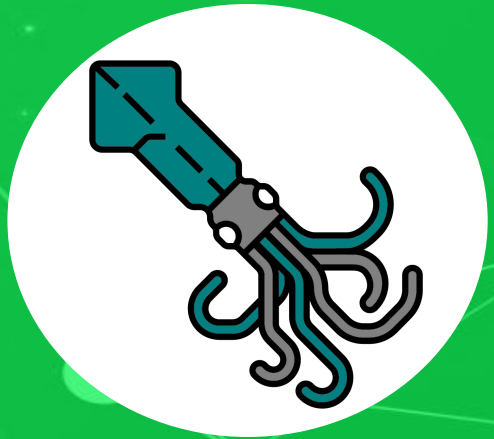
Centre for Genomic Research

UKRI Natural Environment Research Council

NEOF NERC ENVIRONMENTAL OMICS FACILITY

K-mers & Kraken2



UNIVERSITY OF LIVERPOOL

The University Of Sheffield.

Plan

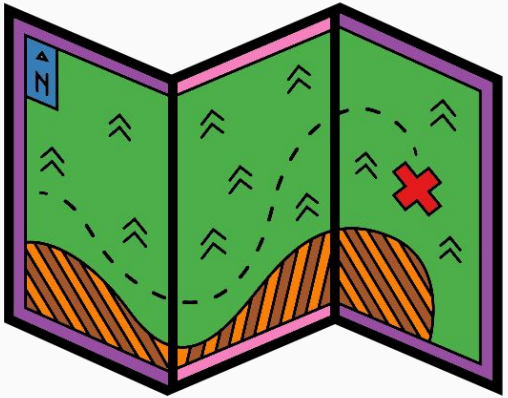
K-mers

Kraken2

LCA

Consensus taxonomy

Optimal K-mer size?



K-mers

K-mer: a sequence of set length

K is the length

Extract all possible k-mers from a sequence

E.g 3-mers

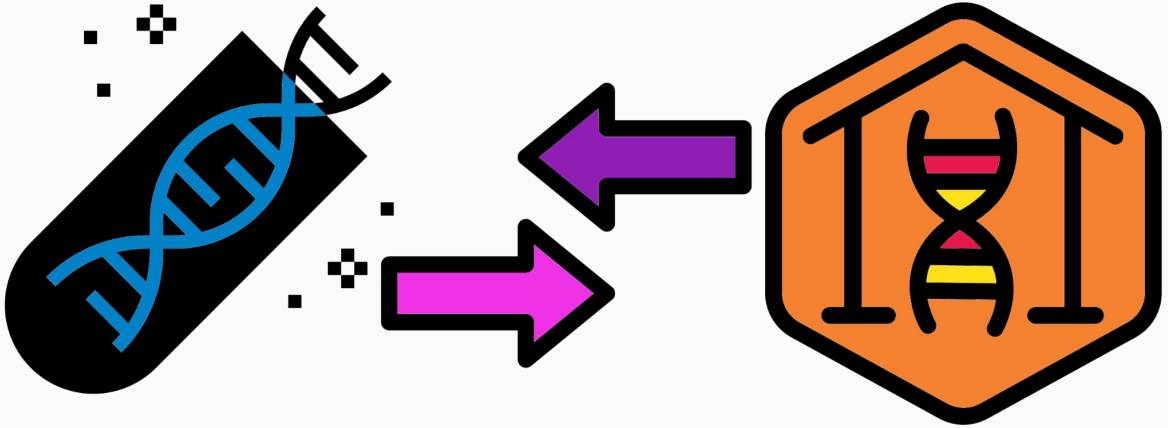
K-mer examples

Sequence	AAGTGCGT
3-mers	AAG | AGT | GTG | TGC | GCG | CGT
4-mers	AAGT | AGTG | GTGC | TGCG | GCGT
5-mers	AAGTG | AGTGC | GTGCG | TGCGT
6-mers	AAGTGC | AGTGCG | GTGCGT
7-mers	AAGTGCG | AGTGCGT
8-mers	AAGTGCGT

Kraken2

Compares

- seq k-mers
- k-mers mapped to a genomic library

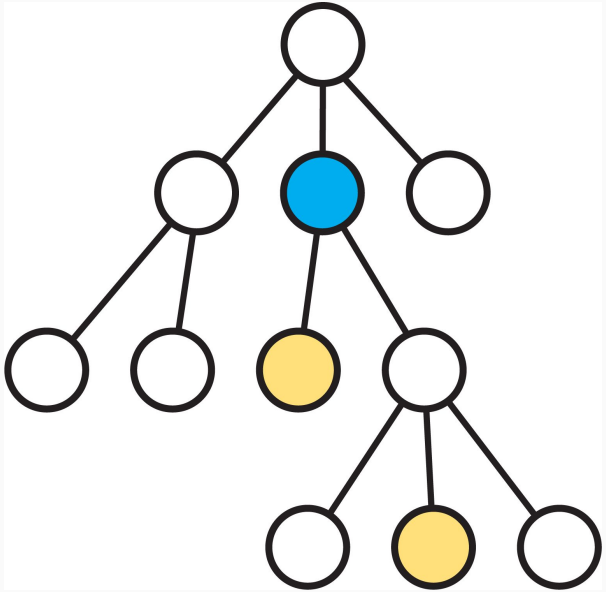


LCA classification

Genomic library

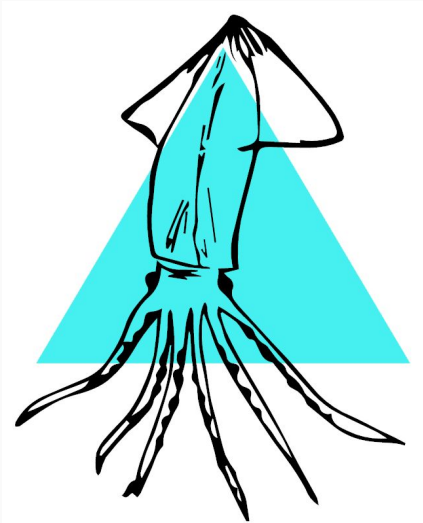
K-mer’s taxonomy

LCA of all the organisms that contain the k-mer

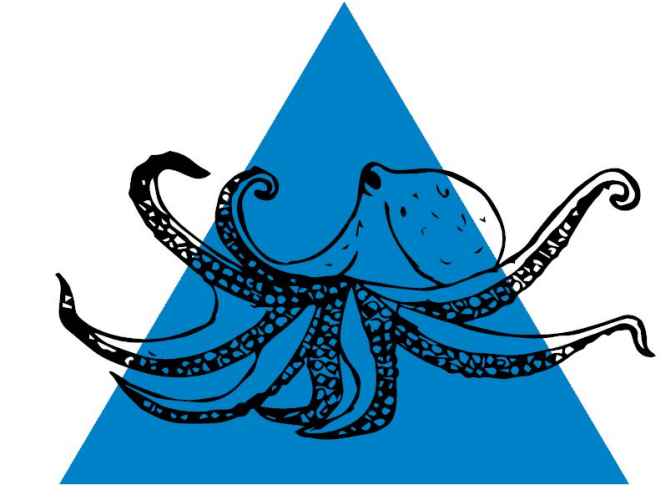


LCA classification

Kingdom: Animalia | Phylum: Mollusca
Class: Cephalopoda | Order: Sepiida



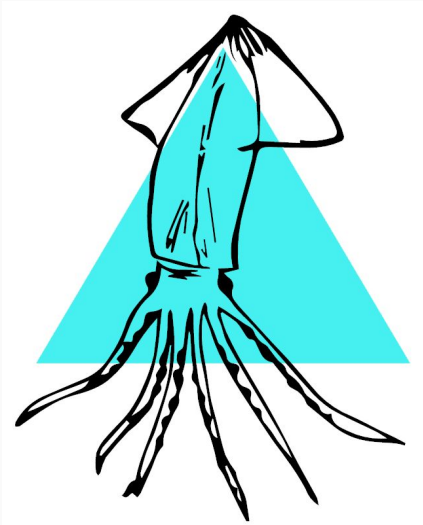
Kingdom: Animalia | Phylum: Mollusca
Class: Cephalopoda | Order: Octopoda



Kingdom: Animalia | Phylum: Mollusca | Class: Cephalopoda

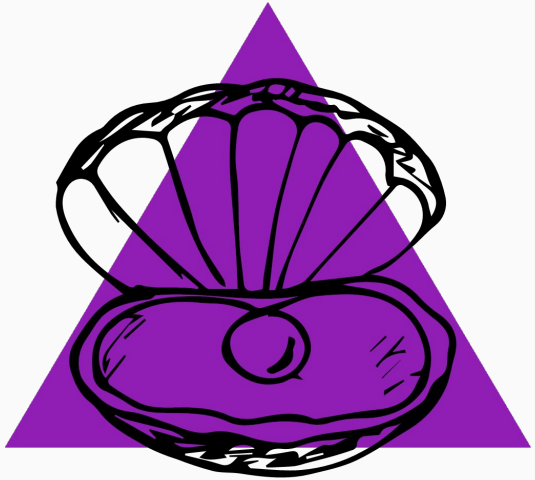
LCA classification

Kingdom: Animalia | Phylum: Mollusca
Class: Cephalopoda | Order: Sepiida



+

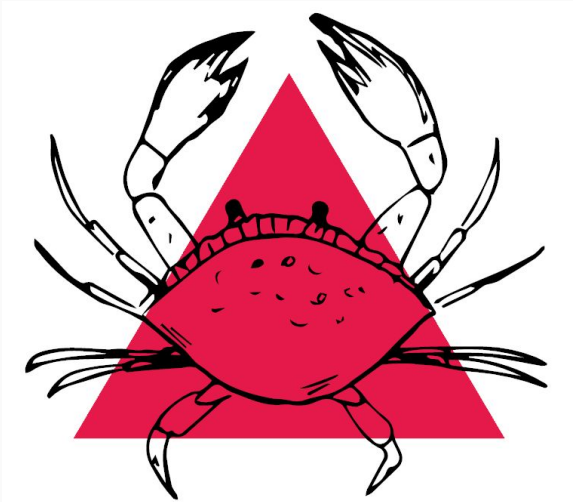
Kingdom: Animalia | Phylum: Mollusca
Class: Bivalva | Order: Venerida



Kingdom: Animalia | Phylum: Mollusca

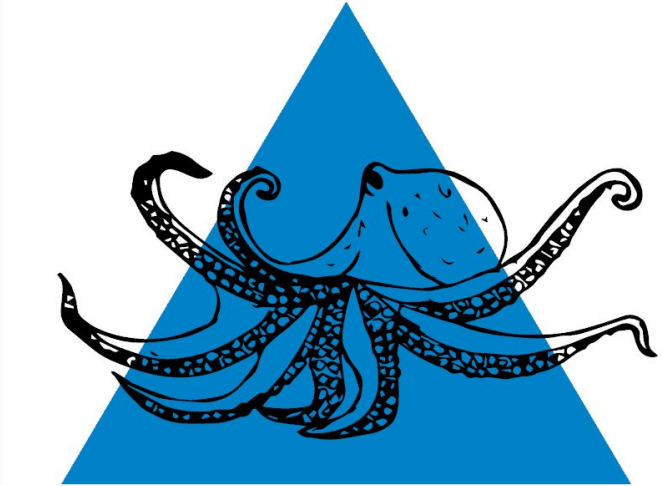
LCA classification

Kingdom: Animalia | Phylum: Arthropoda
Class: Malacostraca | Order: Decapoda



+

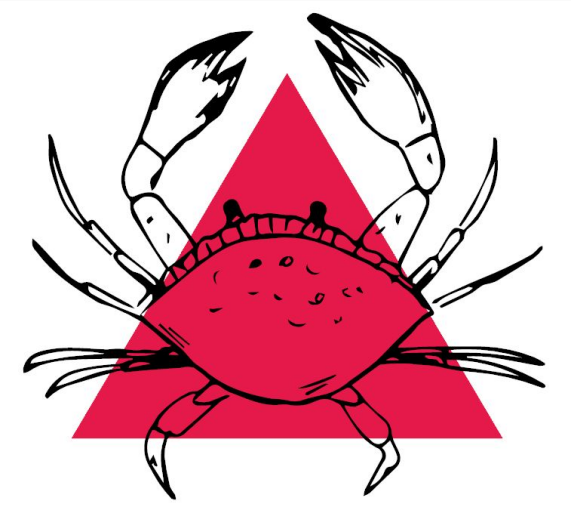
Kingdom: Animalia | Phylum: Mollusca
Class: Cephalopoda | Order: Octopoda



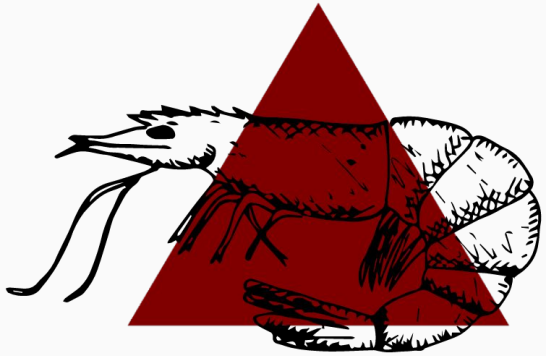
Kingdom: Animalia

LCA classification

Kingdom: Animalia | Phylum: Arthropoda
Class: Malacostraca | Order: Decapoda
Infraorder: Brachyura



Kingdom: Animalia | Phylum: Arthropoda
Class: Malacostraca | Order:Decapoda
Infraorder: Caridea

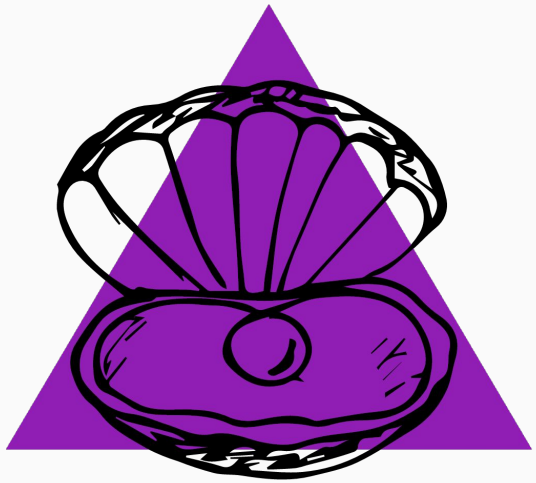


Kingdom: Animalia | Phylum: Arthropoda
Class: Malacostraca | Order: Decapoda

LCA classification

Domain: Eukaryota | Kingdom: Animalia

Domain: Eukaryota | Kingdom: Plantae



Domain: Eukaryota

Genomic library

k-mer	LCA Classification
AAA	Kingdom: Animalia | Phylum: Mollusca
AAC	Kingdom: Animalia | Phylum: Mollusca | Class: Cephalopoda
AAG	Kingdom: Animalia
AAT	Kingdom: Animalia | Phylum: Mollusca
ACA	Domain: Eukaryota
ACC	Kingdom: Animalia | Phylum: Mollusca | Class: Cephalopoda
ACG	Domain: Eukaryota
ACT	Kingdom: Animalia

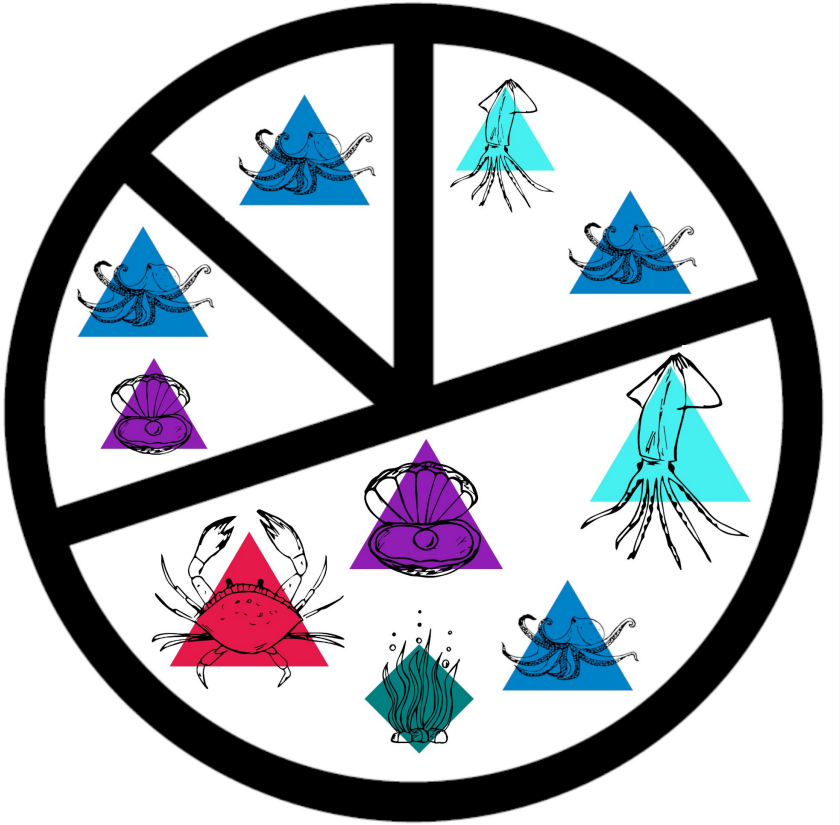
Sequence k-mers

LCA classification in sequence

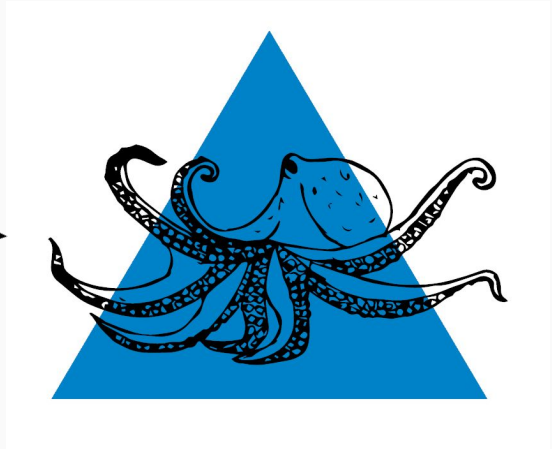
1st 3-mer	AAC	Class: Cephalopoda
2nd 3-mer	ACG	Domain: Eukaryota
3rd 3-mer	CGA	Phylum: Mollusca
4th 3-mer	GAT	Kingdom: Animalia
5th 3-mer	ATT	Order: Octopoda
6th 3-mer	TTA	Order: Octopoda
7th 3-mer	TAC	Phylum: Mollusca
8th 3-mer	ACC	Class: Cephalopoda
9th 3-mer	CCT	Kingdom: Animalia

Sequence classification from k-mer classifications

Percentage of k-mer classifications



Consensus taxonomy classification of read

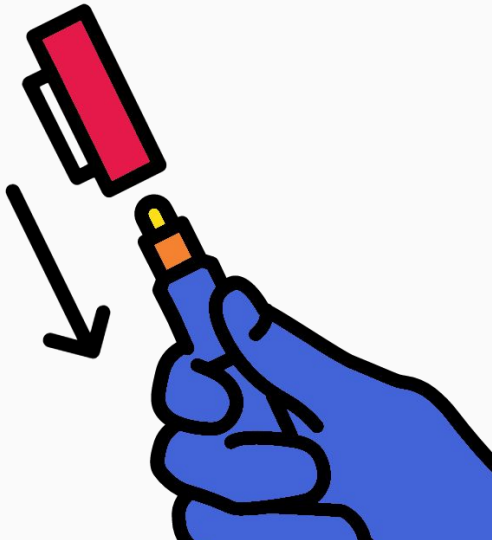


Optimal k-mer size?

- Shorter than Illumina read length (150-300bp)
- Too short and it won't be specific enough
  - There are only 64 unique 3-mers (AAA, AAC, AAG ……. TTT)
  - Therefore only have 64 different taxonomies in genomic library
- Too long and it will be too specific
  - May get few classifications due to differences between individual
  - Finding large k-mers is computational expensive
  - Incredibly large database: >6e+60 unique 101-mers
- Kraken2 uses 35-mers by default
  - > 1.18e+21 unique 35-mers i.e. > 1.18 sextillion
  - 1 sextillion = one thousand million million million

Recap

- K-mers
- Kraken2
- LCA
- Consensus taxonomy
- Optimal K-mer size?



Thank you!

Questions?