**Filtering Reads Made Easy**

**<NAMES>**

1 [The Shmunis School of Biomedicine and Cancer Research](https://en-lifesci.tau.ac.il/lp-en-mcbb), George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

† To whom correspondence should be addressed:

Tal Pupko, Tel: +972 3 640 7693; Fax: +972 3 642 2046; E-mail: talp@tauex.tau.ac.il

Keywords: sequencing, webserver, DNA contamination, interactive webpage

# Abstract

Recent advances in sequencing and informatics technologies have led to a deluge of sequencing data. This has dramatically changed our understanding of the genome and the understating of biology. While it is now relatively easy to sequence, extracting only meaningful reads is still a challenge. Here we introduce a new webserver to facilitate the filtering process. Three of the main features implemented in <Name> are: (1) automated updating of the latest bacteria database; (2) fast comparison of the input reads to the database; (3) a simple interactive web page to further analyze the results. Our goal is to be a one stop shop for cleaning sequencing data from contaminations.

# Introduction

Over the past decades, the price of sequencing has been decreasing rapidly (Check Hayden 2014). This allowed research groups all over the world to sequence different species and doubled the size of the GenBank every few months (Benson et al. 2018). The data provided by those sequencings is a vital resource in biology and may clarify many unknowns.

With the increase of DNA sequencing, confounding contaminant DNA takes place as one of the main challenges. Instead of finding the genome of the required species, some of the reads are contaminated and originate in different species (Steinegger and Salzberg 2020). This contamination can have a significant effect on the later analysis including false positive SNP’s identification (Goig et al. 2020), incorrect labels on sequences in metagenomic studies (Kirstahler et al. 2018).

The process of filtering sequencing data is a laborious process, with many steps, including downloading and maintaining databases, and understanding complicated algorithms. Some research laboratories began implementing their own in-house analysis pipelines, and later, different search engines emerged (Wood, Lu, and Langmead 2019). These applications require specific working environments (i.e., operating systems), computation power (multicore machines), more than basic technological skills (e.g., installation and running), and many parameters to configure. Other laboratories do not filter the reads and hence risk incorrect results.

Here we present <Name> web server. The web server was developed to easily filter genomic reads and make it more accessible for the scientific community. There is no installation required and no other prerequisites are needed. The server automatically updates the databases and provides a simple and interactive graphical user interface (i.e., GUI) to personalizes the user output (Figure 1). Visual and textual results that are ready for publication or further analysis.

# Materials and Methods

## Input

The <Name> web server requires a “fasta” or “fastqc” file containing the reads to be filtered. Future versions will include different databases so, the user may specify which database to search for contaminations against.

## Database

Each month, we download the full databases from NCBI This is an automated process to verify our genomes are up to date. When the download is complete, the data is preprocessed based on the search engine.

## Search Engine

To maximize both speed and accuracy, we use the Kraken 2 search engine (Wood, Lu, and Langmead 2019). With a given read file, each read is split into k-mers and then searched in the relevant database. K-mers are substring of the reads with a k const length. For example, 3-mer for the read: “ATGG” will be: “ATG” and “TGG”. The output of the Kraken 2 search engine is a csv file containing all the reads with a list of species with the number and identity of k-mers found in this read.

## Preprocess Kraken Output

The Kraken algorithm matches k-mers in each read to its database, and classifies the read based on the lowest node in the database's taxonomic tree with matched k-mers. As the Kraken does not provide a confidence score for this classification, two problems arise. The first is that there is no matric for similarity between the read and the databased searched against. In addition, there may be many subspecies in the results, making it difficult to interpret. In order to mitigate these issues, we count matched k-mers by the Kraken and classify the organism based on the most common k-mer's association. We provide a score, based on the precent of the most common k-mer out of all classified k-mers for that organism. This re-classification causes many classifications to be of high taxonomic order such as: kingdom or domain mitigating the second issue. To find a middle ground, we cut the found k-mers at chosen taxonomic order. These results, can be tweaked by the user in order to get custom made output, deciding what should be considered contamination.

Due to the fact that we display specific classification groups, in large datasets there was an issue with the webpage runtime. In order to display the data in graphs fast, we transform the list of the reads (this list may contain millions of reads) into a matrix. The rows of the matrix contain the k-mers percentage, the amount of k-mers of the max species divide by the length of the read. This value represents the similarity of read to the database searched against. The columns of the matrix are the top species names. The cells are the number of reads that fits the specific row and column. This allows us to present user interactive graphs at high speeds.

## User Interactive Analysis

First, the matrix described above is loaded. Then, two graphs are created, one a histogram for the reads amount against the similarities to one of the databases genomes (Figure 2). And the other is a pie chart with all the reads classified by the top contamination genomes (Figure 3). The sum of each row represents the number of reads with this similarity. While the sum of each column represents the number of reads associated with this species. One could find its optimal threshold of k-mer percentage by controlling the slider. In addition, one could control the species list that will be classified as contamination. Those two parameters will immediately change the graphs so, one could see the effect in real time. When the user is ready to export the results, the threshold for the k-mer percentage and the species list will be delivered to classify each of the reads.

## Post Process

The GUI represents only summarized results so filtering the reads by the user choice needs another process. To consider a read contaminant we verify the k-mer percentage is above the k-mer threshold (chosen by the user) and the read is inside the species to contaminate. When the post process is finished, the user will get a “.gz” file containing all the noncontaminated reads.

## Implementation

<NAME> is implemented in Python using Flask framework. The source code is available at: [<GITHUB](https://github.com/orenavram/microbializer) LOCATION>. The web server jobs are processed on ProLiant XL170r Gen9 servers, equipped with 128 GB RAM and 28 CPU cores per node. The Gallery, Overview, and Frequently Asked Questions (FAQ) sections of the web server should help users get the most out of the web server. A running example (different from the case studies analyzed in the Gallery) is also provided.

# Test Case

We have downloaded an RNA-seq of Henneguya Salminicola from NCBI (i.e., SRR7754566). We uploaded the file to <NAME> and analyzed the sequencing reads. The uploaded file contained 32,923,720. Of all the reads, 3,314 had a certain amount of similarity (k-mer percentage in range of (0.24, 1.0)) to the bacteria database. Our results fit the NCBI prediction as a large contamination group is Pseudomonas.

In the results interactive page, we choose k-mer percentage threshold of <XX> and a species list containing: <YY>. The resulting read’s file contained <XX> reads as expected. We reuploaded the file to the webserver to verify the same results. The results of the filtered file contained the same number of reads as excepted, no new species were found (except the ones we did not filter) <MAYBE INSERT FIGURE>.

# Funding

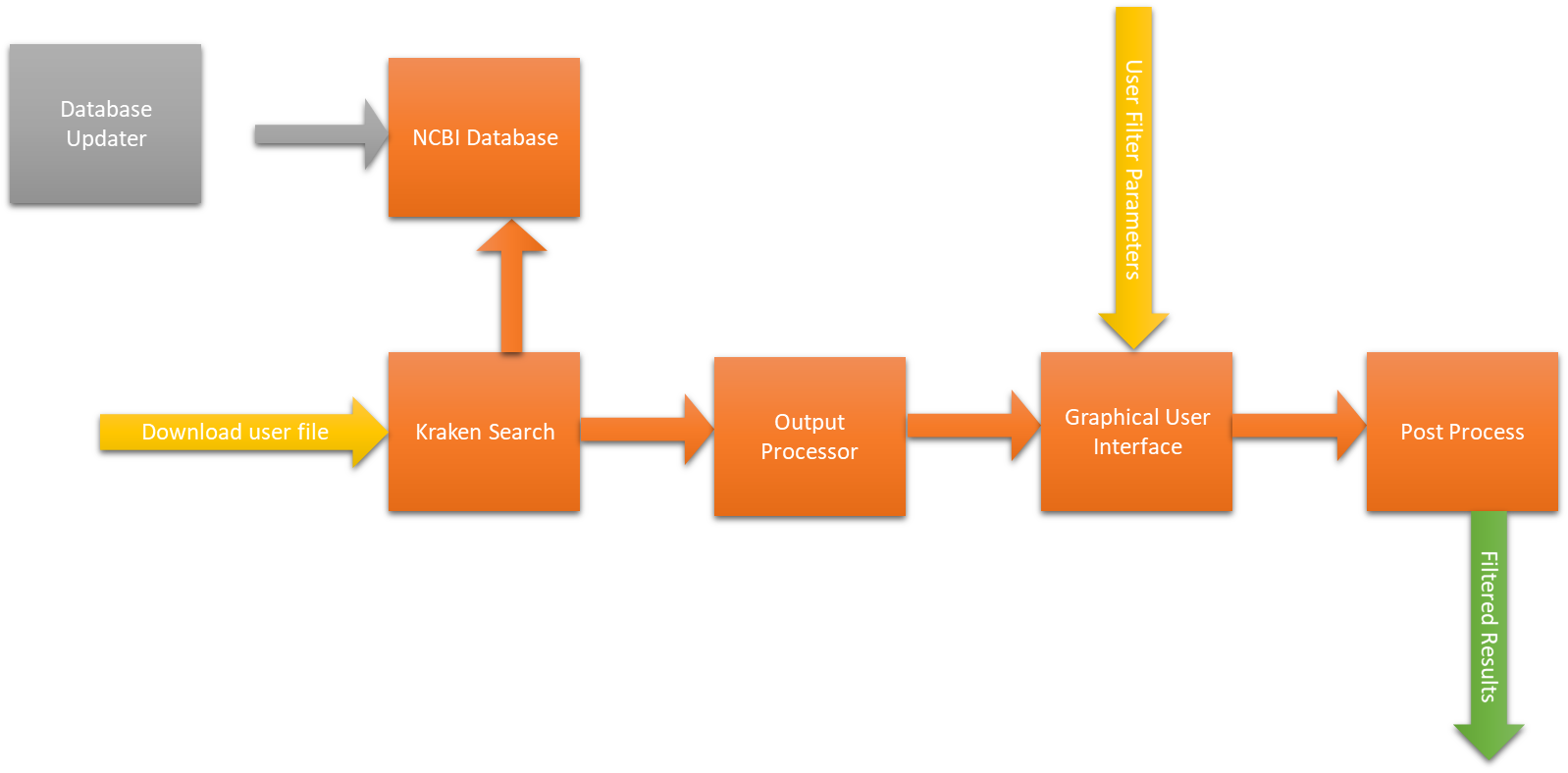
Israel Science Foundation (ISF) [802/16 to T.P.]; Edmond J. Safra Center for Bioinformatics at Tel Aviv University Fellowship (in part). Funding for open access charge: ISF. <VERIFY WITH TAL>

Conflict of interest statement. None declared.

# Figures

## Figure 1.

Web server workflow.



## Figure 2.

This is the histogram graph displayed in the interactive GUI. The x-axis represents the reads similarities to the database searched against (in this example, bacteria database). The red line represents the current k-mer percentage threshold. The blue bars represent non-bacteria reads, and the orange bars represent reads that will be filtered. By clicking on the chart, one can move the k-mer threshold. Note, that some of the reads that are right to the threshold might be filtered, as if are in the species list of contamination which is chosen in the pie-chart.

Chart, histogram

Description automatically generated

## Figure 3

This is the pie-chart graph displayed at the interactive GUI. The graph displays the contamination species and the distribution of the reads. Each color displays a different species, by moving the mouse over the graph, one can see the species name and the number of reads. To choose a species to set as non-contamination, simply uncheck the species box.

Chart, sunburst chart

Description automatically generated

**References**

Benson, Dennis A, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, James Ostell, Kim D Pruitt, and Eric W Sayers. 2018. “GenBank.” *Nucleic Acids Research* 46 (D1): D41–47. https://doi.org/10.1093/nar/gkx1094.

Check Hayden, Erika. 2014. “Technology: The $1,000 Genome.” *Nature* 507 (7492): 294–95. https://doi.org/10.1038/507294a.

Goig, Galo A., Silvia Blanco, Alberto L. Garcia-Basteiro, and Iñaki Comas. 2020. “Contaminant DNA in Bacterial Sequencing Experiments Is a Major Source of False Genetic Variability.” *BMC Biology* 18 (1): 24. https://doi.org/10.1186/s12915-020-0748-z.

Kirstahler, Philipp, Søren Solborg Bjerrum, Alice Friis-Møller, Morten la Cour, Frank M. Aarestrup, Henrik Westh, and Sünje Johanna Pamp. 2018. “Genomics-Based Identification of Microorganisms in Human Ocular Body Fluid.” *Scientific Reports* 8 (1): 4126. https://doi.org/10.1038/s41598-018-22416-4.

Steinegger, Martin, and Steven L. Salzberg. 2020. “Terminating Contamination: Large-Scale Search Identifies More than 2,000,000 Contaminated Entries in GenBank.” *Genome Biology* 21 (1): 115. https://doi.org/10.1186/s13059-020-02023-1.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *Genome Biology* 20 (1): 257. https://doi.org/10.1186/s13059-019-1891-0.