ContFree-NGS: Removing Reads from Contaminating Organisms in NGS data

Felipe Vaz Peres[1,2] and Diego Mauricio Riaño-Pachón[2]
1 - Federal University of São Carlos, Araras, SP, Brasil
2 - Laboratory of Computational, Evolutionary and Systems Biology, Center of Nuclear Energy in Agriculture, University of São Paulo, Piracicaba, SP, Brasil

Contamination problem

A contaminating sequence is one that does not faithfully represent the genetic information from the biological source organism because it contains one or more sequence segments of foreign origin, and they could cause several problems in downstream analysis.

The primary consequences of contamination are:

- time and effort wasted on meaningless analysis
- erroneous conclusions drawn about the biological significance of the sequence
- misassembly of sequence contigs and false sequence clustering
- delay in the release of the sequence in public databases
- pollution of public databases

National Center for Biotechnology Information 2016, Contamination in Sequence Databases. https://www.ncbi.nlm.nih.gov/tools/vecscreen/contam/

Contamination in public RNA-Seq datasets

ContFree-NGS was developed to solve a problem I faced when I was working with sugarcane RNA-Seq datasets.

%	reads	species
2.685045	1.187.361	Acinetobacter baumannii
1.004987	444.418	Coccomyxa subellipsoidea
0.639131	282.632	Arthrobacter sp. Y81
0.562073	248.556	Paenibacillus sp. IHB B 3415

Contamination removal

Recently, some tools have been made available that aim to remove sequences from contaminating organisms in next generation sequencing (NGS) datasets:

- DecontaMiner is a tool to unravel the presence of contaminating sequences in the set of reads that do not map to a reference genome
- Conterminator removes contaminating sequences from contigs exploiting a taxonomic assignment file
- QC-Blind is an automatic tool to do unsupervised assembly and contig binning to identify and remove putative contaminants

These tools have in common that they either require a reference genome of the source organism, or need to perform assembly prior contaminant detection.

Sangiovanni, M., Granata, I., Thind, A. et al. From trash to treasure: detecting unexpected contamination in unmapped NGS data. BMC Bioinformatics 20, 168 (2019). https://doi.org/10.1186/s12859-019-2684-x
Steinegger, M., Salzberg, S.L. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol 21, 115 (2020). https://doi.org/10.1186/s13059-020-02023-1
Xi, W., Gao, Y., Cheng, Z. et al. Using QC-Blind for Quality Control and Contamination Screening of Bacteria DNA Sequencing Data Without Reference Genome. Frontiers in Microbiology 10 (2019): 1560. https://doi.org/10.3389/fmicb.2019.01560

Our goal

Our goal was to develop a simpler tool to remove contaminated sequences directly from unassembled reads, without mapping, exploiting fast k-mer analysis implemented in taxonomic assignment engines commonly used in metagenomics.

KAIJU

Fast and sensitive taxonomic classification for metagenomics

Kraken 2

Taxonomic Sequence Classification System

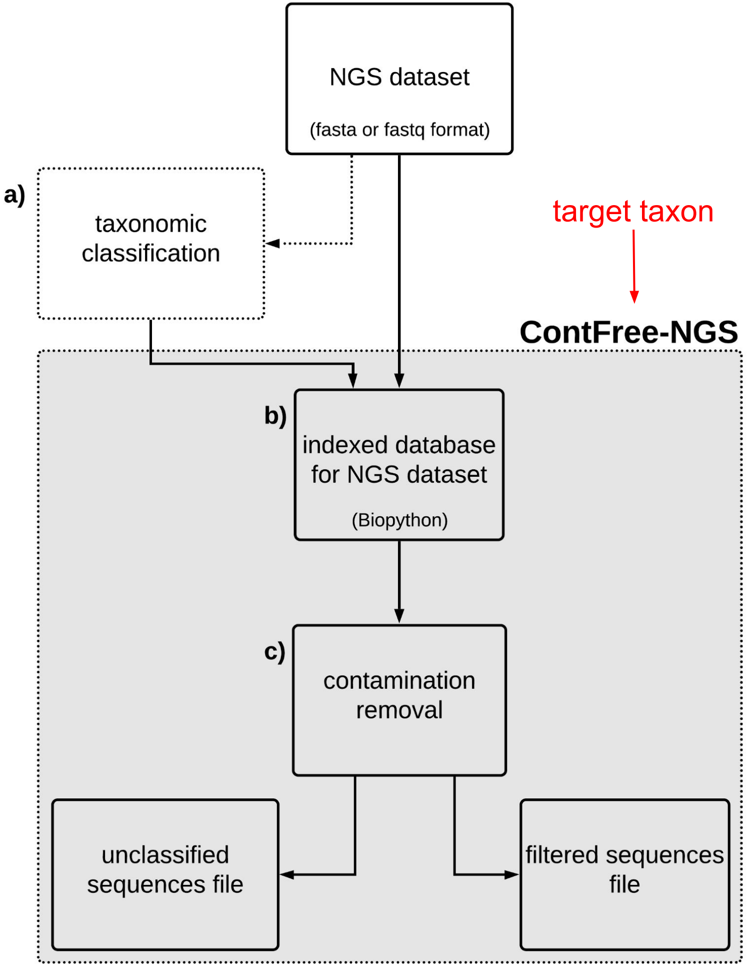
Menzel, P., Ng, K. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7, 11257 (2016). https://doi.org/10.1038/ncomms11257

Wood, D.E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. Genome Biol 20, 257 (2019). https://doi.org/10.1186/s13059-019-1891-0

Our approach

ContFree-NGS was implemented as a single Python v3 (>3.6) script, using the biopython module and the Python Environment for Tree Exploration (ETE).

In order to assess contamination, ContFree-NGS exploits a taxonomic assignment file containing the read identifier and a NCBI taxonomic identifier for every sequence in the dataset.



Taxonomic classification

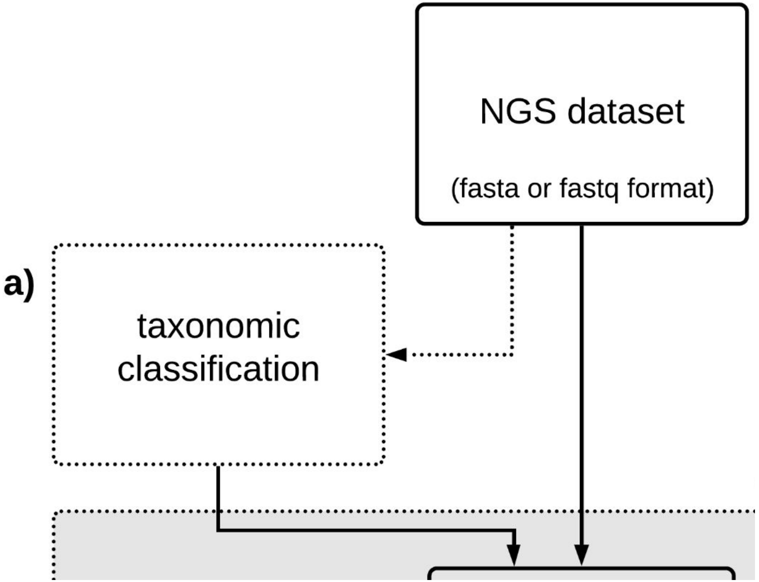
As ContFree-NGS exploits the results from a taxonomic assignment engine, users must use the proper switches to achieve an accurate classification, for instance a proper value of the --confidence switch in Kraken2.

To perform the taxonomic classification with Kraken2, we built a custom database containing the following reference libraries:

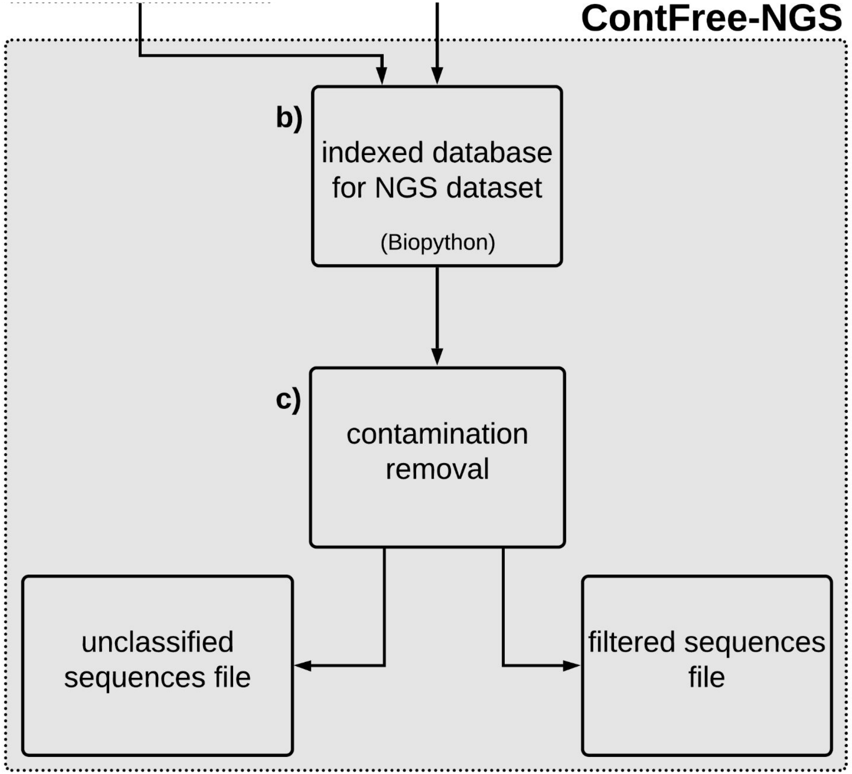
- Archaea
- Bacteria
- Viral
- Human
- Fungi
- Plant
- Protozoa
- NCBI non-redundant nucleotide database

Kraken output:

```
C	001	33090
U	002	0
```



Indexed database



The first step of ContFree-NGS is to create a SQL database to store the sequence index to make the contaminant removal process faster.

This is done using the Bio.SeqIO.index_db function from biopython.

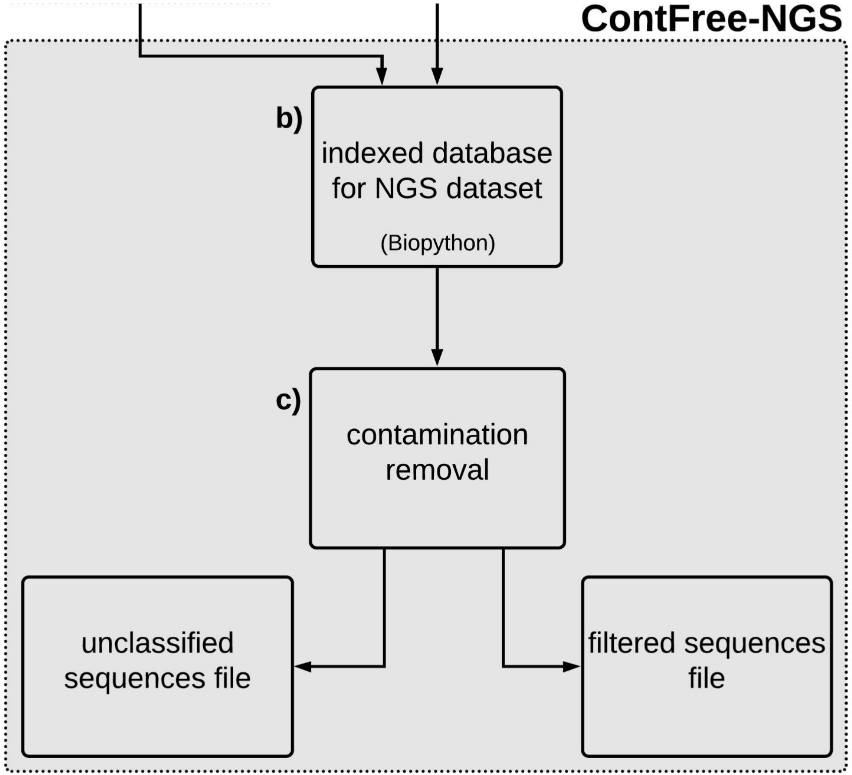
Contamination removal

ContFree-NGS creates a list with the NCBI identifier for the target taxon and all the identifiers of its descendants according to the NCBI taxonomy database.

Example: [**33090**, 128810]

Then, ContFree-NGS iterates over the taxonomy assignment file.

- If the read was not assigned to any taxa it is saved in a fastq file for unclassified reads.
- If the read was assigned, it will check if its taxon is found in the list of the target taxon descendants, if so, will save the read to a fastq file for filtered reads, otherwise the read will be discarded.

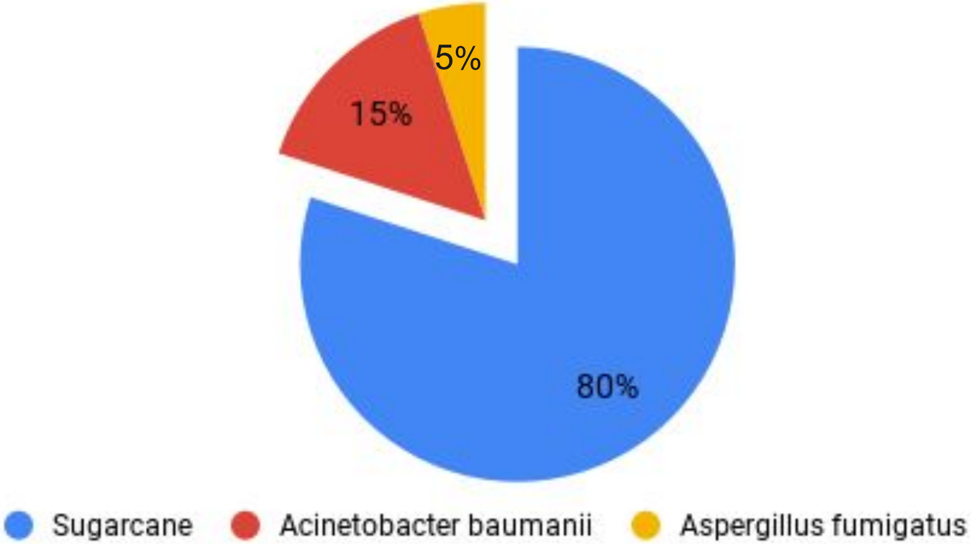


Evaluation on artificially contaminated datasets

We evaluated ContFree-NGS on three sugarcane artificially contaminated datasets:

- A: 50.000 paired end reads
- B: 250.000 paired end reads
- C:1.250.000 paired end reads

In all datasets 80% of the reads came from sugarcane (SRR1774134), 15% came from Acinetobacter baumanii (SRR12763742) and 5% came from Aspergillus fumigatus (DRR289670).



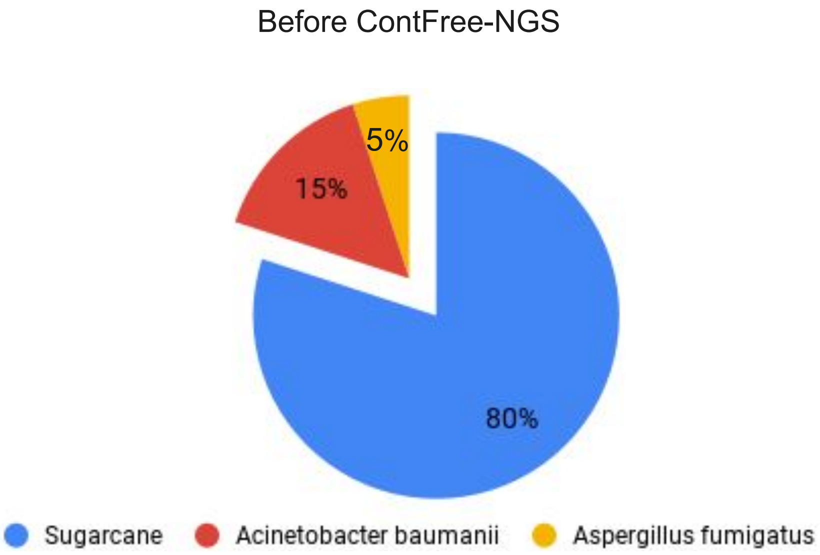
Evaluation on artificially contaminated datasets

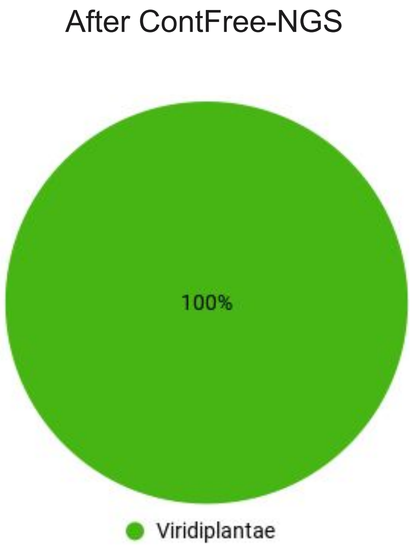
Running Kraken 2 with --confidence 0.05, the following sequences was classified in some taxon:

- A: 25.547 reads
- B: 128.396 reads
- C: 664.270 reads

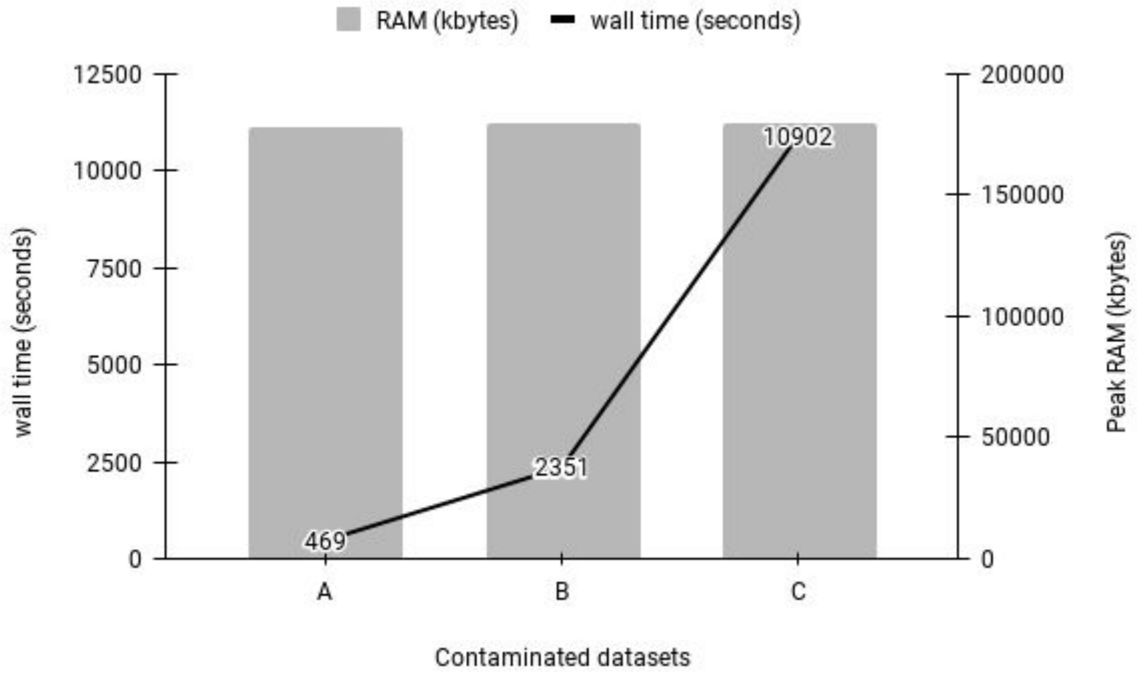
For the three datasets, nearly half of all sequences were assigned to a taxon.

Evaluation on artificially contaminated datasets

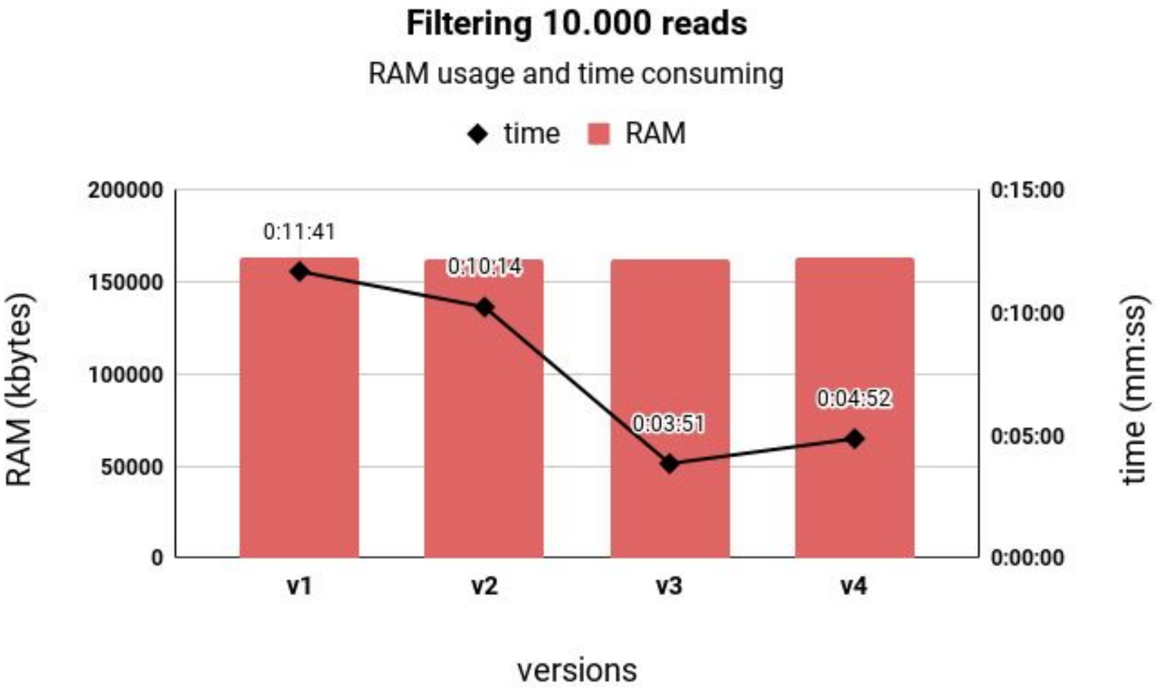




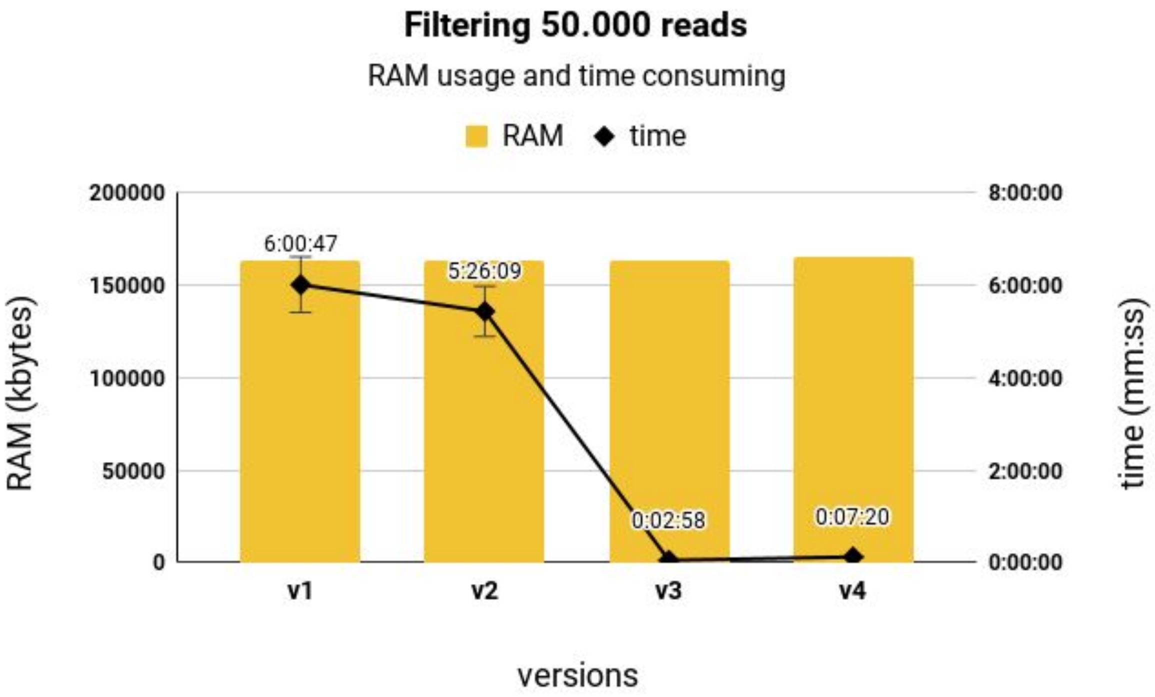
RAM usage and time consuming



Performance



Performance



Conclusions

ContFree-NGS is a very simple filter and useful tool that removes sequences from contaminating organisms in a NGS dataset.

Memory usage is low and independent of the number of classified sequences and wall time scales rapidly with the number of classified sequences.

As ContFree-NGS exploits the results from a taxonomic assignment engine, users must use the proper switches to achieve an accurate taxonomic classification.

Thanks!

Have any questions or suggestions?
Contact: felipe.vzps@gmail.com