bioRxiv pre-print LaTeX v0.2

Title:

Recentrifuge: robust comparative analysis and contamination removal for metagenomic data

Running title:

Recentrifuge

Jose Manuel Martí[1,#]

[1]Institute for Integrative Systems Biology (I[2]SysBio), Valencia, Spain.

October 24, 2017

Words count for the Abstract section: 152

Number of floats: 7

Words count for the rest of text (NOT inc.refs): 2153

- 180 words in TitlePage
- 8 words in Keywords
- 941 words in Section: Introduction
- 235 words in Section: Methods
- 299 words in Section: Results
- 490 words in Section: Discussion

[#] Contact: jose.m.marti@uv.es

DRAFT

bioRxiv pre-print LaTeX v0.2

Abstract

Metagenomics sequencing is becoming more and more widespread in biomedical and biological research, and the pace is increasing thanks to nanopore sequencing. With a rising number of samples and data per sample, the challenge of efficiently comparing results within a sample and between samples arises. Such analysis is complicated by laboratory and host contaminants which demands a reliable method for the removal of negative control samples from the set of samples. This is critical in samples with low microbial biomass, where contamination can comprise most of a sample if not all. With Recentrifuge, researchers can analyze results from the Centrifuge classifier using interactive hierarchical pie charts with special emphasis on the score level of the classifications. In addition to control-substracted samples, shared taxa and exclusive taxa per sample are also provided with scores for each taxonomic level of interest, thus enabling robust comparative analysis of multiple samples in low biomass metagenomics studies.

***Keywords*—** metagenomics, comparative analysis, sample contamination, low biomass, Centrifuge

Introduction

Studies of microbial communities are becoming more and more popular in both clinical and biological arenas. These studies are boosted by new DNA and RNA sequencing technologies which are dramatically decreasing the cost per sequenced base. With the development of nanopore sequencing, portable and affordable real-time shotgun metagenomic sequencing (SMS) is a reality (*1*). With these technologies, sets of sequences belonging to microbial communities from different sources and times can be retrieved and analyzed to unravel spatial and temporal patterns in the microbiota (see Figure 1 for an example). In those studies involving SMS, DNA is extracted from each sample using a commercial kit or an optimized custom

bioRxiv pre-print LaTeX v0.2

protocol, the purified DNA is then sequenced and the sequences analyzed by a bioinformatics pipeline. This entire procedure is summarized in Figure 2 but the core of the analysis is detailed in Figure 3.

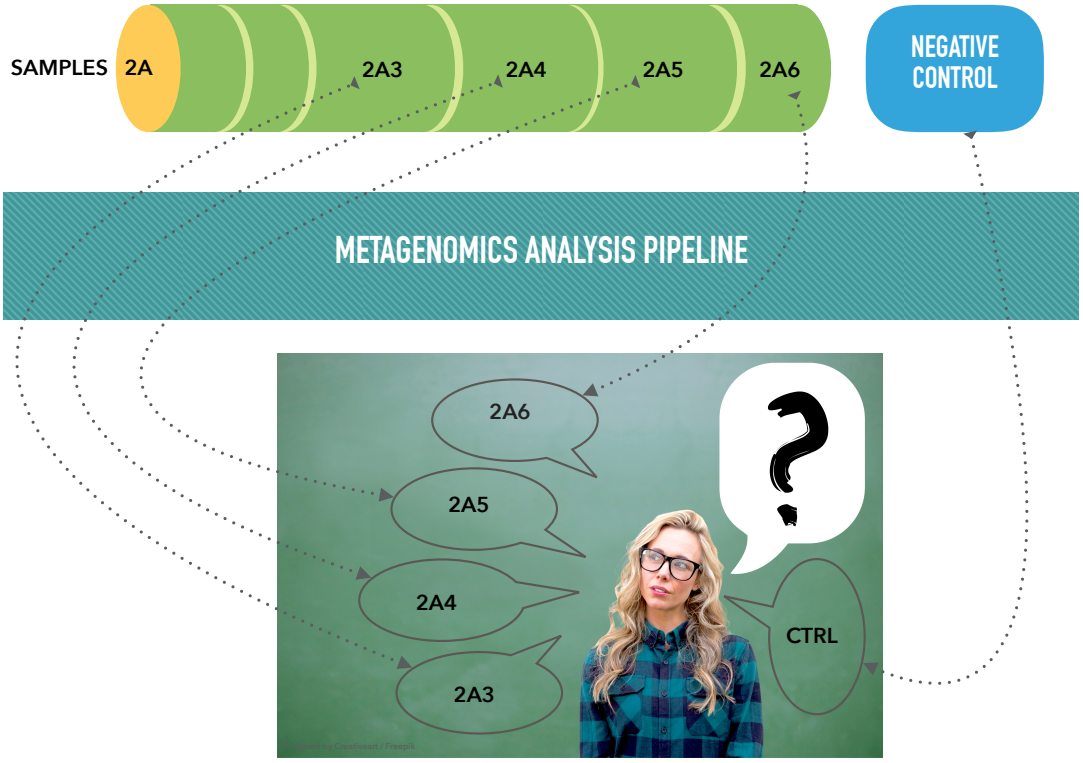

Figure 1. This is an example outlining the problem of comparing different but related samples in a SMS study. The sample named 2A is subdivided longitudinally in six subsamples whose DNA is extracted along with a negative control sample. The purified DNA is then sequenced and the generated sequencing reads are processed through a metagenomics analysis pipeline (as the one detailed in Figure 2). A collection of seven different datasets are finally generated, which should be adequately compared in order to elucidate lengthwise patterns in the microbiota within the 2A sample.

In case of low biomass samples there is very little microbial DNA; the library preparation and sequencing methods will return sequences whose main source is contamination (2). From the data science perspective, this is just another instance of the importance of keeping a good *signal to noise ratio*, otherwise, when the *signal* (real DNA from the sample) approaches the order of magnitude of the *noise* (DNA from contamination), special methods are required to try to tell them apart.

bioRxiv pre-print LaTeX v0.2

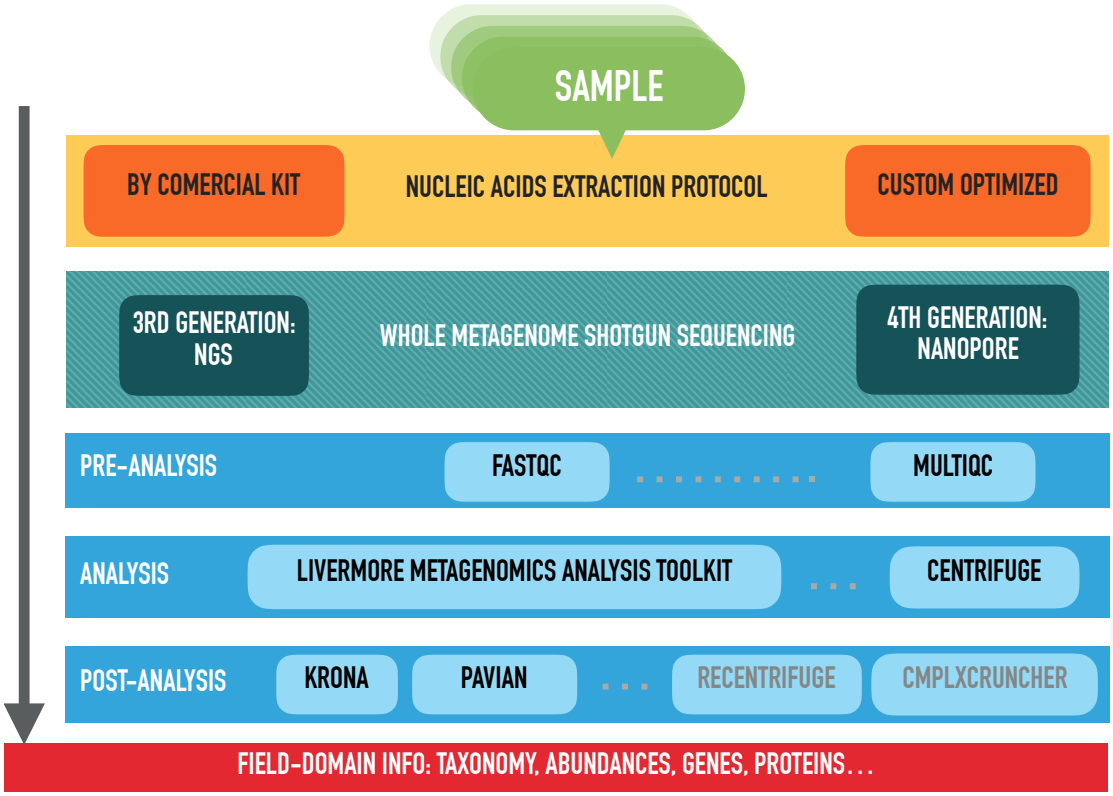

Figure 2. A study involving SMS, like the one illustrated by Figure 1, spans in a number of stages to extract valuable field-domain information starting from the original samples. For each sample, the DNA is extracted using a commercial kit, a custom protocol optimized for the type of sample or a combination of both. Next, a library matching the sequencing technology to be used is prepared with the purified DNA, which is then sequenced. The reads provided by the sequencer are processed through a bioinformatics pipeline that we could roughly separate in three consecutive steps: in the pre-analysis the reads are quality checked using codes like FastQC (Babraham Bioinformatics, 2016) and MultiQC (*3*); in the analysis stage, the most computational intensive, the reads are taxonomically or functionally classified using software packages like LMAT (*4*, *5*) and Centrifuge (*6*) (see Figure 3 for details); finally, in the post-analysis step, the results are further processed to enable deeper analysis and improved visualization by using different tools like Krona (*7*), Pavian (*8*) and others.

The roots of contaminating sequences are diverse, as they can be traced back to nucleic acid extraction kits (the *kitome*), reagents, airborne particles, crossover between samples, or other sources. In fact, such *kitome* varies even within different lots of the same kit. For example, the PowerSoil DNA Isolation Kit, a kit that usually provides high yields and has been widely used, including Earth Microbiome Project and Human Microbiome Project, usually yields a background contamination by no means negligible (*2*). The lower the biomass in the sam-

bioRxiv pre-print LaTeX v0.2

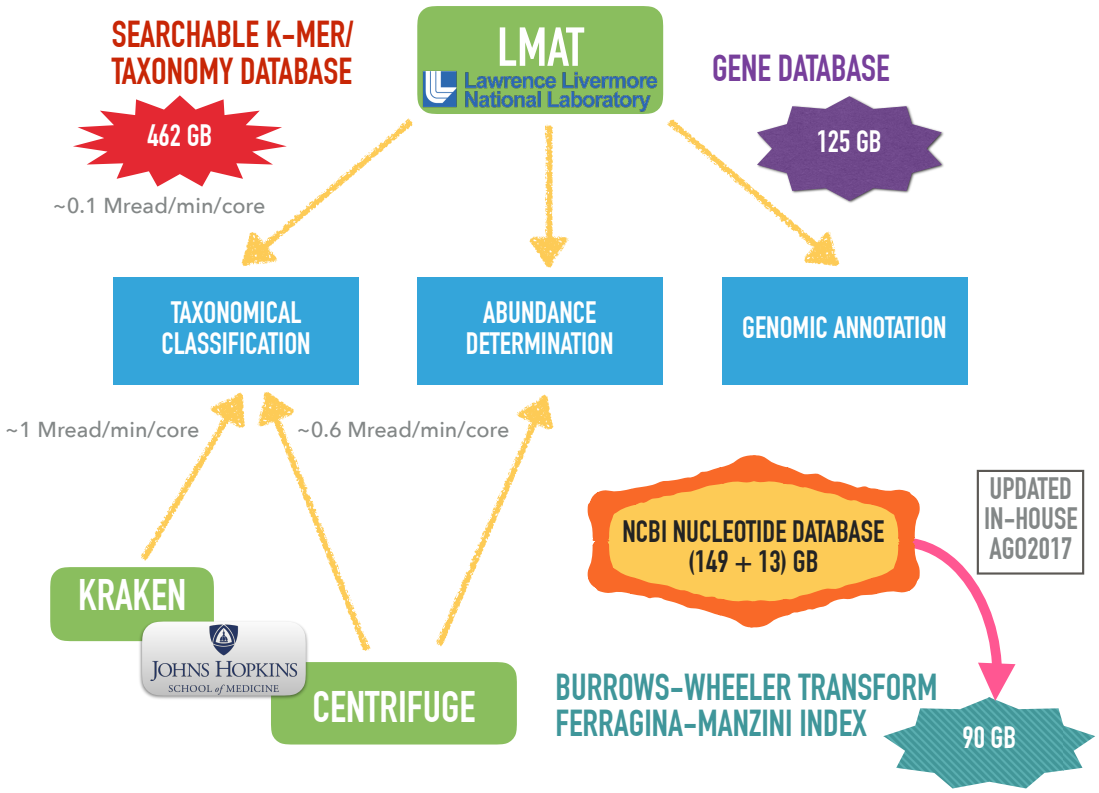

Figure 3. The core phase of a metagenomics analysis pipeline (see Figure 1 and 2 for the outline of the bioinformatics phases) is carried out by high performance computing software. These are intensive codes in both CPU and memory (sometimes, they are input/output intensive too), like LMAT (*4*), Kraken (*9*) and, more recently, Centrifuge (*6*). All these tools are performing taxonomic classification and abundance estimation, whereas LMAT is also able to annotate genes. For the taxonomic classification, both LMAT and Kraken use an exact k-mer matching algorithm with large databases ($\sim$ 100 GB) while Centrifuge use compression algorithms to reduce the databases size ($\sim$ 10 GB) but at some speed expense. The most complete LMAT database is approaching half terabyte of required memory while the Centrifuge database generated in-house in August 2017 from the NCBI Nucleotide (*10*) *nt* database (149 GB plus 13 GB in indexes) is occupying just 90 GB.

ples, the more essential is to collect negative control samples to help in the contamination background assessment, because without them it would be almost impossible to distinguish true microbial from contamination signals.

Once the real DNA from the samples has been told from the contaminating DNA, the problem of performing a reliable comparison between samples remains. The sequence reads are generally assigned to different taxonomic ranks, specially when using a more conservative

bioRxiv pre-print LaTeX v0.2

approach like Lowest Common Ancestor (LCA) (4). Even if LCA is not used, the reads are classified with diverse scores. Despite this difficulties, it is still more challenging to compare samples with very different DNA yields, for instance, low biomass samples against higher microbial DNA ones, because of the different resolution in the taxonomic levels. This sort of problem is also true when the samples, even with DNA yields in the same order of magnitude, have a quite different microbial structure so that the minority and majority microbes are essentially different between them.

Recentrifuge enables researchers to analyze results from the Centrifuge classifier using Krona-like charts (7), but placing emphasis on the score level of the classifications by the use of a scaled color-map. In addition to the scored plots for the raw samples, four different sets of scored plots are generated for each taxonomic level of interest: control-subtracted samples, shared taxa (with and without subtracting the control) and exclusive taxa per sample. This battery of analysis and plots permits robust comparative analysis of multiple samples, specially in the case of low biomass metagenomics studies.

Methods

The visualization part of Recentrifuge is based on interactive hierarchical pie charts. The Krona (7) 2.0 JavaScript (JS) library developed at the Battelle National Biodefense Institute has been adapted and extended to fit the particular needs of the problem.

The following list abridges the implementation details concerning the Recentrifuge computing kernel:

1. It is a full statically annotated Python 3.6 code. PEP-8 (11), PEP-484 (12) and PEP-526 (13) compliant. Written following the Google python style guide (14). Pylint and mypy checked.

bioRxiv pre-print LaTeX v0.2

2. Coded using python multi-platform parallelization (both by input file and by taxonomic level) in order to reduce the elapsed time when dealing with huge datasets.

3. Intensive use of recursive methods to cope with tree-arithmetics, which enables robust comparison between taxonomic trees at any rank. The code is informed of the 30 different taxonomic levels used in the NCBI taxonomy (*16*).

4. Implemented under a object-oriented paradigm to ease future extensions targeting new or improved uses; it is prepared for other taxonomies different from NCBI and for additional input formats. A summarized UML (Unified Modeling Language) (*15*) class diagram of Recentrifuge is shown in Figure 4.

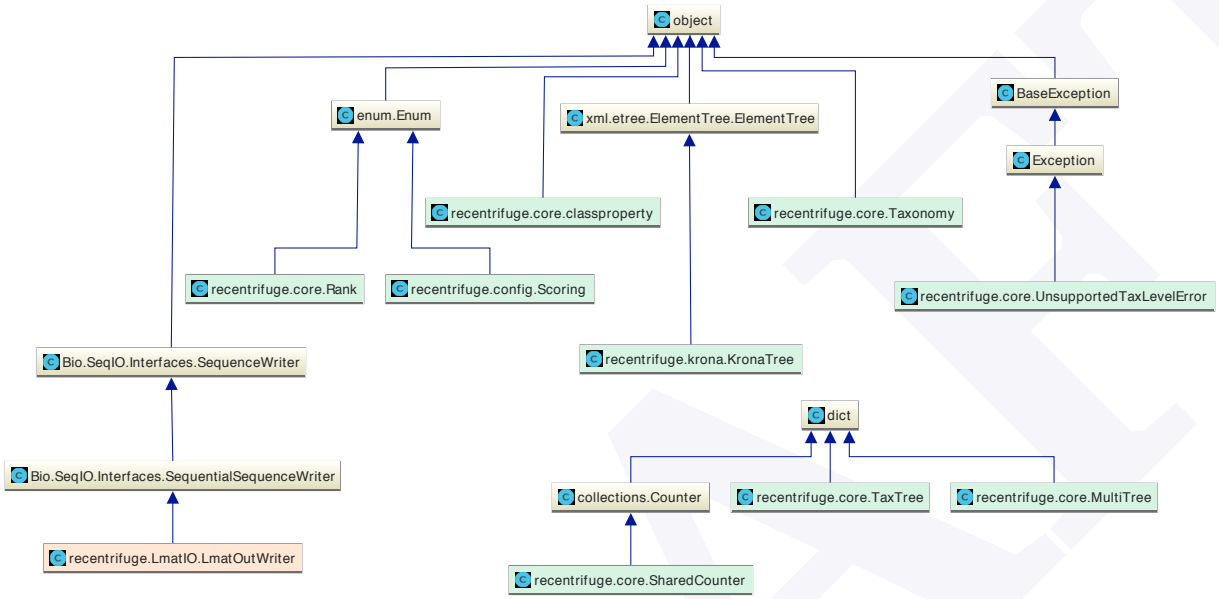

Figure 4. Unified Modeling Language class diagram summarizing relationships between classes in the Recentrifuge package. The green background denotes developed classes currently used in the package, while the orange background indicates classes developed for future use. The rest of classes in the figure, the parent classes from which the Recentrifuge ones are derived, belong to Python Standard Library and BioPython.

bioRxiv pre-print LaTeX v0.2

Results

Recentrifuge is a metagenomics analysis software with the computing kernel implemented and parallelized using Python 3.6, and the visualization part in JavaScript, adapted from the Krona (*7*) JavaScript library. The Figure 5 summarizes the package context and data flows. Currently, Recentrifuge straightforwardly accepts Centrifuge (*6*) output files, thus enabling a scored oriented analysis and visualization.

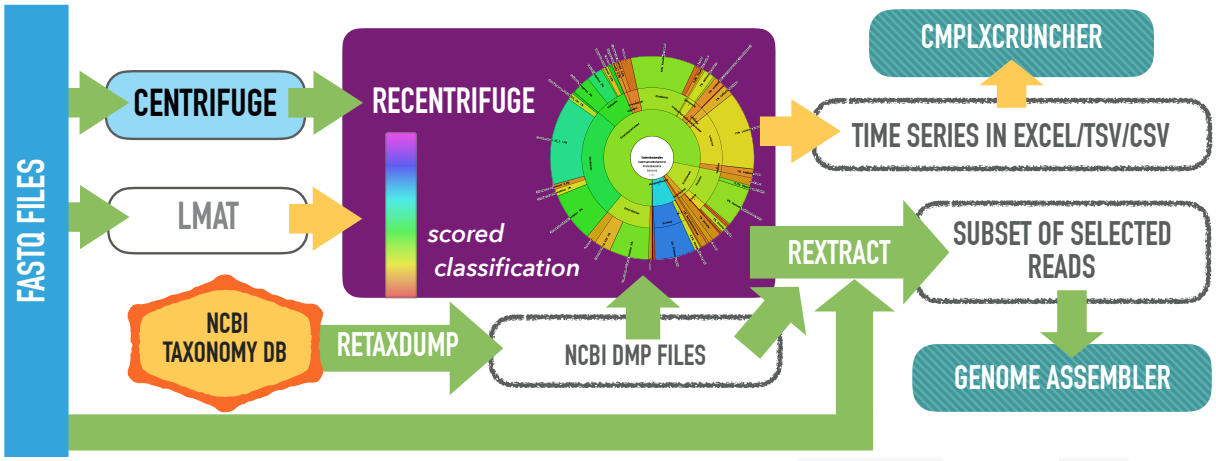

Figure 5. Outline of the Recentrifuge package with its context and main data flows. The green arrows show fully implemented pipelines. Currently, Recentrifuge accepts Centrifuge (*6*) direct output files, thus enabling a scored oriented visualization. The yellow arrows indicate pipelines only partially developed: LMAT (*4*) results as input for analysis, and output of Recentrifuge to different text field formats to enable further longitudinal (time or space) series analysis, for example, using CmplxCruncher (in development). The NCBI Taxonomy (*16*) dump databases are easily retrieved using *Retaxdump* script. *Rextract* utility extracts a subset of reads of interest from the paired-ends FASTQ input files, which can be used in any downstream application, like genome assembling and visualization, for instance, using IGV (Broad Institute).

To ensure the widest portability for the interactive visualization of the results, the central outcome of Recentrifuge is a stand-alone HTML file which can be loaded by any JavaScript enabled browser. The package also provides statistics about the reads and their classification performance. In addition, the user can choose between different score visualization algorithms, some of which are more interesting for datasets containing variable length reads, like the ones generated by Oxford Nanopore sequencers.

bioRxiv pre-print LaTeX v0.2

Different derived products are obtained using auxiliary tools in the package, like *Rextract*, a script which helps in extracting a subset of classified reads of interest from the paired-ends FASTQ input files. These reads can be used in any downstream application, like genome visualization and assembling.

Recentrifuge is currently available in GitHub licensed under different open-source licenses.

Discussion

Recentrifuge enables score-oriented robust comparative analysis of multiple samples, specially in low biomass metagenomics studies, where contamination removal is a must.

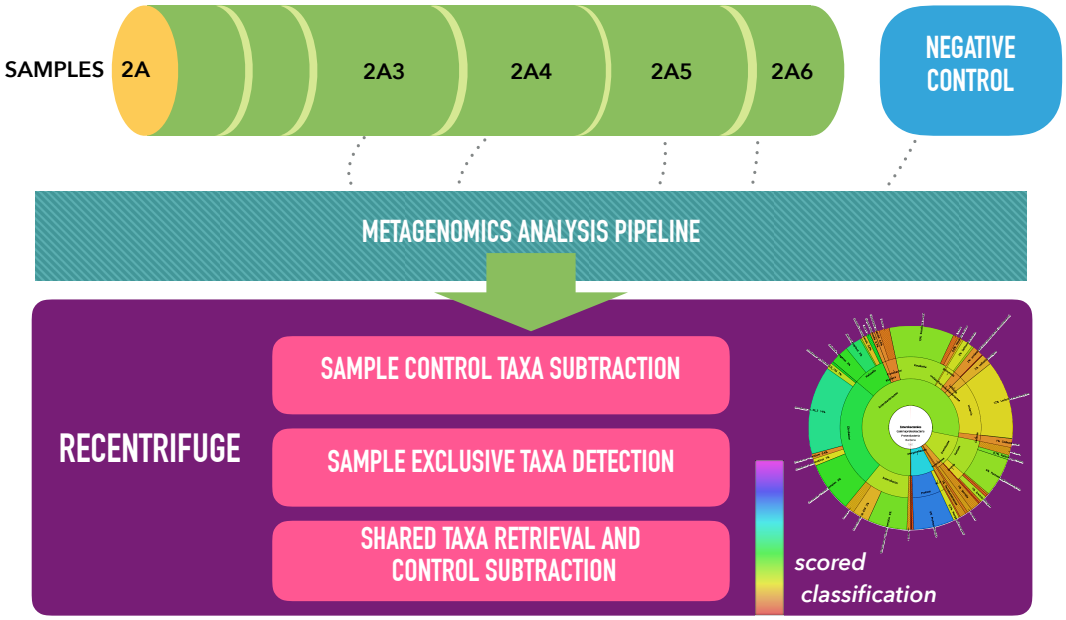

Figure 6. Summary of immediate advantages of applying Recentrifuge to the example of Figure 1, a study involving SMS of different but related samples, including a negative control. Four different sets of scored charts are generated for each taxonomic level of interest in addition to the scored plots for the raw samples: samples with the control taxa subtracted, the exclusive taxa per sample and the shared taxa with and without control taxa subtracted. This battery of analysis and plots permits robust comparative analysis of multiple samples in low biomass metagenomics studies.

The Figure 6 summarizes the instant benefits of introducing Recentrifuge in the SMS example

bioRxiv pre-print LaTeX v0.2

study of Figure 1, while Figure 7 shows a screenshot of the corresponding Recentrifuge HTML interface in such case with scored classification enabled and control taxa subtracted.

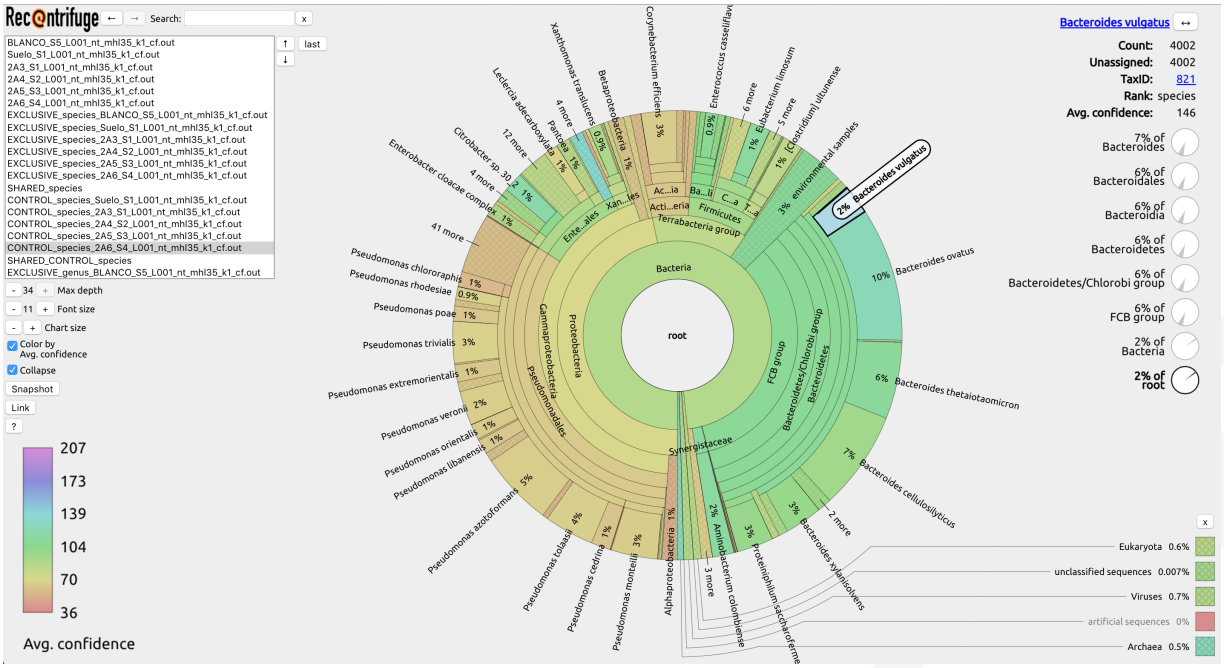

Figure 7. Screenshot of the Recentrifuge HTML interface in the SMS example study of Figure 1 with results summarized in Figure 6. The sample 2A6 after control removal at species level is selected (see the sample selection box in the top left under the Recentrifuge logo). In the center, the corresponding hierarchical pie chart is drawn. For each taxon, the background color reflects the average confidence level of the taxonomic classification following the scale plotted in the bottom left of the figure. In this particular case, the taxon *Bacteroides vulgatus* is selected in the pie chart, thus prompting the display of related statistics and links in the top right of the figure. The current links are to Wikipedia and NCBI Taxonomic Browser (*16*), while the statistics include the number of reads assigned to this or lower taxonomic levels and their average confidence, the number of reads just assigned to this level, the NCBI taxid and rank, and some relative frequencies data.

Just as it is important to accompany any physical measurement by a statement of the associated uncertainty, it is desirable to accompany any read classification by a confidence estimation of the taxon that has been finally assigned. Recentrifuge reads the score given by Centrifuge to any read and uses this valuable information to calculate an averaged confidence level for each classification node. This value may be also a function of further parameters, such as read quality or length, which is specially useful in case of notable variations in the length of the reads, like in the datasets generated by nanopore sequencers.

bioRxiv pre-print LaTeX v0.2

The calculated score is kept in all the downstream analysis and comparisons. That is probably the main difference between Recentrifuge and Pavian (8), which not includes the score of the reads in its analysis. Only a few codes, such as Krona (7) and MetaTreeMap (17), are hitherto able to handle a score assigned to the classification nodes.

By enabling robust comparative analysis of multiple samples in low biomass metagenomics studies, Recentrifuge could have a key role in the analysis of oligotrophic microbes both in clinical or environmental samples.

Acknowledgments

I thank Carlos P. Garay (I^2SysBio), Victoria Fernández (Sistemas Genómicos, S.L.), Jose Fco. Martí and Jose L. García (I^2SysBio) for helpful conversations and suggestions about Recentrifuge.

Funding Information

No funding applies to this work.

Competing interests

The author declares no competing interests.

DRAFT

bioRxiv pre-print LaTeX v0.2

References

1. **Edwards et. al**. 2016. Extreme metagenomics using nanopore sequencing: a field report from Svalbard, 78°N, BioRxiv **073965**.

2. **Kim D, Hofstaedter CE, Zhao C, et al.** 2017. Optimizing methods and dodging pitfalls in microbiome research. Microbiome **5**:52.

3. **Ewels P, Magnusson M, Lundin S, Käller M**. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics **32**:3047-8.

4. **Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, Allen JE**. 2013. Scalable metagenomic taxonomy classification using a reference genome database. Bioinformatics **29**:2253-60.

5. **Ames SK, Gardner SN, Marti JM, Slezak TR, Gokhale MB, Allen JE**. 2015. Using populations of human and microbial genomes for organism detection in metagenomes. Genome Res. **25**:1056-67.

6. **Kim D, Song L, Breitwieser FP, Salzberg SL**. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. **26**:1721-9.

7. **Ondov BD, Bergman NH, and Phillippy AM**. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics **12**:385.

8. **Breitwieser FP, Salzberg SL**. 2016. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. BioRxiv **084715**.

9. **Wood DE, Salzberg SL**. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol **15**:R46.

10. Nucleotide [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (NCBI). 1988. Available here.

bioRxiv pre-print LaTeX v0.2

11. **Van Rossum G, Warsaw B, Coghlan N**. 2013. PEP-8: Style Guide for Python Code [Internet]. Python Software Foundation. Available here.

12. **Van Rosum G, Lehtosalo J, Langa L**. 2015. PEP4-84: Type Hints [Internet]. Python Software Foundation. Available here.

13. **Gonzalez R, House P, Levkivskyi I, Roach L, Van Rosum G**. 2016. PEP-526: Syntax for Variable Annotations [Internet]. Python Software Foundation. Available here.

14. **Patel A, Picard A, Jhong E, Hylton J, Smart M, Shields M**. Rev. 0.59. Google Python Style Guide [Internet]. Google. Available here.

15. International Organization for Standardization. 2012 (rev 2017). ISO/IEC Standard No. 19505-1:2012: Information technology – Object Management Group Unified Modeling Language (OMG UML). Available here.

16. Taxonomy [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (NCBI). 1988. Available here.

17. **Hebrard M, Taylor TD**. 2016. MetaTreeMap: An alternative visualization method for displaying metagenomic phylogenic trees. PloS one **11**:6.

DRAFT