12. Feb. 21

Metagenomics

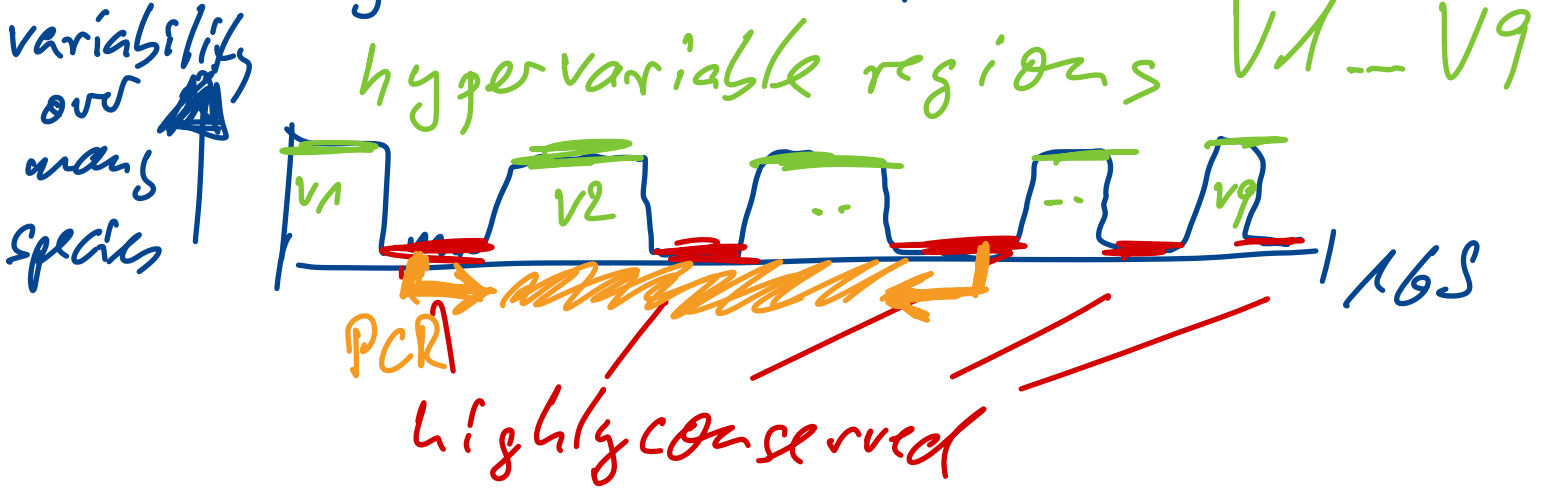
Goal: learn about a microbial community

- taxonomical classification
  (tax. ranks: superkingdom, phylum, class, order, family, genus, species)
- functional classification
- (unknown) individual members of the community
- metagenomic assembly

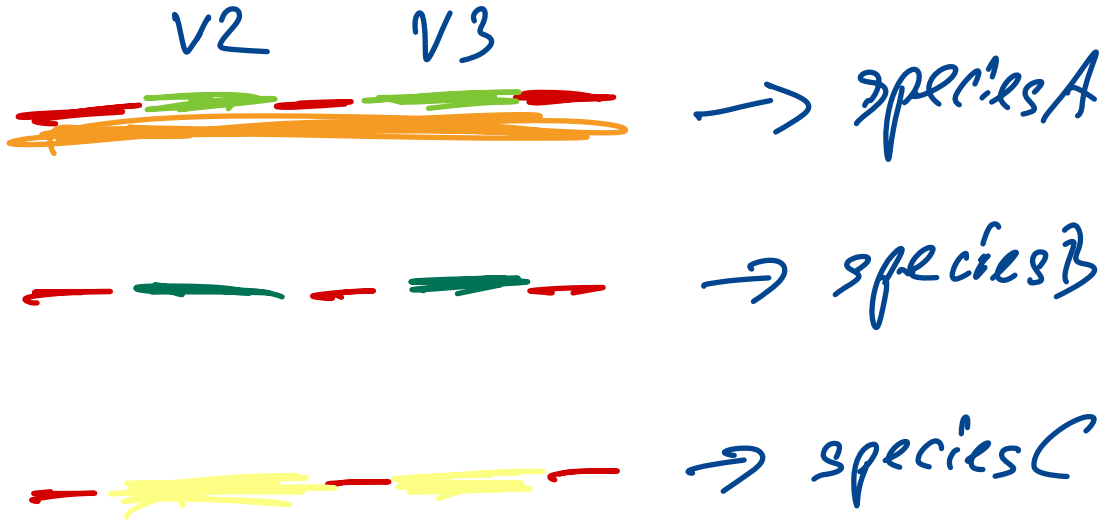
I. Marker-based methods (taxonomical classification)

Idea: use marker genes that appear in many species

e.g. 16S rRNA



↳ amplicon-based metagenomics



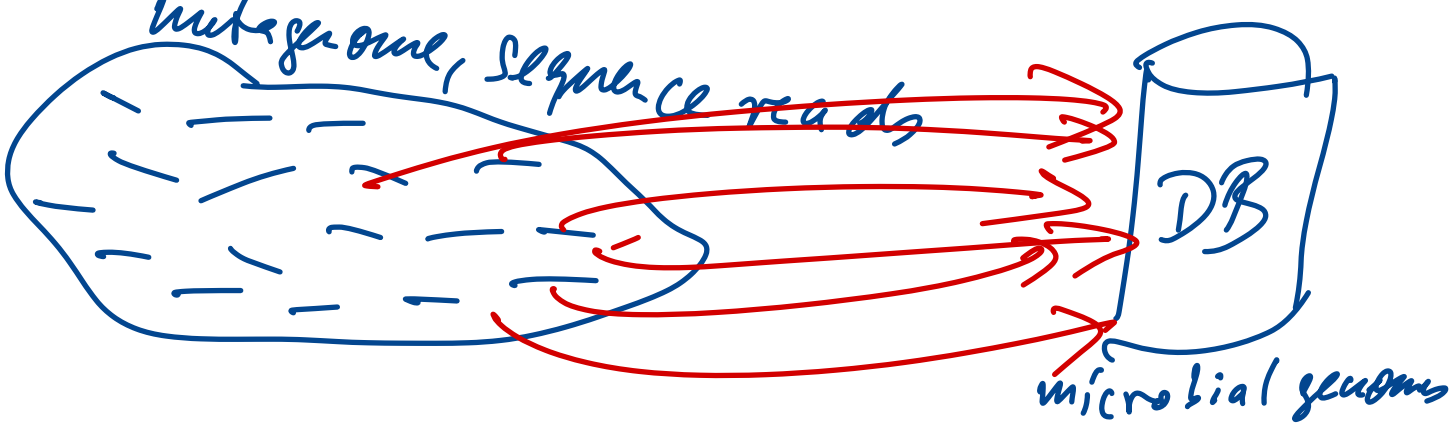
II. Whole-metagenome sequencing

1. Composition-based Methods

e.g. sequence features like GC-content or k-mer frequencies

2. Comparison-based Methods

- rely on homology information via database searches



Bioinformatics challenges: high coverage

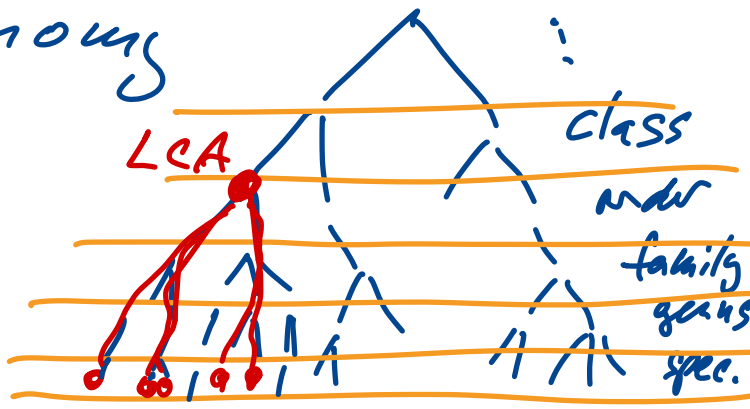
- huge datasets (deep sequencing)
  → 100s M reads is typical per sample
  → often multiple samples (timepoints, other conditions)
- algorithms need to be efficient

Read to database mapping in metagenomics

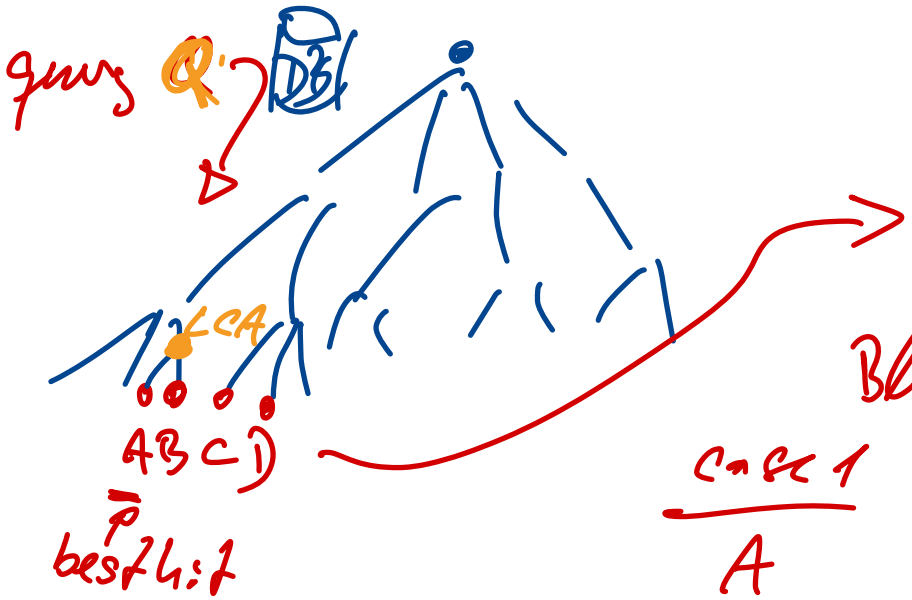
- **MG-RAST** (2008...)
  - for each read, collect best BLAST hit
  - simplistic, therefore classify up to rank family (only if it's a good hit)

- **MEGAN** (2007...)
  - run BLAST for each read
  - top 10 hits, as long as they have a bit-score ≥ 35
  - LCA in taxonomy

lowest common ancestor



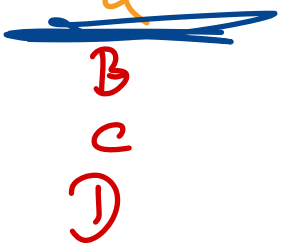
- **SORT-ITEMS** (2009...)
  - starts like MEGAN
  - add a reciprocal search



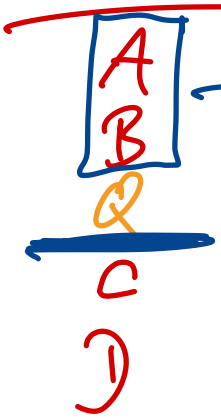
DB$_{new}$ = {Q, A, B, C, D}

Blast A against DB$_{new}$

case 1 → select just A (species)



case 2

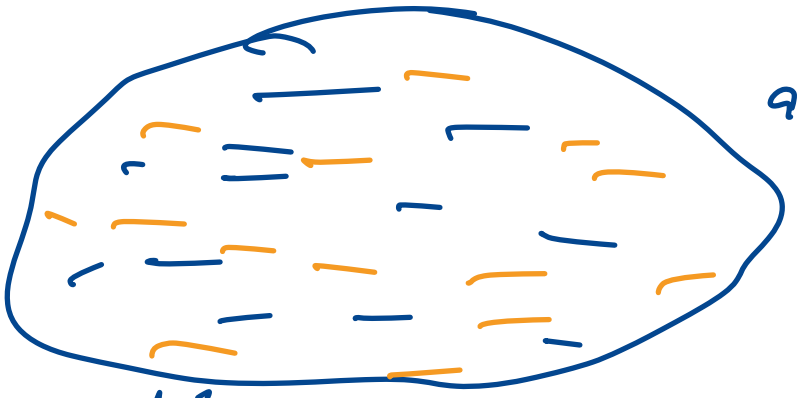


→ lca (no species prediction)

- further development: **CARMA3**

III. Metagenomic Assembly

1. Ideally: run a standard assembler
   ↪ each contig is a genome



assembly ⇒

"MAG"

metagenome assembled genome

problems:

- metagenomes are complex
- coverage may still be too low, especially for low-abundant genomes
- genomes are not "pure", there may be many variants of some species

Tricks:

- binning of reads before assembly

  hope: all reads in one bin belong to the same MAG.

- enhance binning by single-cell sequencing and read recruitment