Monday Meeting presentation

05/12/2022

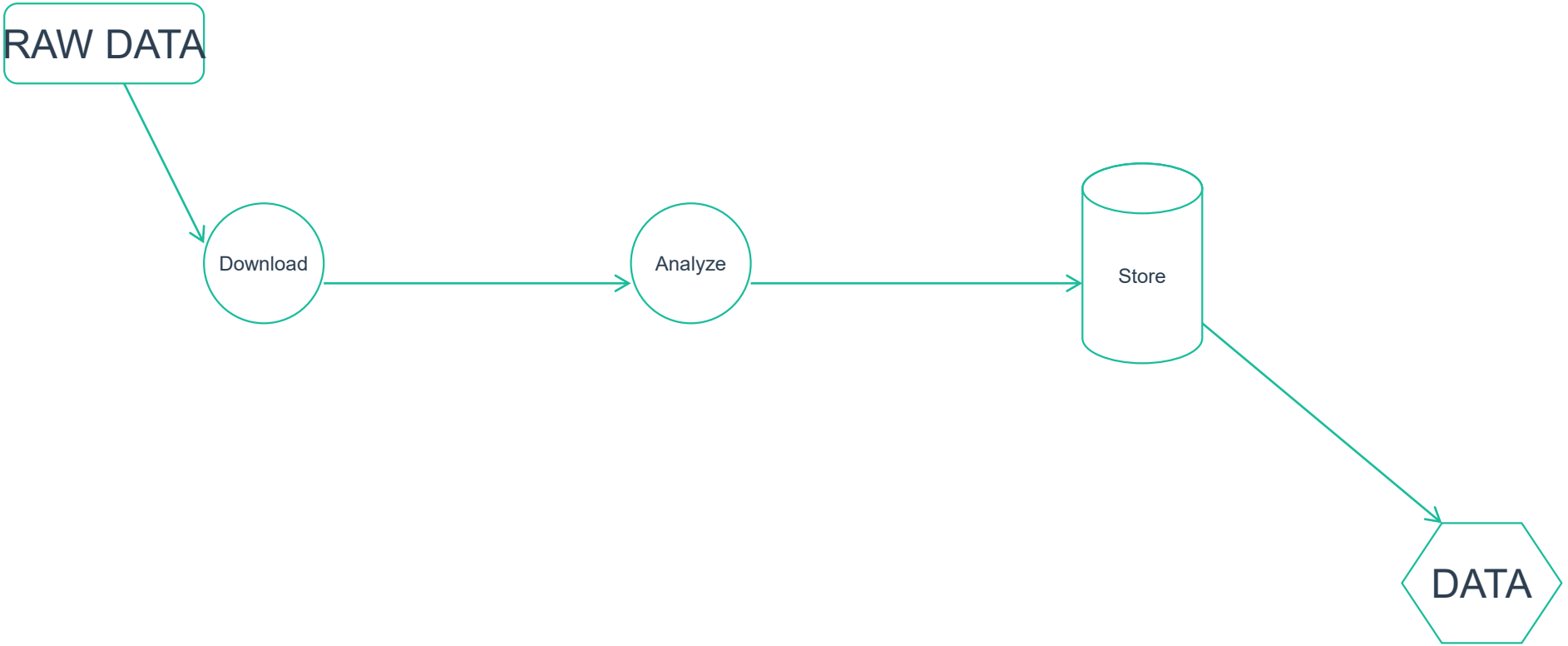
Nikiforos Pyrounakis

Research assistant

Downloading all the metagenomic datasets

- Lets analyze all the available metagenomic data once more
- Previous years 214,095 metagenomic datasets were analyzed
- End of November: 359,603 metagenomic datasets

Dummy pipeline



Before pressing the button

.Before we start downloading Petabyte of data:

–Is there room for improvement ?

–Can new tools realistically lead to better results?

. Targeted + functional software

1)Better results

2)Less computational time

✓Less money spent

So far...

.Taxonomy profiling

.Seed based metagenomic assembly

Tools we tested

.Metaphlan 4

.Kaiju

.Metacherchant

Metaphlan 4

.Profiling the composition of microbial communities (Bacteria, Archaea, Eukaryotes)

.~5.1M unique clade-specific marker genes identified from ~1M microbial genomes spanning 26,970 species-level genomes bins

Kaiju

- Sensitive taxonomic classification of high-throughput sequencing reads
- NCBI taxonomy + A reference database containing microbial and viral protein sequences
- Available genomes from NCBI RefSeq or the microbial subset of the NCBI BLAST non-redundant protein db, *nr*

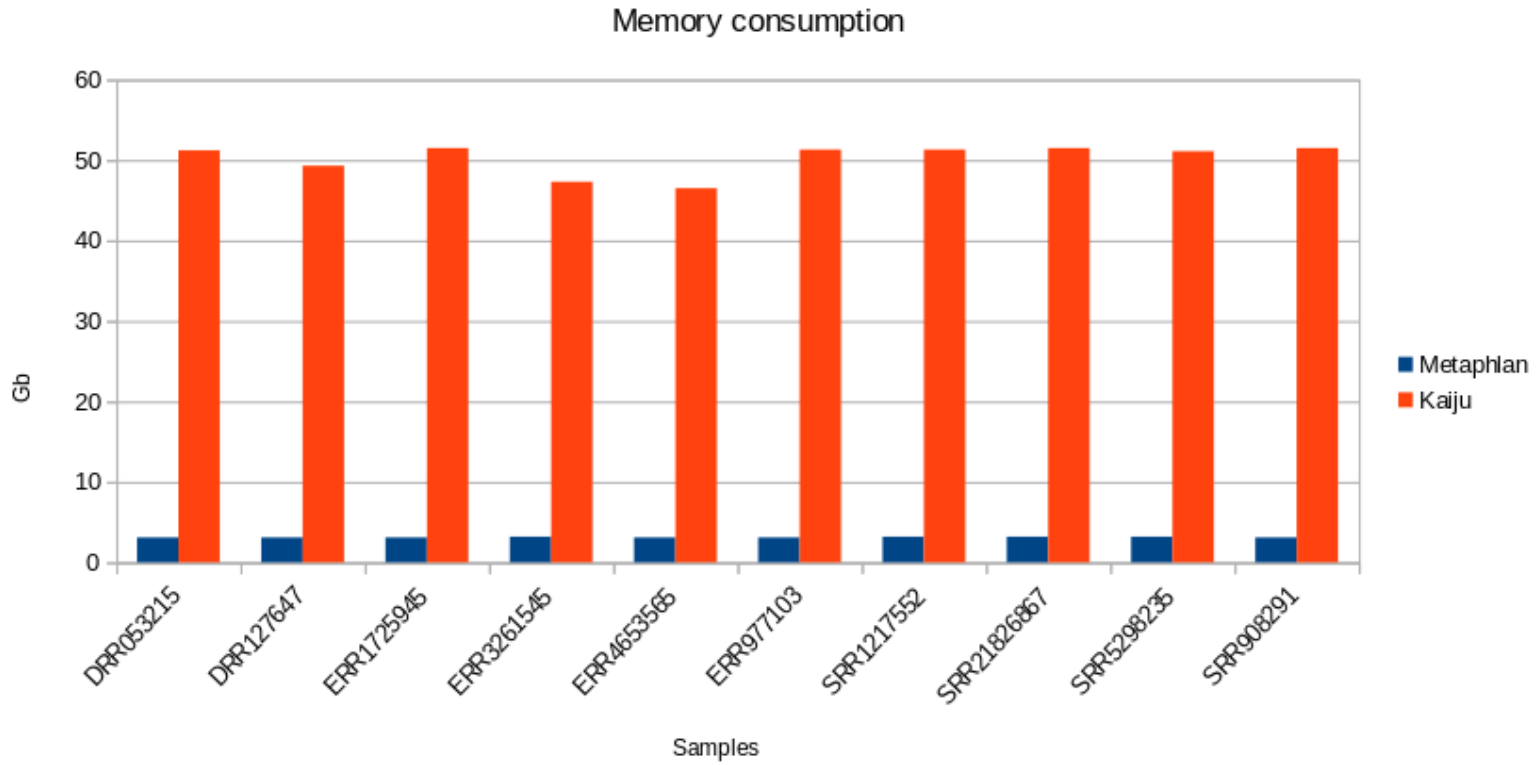
Testing grounds

.Computerome --> Thin node – 40 cores + 185 gb

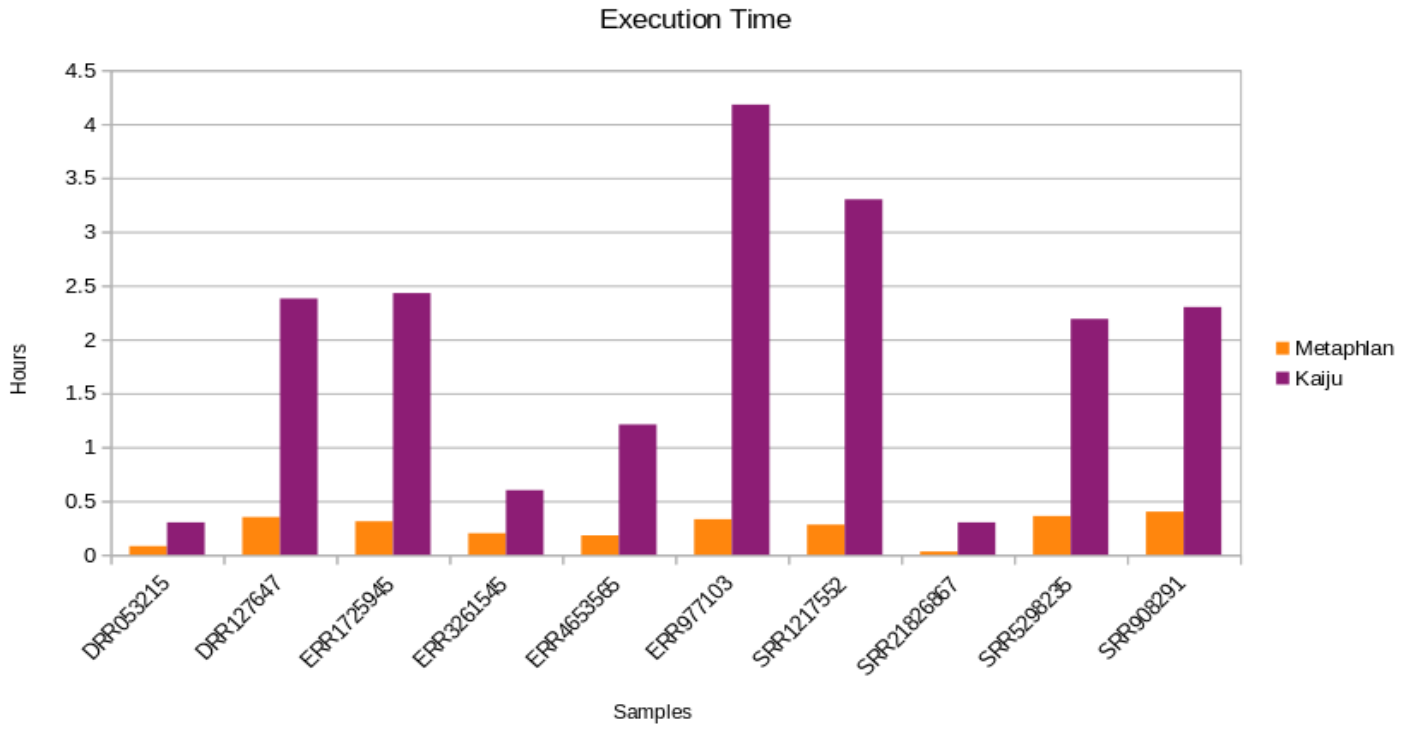
.10 metagenomic samples of various sizes(Pig, Chicken, Sewage, Nasal, Human feces, Soil, Fresh water, Marine water,...)

.*usr/*bin/time + Snakemake built in benchmarking feature

Metaphlan 4 vs Kaiju - Memory



Metaphlan 4 vs Kaiju - Time



Metaphlan 4 vs Kaiju

.Cases where both tools reported same species as the most abundant: 4

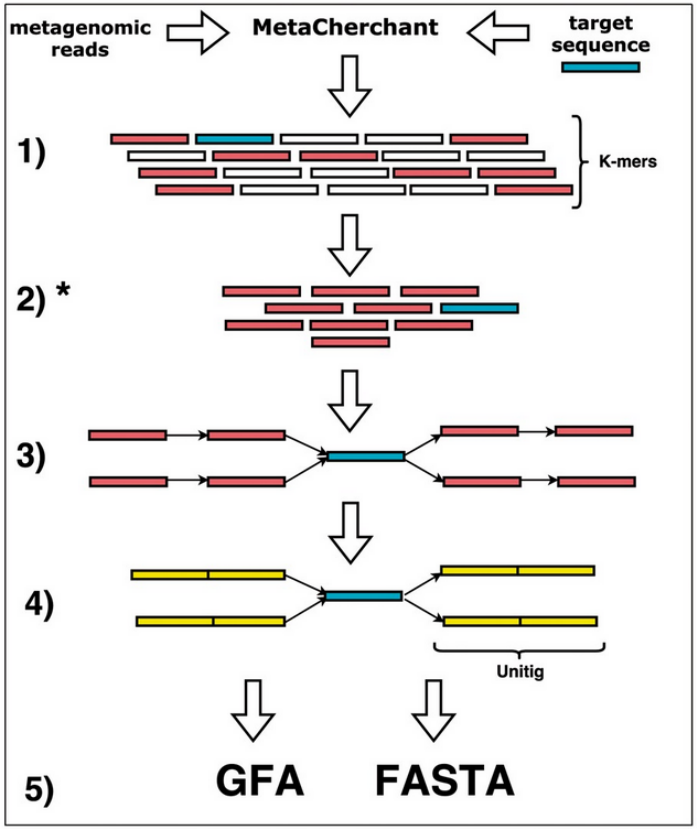
.Case where tools reported different species as most abundant: 2

.Cases where Kaiju did not report species: 0

.Cases where Metaphlan did not report species: 4

Metacherchant

- Algorithm for extracting the genomic environment of antibiotic resistance genes
- Performs sensitive taxonomic classification of sequencing reads from metagenomes



Metacherchant

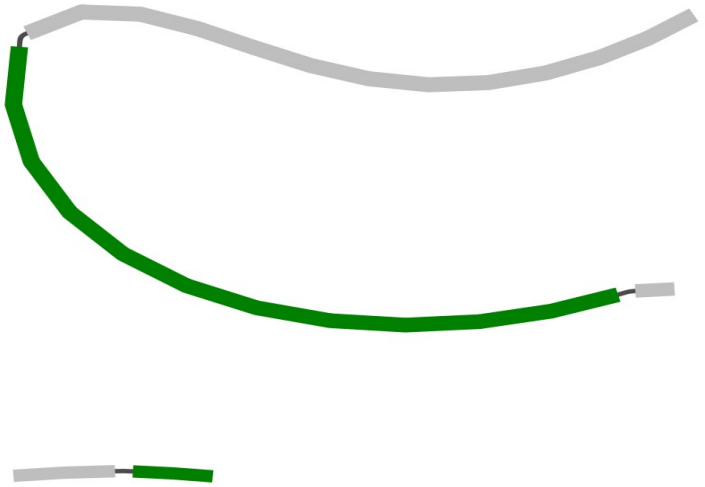
- Input raw reads : ERR3261545
  - Nasal.

- Seed: *blaTEM* gene (~860 nucl.)
  - One of the most commonly encountered β-lactamase with more than 150 variants.

- Kmer size : 30

Metacherchant results

- Fast
- Lightweight
- Visualization options with Bandage



Bandage output. Green nodes correspond to the seed sequence. Gray nodes is the r

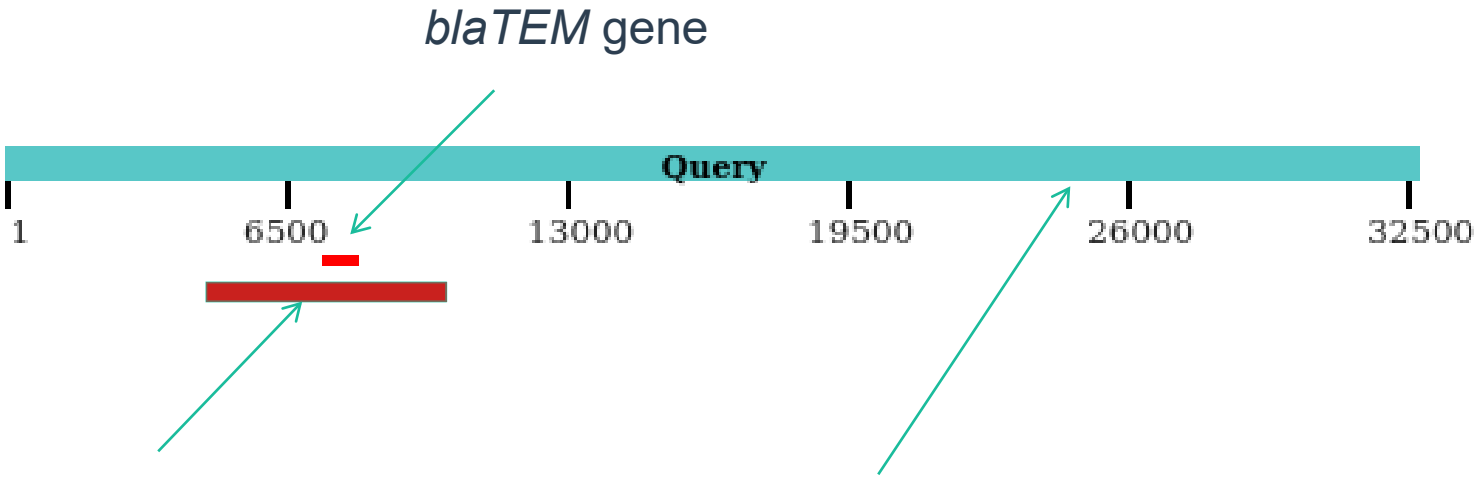
Validation of Metacherchant results

.We used Spades to assemble the same sample

.We visualized the .graph file of Spades with Bandage

.We extracted the node containing the *blaTEM* gene and its environment

If everything worked fine...

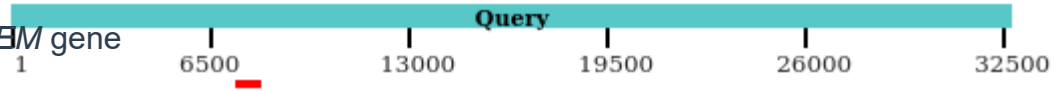


Expected *blaTEM* gene with prolonged flanks

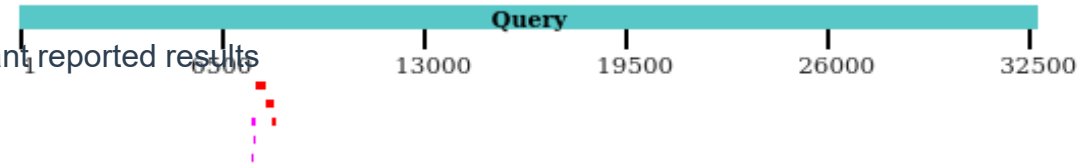
Spades node containing blaTEM gene and its environment

Validation of Metacherchant results

des node containing *blaTEM vs blaTEM* gene



containing *blaTEM vs* Metacherchant reported results



Future work

.Which databases should we use? Continue testing and benchmarking

.Using snakemake framework for downloading and analyzing metagenomic datasets

Thank you!