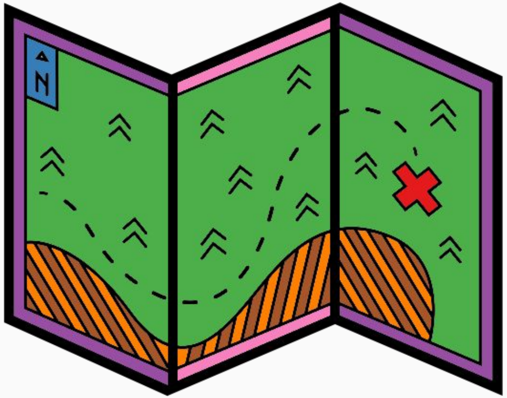
Plan

What is shotgun Metagenomics?

Examples

Read approach

- QC
- Taxonomy classification



Shotgun Metagenomics

All DNA in a sample

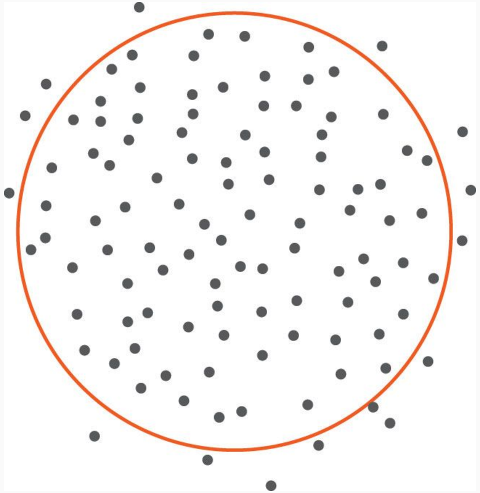
All organisms present

DNA broken up randomly

Numerous small pieces

Sequence all the sheared bits of DNA

Named after quasi-random firing pattern of shotgun



Benefits

Taxonomy

Genes sequenced

Random

Most low coverage organisms will be represented

Culture free

Captures unculturable organisms

Drawbacks

Expensive/less samples

Complex analysis

Difficult to elucidate abundance of organisms

Classification only as good as databases

Salmon and eDNA

Environmental DNA

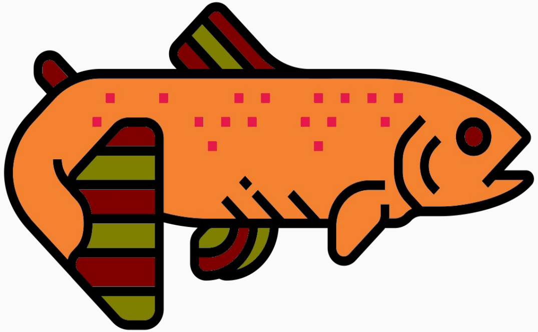
Track presence of species

Trace amounts of DNA (scales, excretions etc.)

Track migration of Salmon upstream

Successfully time migration times

Week after migration saw more. Why?



Adaptation to pollutants by Crustaceans

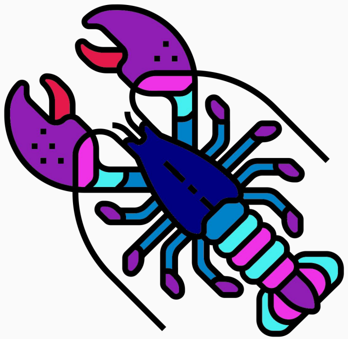
eDNA capture of various clean and polluted waters

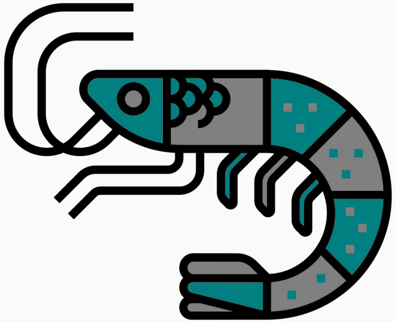
Detect presence of different crustaceans

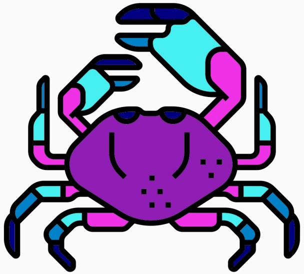
Detect presence of food of crustaceans

What genes are present in polluted waters?

What genes are present in clean waters?







Dampwood vs Drywood termites

Dampwood Termites infest wood with high moisture

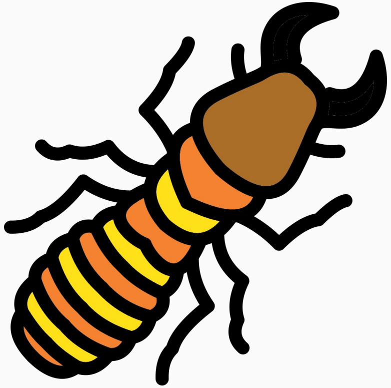
Drywood Termites can infest wood with low moisture. Can live in wood structures.

Termite genes for low water?

Gut microbiome allows digestion of wood.

Different organisms?

Any metabolic processes of bacteria allow for low water?



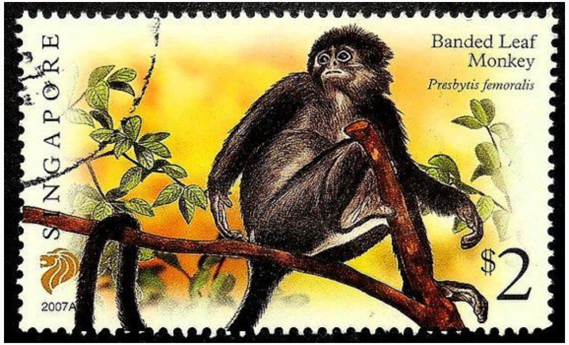
Fecal metagenomics

Banded leaf monkey (Presbytis femoralis)

Endangered

Diet of 53 plants, 33 families

Broadly consistent diet (>2 years observation)



Sequenced P. femoralis mitochondria for phylogeny

Presence of gut parasites confirmed.

Fecal metagenomics for the simultaneous assessment of diet, parasites, and population genetics of an understudied primate: Srivathsan et al 2016

Read approach

1. Raw data
2. Quality control
3. Host removal
4. Taxonomic profiling
5. Functional Profiling

Raw data

Introduction to data

Linking the data

Quality check

- FastQC
- MultiQC

Quality control

- Quality trimming
  - Trim Galore!
- Check quality trimmed data
  - FastQC
  - MultiQC

Host removal

Reference genome assembly

Bowtie2

Index reference

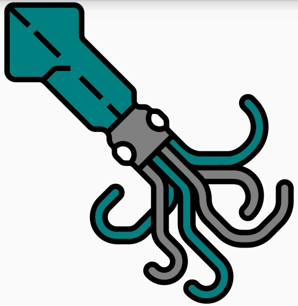
Align reads

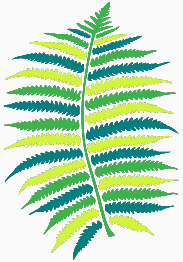
Extract unmapped reads

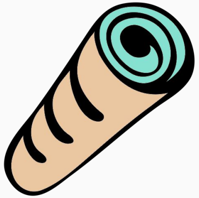
Re-pair reads

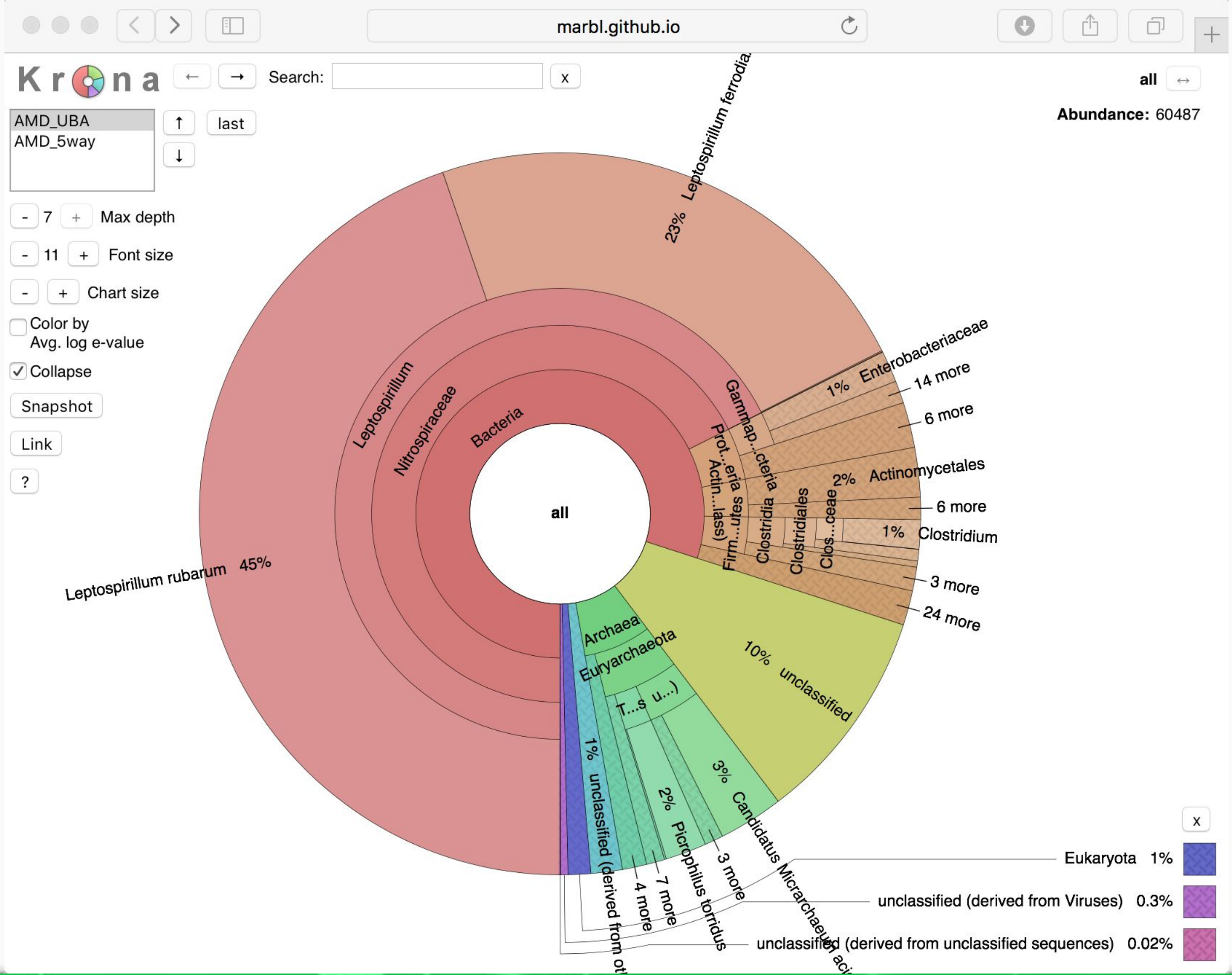
Taxonomic Profiling

- Kraken2
  - Taxonomic classification
- Krona
  - Visualisation of taxonomy
- Bracken
  - Abundance estimation of classified taxonomy
- LefSe
  - Biomarker detection









marbl.github.io

Search:

AMD_UBA
AMD_5way

Max depth: 7

Font size: 11

☐ Color by Avg. log e-value

☑ Collapse

all

Abundance: 60487

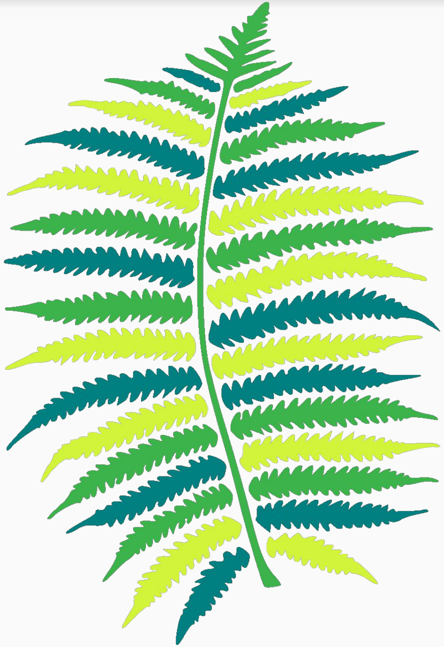
Bracken

Kraken2 detects presence

Bracken utilises Kraken2 output

Bayesian Reestimation of Abundance with KrakEN

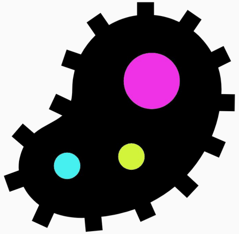
Statistical method to compute abundance of species



Bracken and genome length

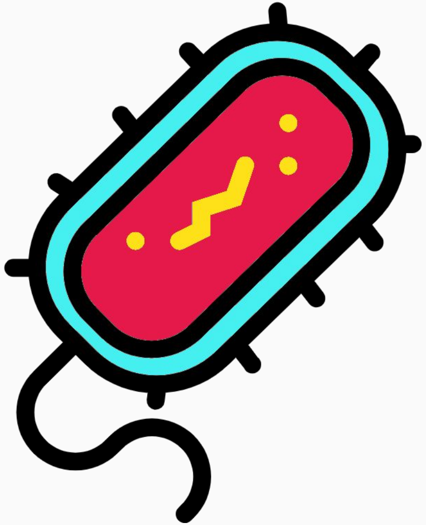
Genome size: 2Mb

Classified reads: 200,000 (100bp*2)



Genome size: 4Mb

Classified reads: 200,000 (100bp*2)



Same level of classification after Kraken2

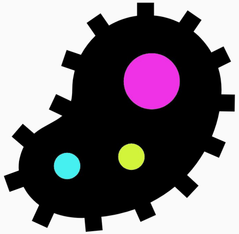
Bracken and genome length

Genome size: 2Mb

200,000 (100bp*2)

40Mb / 2Mb

Abundance reestimation = 20

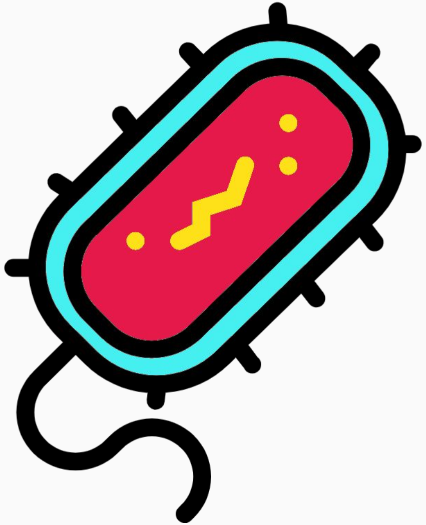


Genome size: 4Mb

200,000 (100bp*2)

40Mb / 4Mb

Abundance reestimation = 10



Double the amount of smaller organism

LEfSe

LEfSe (Linear discriminant analysis (LDA) Effect Size)

Method to detect organisms that are differentially abundant between sample groups

For example

Disease causing organisms

Survival of organisms

Reminders and Tips

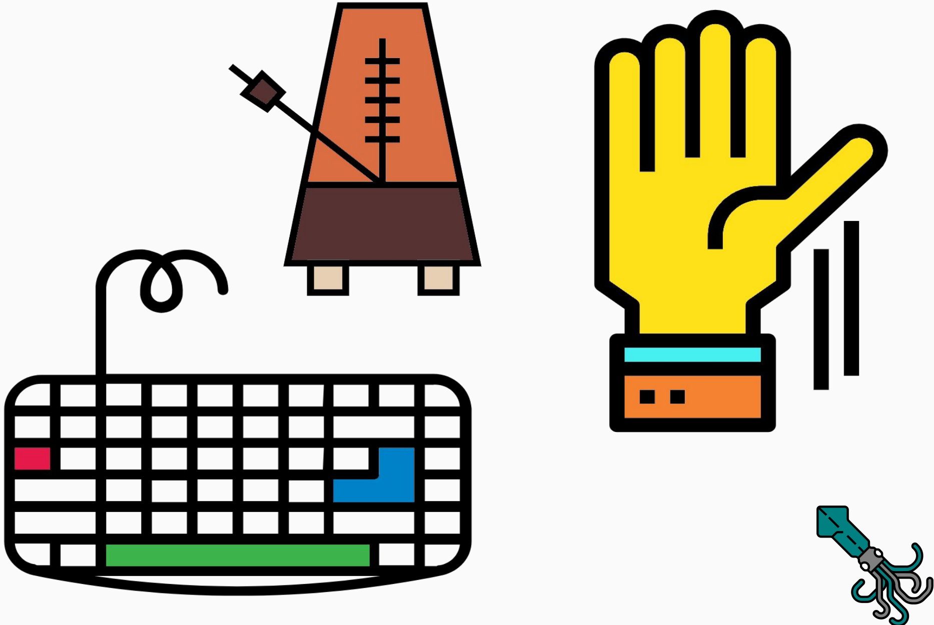
Work at your own pace

Typos

Ask questions

Breaks are important

Tab, space, and enter



During sessions

Zoom - Ask via microphone if no question currently being asked/answered

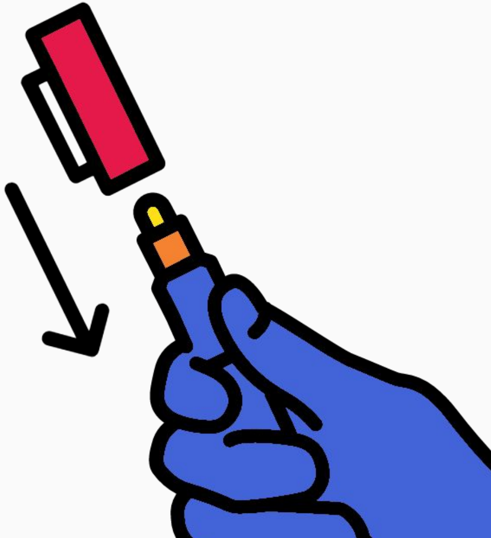
Slack - Ask questions via the channel or ask to go into a zoom breakout room with one of us

WebVNC - We can connect to your webVNC to see and help with issues.

Breakout rooms upon request

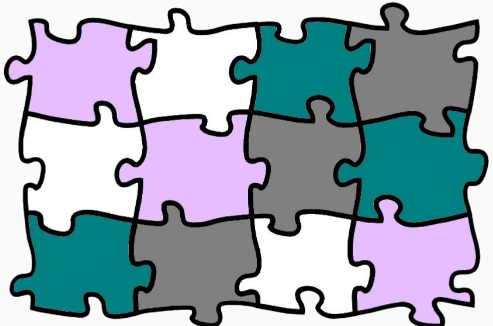
Recap

- What is Shotgun Metagenomics
  - Taxonomy & Functional data
- Examples
  - Salmon, Crustaceans, Termite, and Monkeys
- Read approach
  - Raw data, Trimming, and Taxonomy



Thursday's presentation

- 10am
- Assembly approach
  - Stitch reads, assembly, binning + annotation
- Read vs Assembly approach
- Eukaryotic analysis
- Short read vs long read



Thank you!

Questions?