bioRxiv pre-print LaTeX v0.3

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.

January 4, 2018

Words count for the Abstract section: 150

Number of floats: 7

Words count for the rest of text (NOT inc.refs): 3313

- 178 words in TitlePage
- 9 words in Keywords
- 1107 words in Section: Introduction
- 648 words in Section: Methods
- 471 words in Section: Results
- 900 words in Section: Discussion

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

PREPRINT

bioRxiv pre-print LaTeX v0.3

Abstract

Metagenomic sequencing is becoming more and more widespread in environmental and biomedical research, and the pace is increasing even more 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 reagents, laboratory and host related contaminants which demands a reliable method for the removal of negative control taxa from the rest of samples. This is particularly critical in low biomass body sites and environments, where contamination can comprise most of a sample if not all. With Recentrifuge, researchers can analyze results from Centrifuge and LMAT classifiers using interactive hierarchical pie charts with special emphasis on the confidence level of the classifications. In addition to control-subtracted samples, shared and exclusive taxa per sample are provided with scores, thus enabling robust comparative analysis of multiple samples in any metagenomic study.

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

Introduction

Studies of microbial communities by metagenomics are becoming more and more popular in different biological arenas, like environmental, clinical, food and forensic studies (*1*–*3*). These works are boosted by new DNA and RNA sequencing technologies that 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 (*4*). 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

PREPRINT

bioRxiv pre-print LaTeX v0.3

extracted from each sample using a commercial kit or an optimized custom 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.

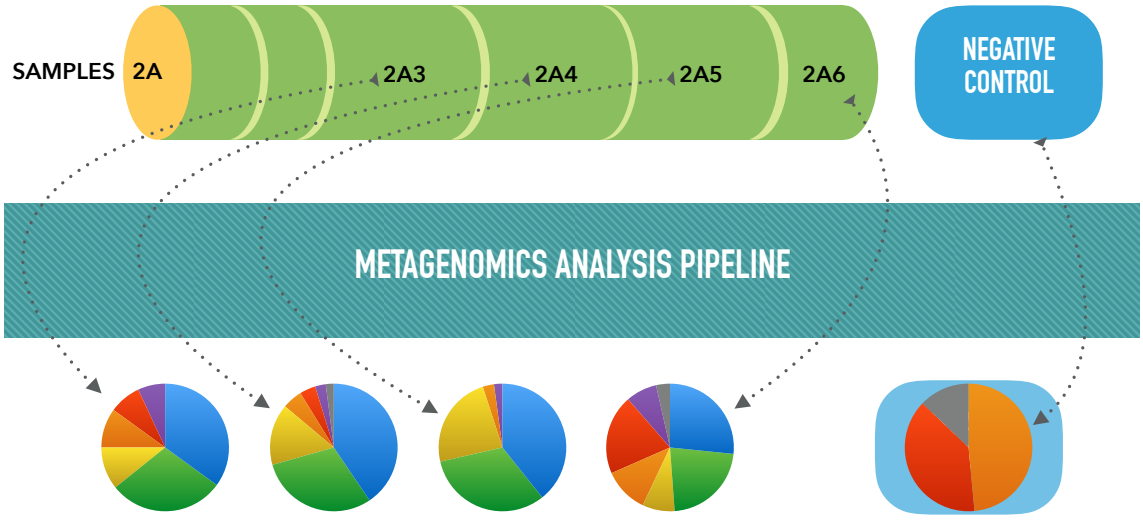

Figure 1. This is an example outlining the problem of comparing different but related samples in a SMS study. The sample named 2A is subdivided longitudinally in six subsamples whose DNA is extracted along with a negative control sample. The purified DNA is then sequenced and the generated sequencing reads are processed through a metagenomics analysis pipeline (as the one detailed in Figure 2). A collection of 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 native microbial DNA; the library preparation and sequencing methods will return sequences whose main source is contamination (*12*, *13*). From the data science perspective, this is just another instance of the importance of keeping a good *signal to noise ratio* (*14*), otherwise, when the *signal* (inherent DNA, target of the sampling) approaches the order of magnitude of the *noise* (acquired DNA from contamination), special methods are required to try to tell them apart.

The roots of contaminating sequences are diverse, as they can be traced back to nucleic acid extraction kits (the *kitome*) (*15*), reagents (*16*), the host (*17*) and the post-sampling environment (*18*), where contamination arises from different origins such as airborne particles,

bioRxiv pre-print LaTeX v0.3

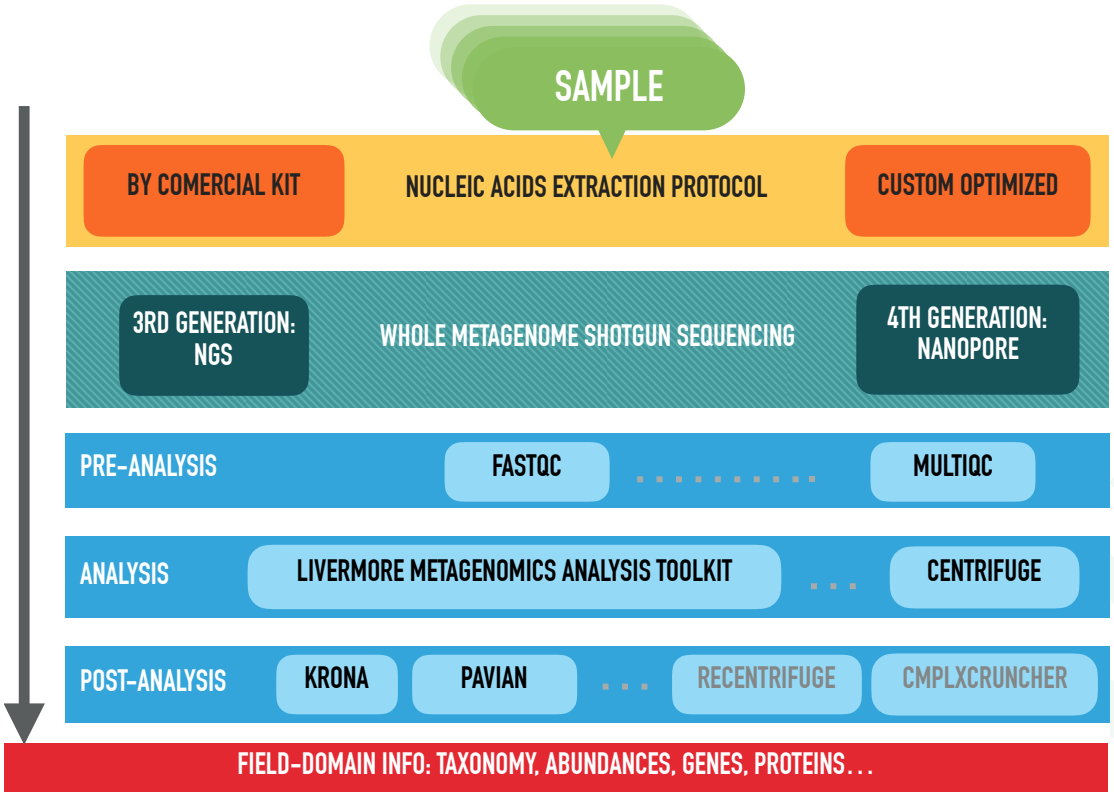

Figure 2. A study involving SMS, like the one illustrated by Figure 1, spans in a number of stages to extract valuable field-domain information starting from the original samples. For each sample, the DNA is extracted using a commercial kit, a custom protocol optimized for the type of sample or a combination 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 (*5*); in the analysis stage, the most computational intensive, the reads are taxonomically or functionally classified using software packages like LMAT (*6*) and Centrifuge (*7*) (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 (*8*), Pavian (*9*) and others.

crossover between current samples, or DNA left behind from previous sequencing experiments (*19*). Variable amounts of DNA from these sources are sequenced along with native microbial DNA, which could lead to severe bias in coverage and quantification of microbiota, especially in low biomass body sites and environments (*20*).

Regarding the *kitome*, it 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

bioRxiv pre-print LaTeX v0.3

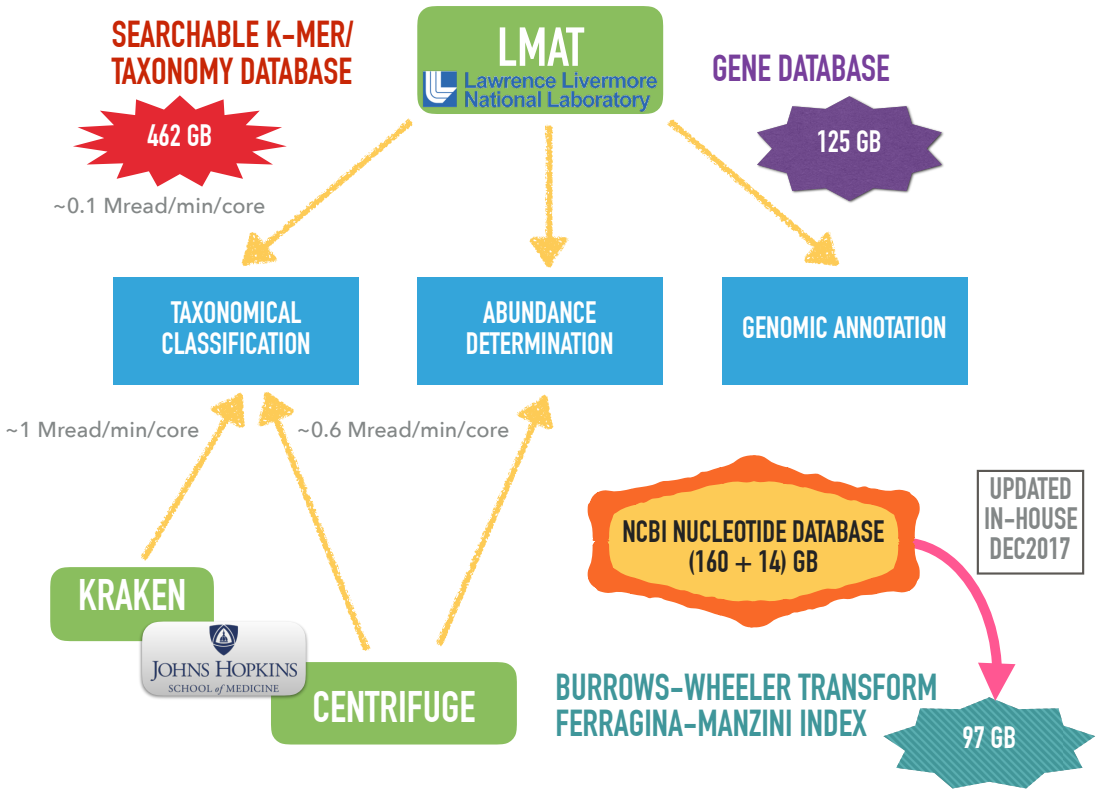

Figure 3. The core phase of a metagenomics analysis pipeline (see Figure 1 and 2 for the outline of the bioinformatics phases) is carried out by high performance computing software. These are intensive codes in both CPU and memory (sometimes, they are input/output intensive too), like LMAT (*6*), Kraken (*10*) and, more recently, Centrifuge (*7*). 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 December 2017 from the NCBI Nucleotide (*11*) *nt* database (160 GB plus 14 GB in indexes) is occupying just 97 GB.

used, including Earth Microbiome Project and Human Microbiome Project, usually yields a background contamination by no means negligible (*13*). The lower the biomass in the samples, the more essential it is to collect negative control samples to help in the contamination background assessment, because without them it would be almost impossible to distinguish inherent microbiota in a sample (signal) from contamination (noise).

Once the native DNA from the samples has been told from the contaminating DNA, the prob-

bioRxiv pre-print LaTeX v0.3

lem of performing a reliable comparison between samples remains. The sequence reads are generally assigned to different taxonomic ranks, especially when using a more conservative approach like Lowest Common Ancestor (LCA) (*6*). While LCA drastically reduces the risk of false positives, it usually spreads the taxonomic level of the classifications from the more specific to the more general. Even if LCA is not used, each read is assigned a particular score or confidence level by the classifier, which should be taken into account by any downstream application as a reliability estimator of the classification.

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 (*20*). A closely related problem arises in bioforensic studies and environmental surveillance, where it is essential to be able to reliably detect the presence of a particular organism in metagenomic samples even if the organism is only present in a trace amount (*21*, *22*).

Recentrifuge enables researchers to analyze results from the Centrifuge and the LMAT classifiers using Krona-like charts (*8*), 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 in metagenomics studies, especially useful in cases of low biomass environments or body sites.

PREPRINT

bioRxiv pre-print LaTeX v0.3

Methods

The visualization part of Recentrifuge is based on interactive hierarchical pie charts. The Krona (8) 2.0 JavaScript (JS) library developed at the Battelle National Biodefense Institute was adapted and extended to fit the particular needs of the problem.

The computing kernel of Recentrifuge is a parallel code written in Python following a Object-Oriented paradigm. For each sample a logical taxonomic tree is populated, with the leafs usually belonging to the lower taxonomic levels, like species. The methods involving trees were implemented as recursive functions for compactness and robustness, making the code less error-prone. For example, there is a recursive method that "folds the tree" for any sample if the number of assigned reads to a taxon is under the 'mintaxa' (minimum reads assigned to a taxon to exist in its own) setting, or because the taxonomical level of interest is over the assigned taxid (taxonomical identifier). This method does not just "prune the tree", on the contrary, it accumulates the counts of a taxon in the parent one and recalculates the parent score as a weighted average taking into account the counts and score of both. This is done recursively until the desired conditions are met. This method is applied, at a given taxonomic level, to the trees of every sample before being compared in search for the shared and exclusive taxa. Tuning the 'mintaxa' parameter is key to trim crossover between samples (it is set to 10 by default, as in LMAT).

In general, in addition to the input samples, Recentrifuge includes some derived samples in its output. After parallel calculations for each taxonomic level of interest, it includes hierarchical pie plots for `CTRL` (control subtracted samples), but also for `EXCLUSIVE`, `SHARED` and `SHARED_CONTROL` samples, defined below.

Let $\mathbb{T}$ mean the set of taxids in the NCBI Taxonomy (23) and T_s the collection of taxids present in a sample s. If R_s stands for the set of reads of a sample s and C_s for the group of them classifiable, then the taxonomic classification c is a function from C_s to $\mathbb{T}$, i.e., $C_s \xrightarrow{c} \mathbb{T}$, where $C_s \subseteq R_s$ and $c[C_s] = T_s \subseteq \mathbb{T}$. The set L of the 30 different taxonomic levels used in the NCBI is ordered in accordance with the taxonomy, so $(L, <)$ is a strictly ordered set, since $forma < varietas < subspecies < \cdots < domain$. Then, $T_s = T_s^{forma} \bigcup \cdots \bigcup T_s^{domain} = \bigcup^L T_s^l$, where T_s^l represents the collection of taxa belonging to a sample

bioRxiv pre-print LaTeX v0.3

s for a particular taxonomic rank or level l.

For a taxonomic rank k of interest, in a series of m samples where the negative control is the sample 0, the sets of taxa in the derived samples CTRL $\left({}^{\texttt{CTRL}}T_s^k\right)$, EXCLUSIVE $\left({}^{\texttt{EXCL}}T_s^k\right)$, SHARED $\left({}^{\texttt{SHARED}}T^k\right)$ and SHARED_CONTROL $\left({}^{\texttt{SHRD_CTRL}}T^k\right)$ are computed by Recentrifuge as follows:

$$
\begin{aligned}
{}^{\texttt{CTRL}}T_s^k &= \bigcup_{l\geq k}^{L} T_s^l - T_0^k \\
{}^{\texttt{EXCLUSIVE}}T_s^k &= T_s^k - \left(\bigcup_{n\neq s}^{m} T_n^k\right) \\
{}^{\texttt{SHARED}}T^k &= \bigcap_{n=0}^{m} T_n^k \\
{}^{\texttt{SHARED_CONTROL}}T^k &= \bigcap_{n>0}^{m} T_n^k - T_0^k
\end{aligned}
$$

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

1. 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.

2. 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 (*23*).

3. It is a full statically annotated Python 3.6 code. PEP-8 (*24*), PEP-484 (*25*) and PEP-526 (*26*) compliant. Written following the Google python style guide (*27*). Pylint and mypy checked.

4. Implemented under a object-oriented paradigm to ease future extensions targeting new or improved uses; it is prepared for additional input formats and other taxonomies different from NCBI —by direct support extending the base class or indirectly by using a mapping software like CrossClassify (*28*)—. A summarized UML (Unified Modeling Language) (*29*) class diagram of Recentrifuge is shown in Figure 4.

bioRxiv pre-print LaTeX v0.3

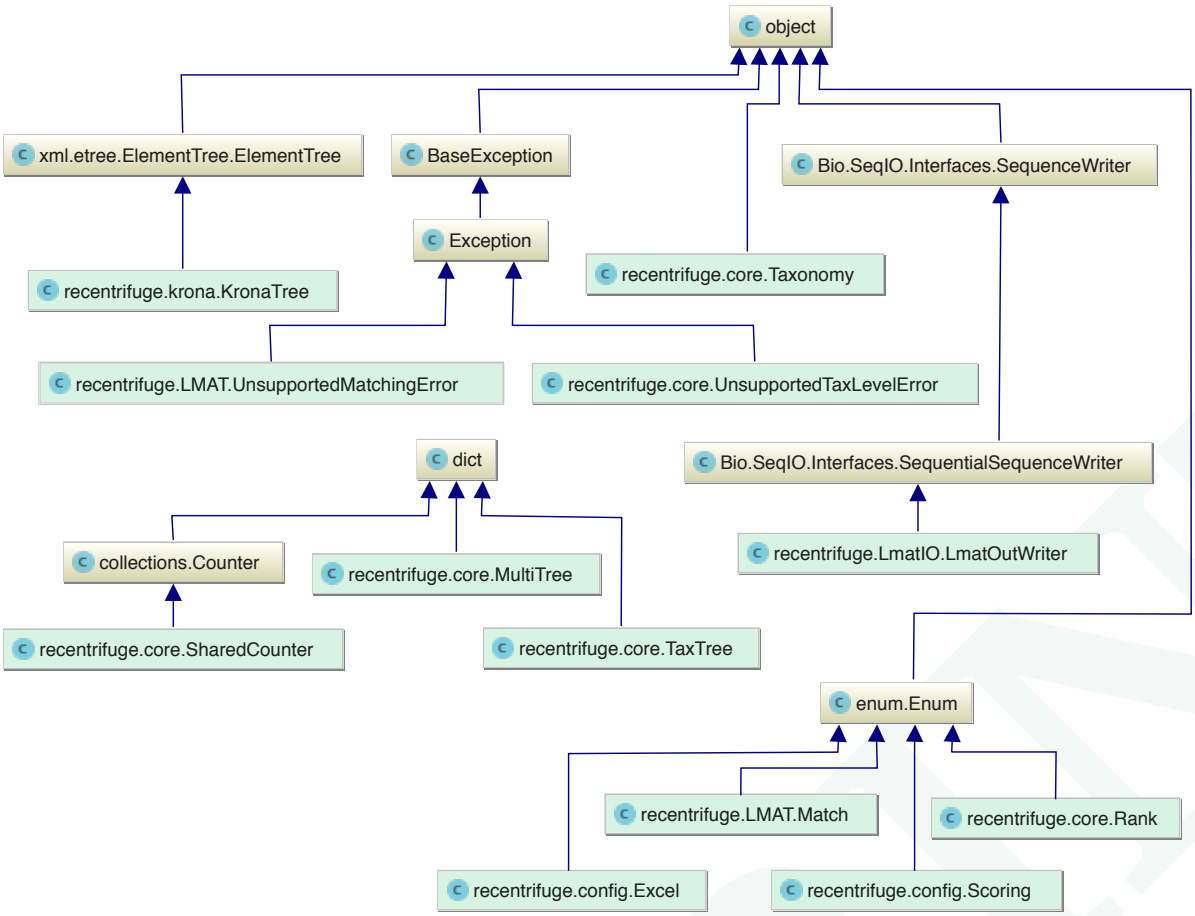

Figure 4. UML class diagram summarizing relationships between classes in the Recentrifuge package. The green background denotes developed classes currently used in the package. The rest of classes in the figure, the parent classes from which the Recentrifuge ones are derived, belong to the Python Standard Library and BioPython.

Results

Recentrifuge is a metagenomics analysis software with two different main parts: the computing kernel, implemented and parallelized from scratch using Python, and the interactive interface, written in JavaScript as an extension of the Krona (*8*) JS library to take full advantage of the classifications confidence level. Figure 5 summarizes the package context and data flows. Recentrifuge straightforwardly accepts both Centrifuge (*7*) and LMAT (*6*) output files, thus enabling scored oriented analysis and visualization for these tools. LMAT plasmids

bioRxiv pre-print LaTeX v0.3

assignment system (17) is also supported by Recentrifuge.

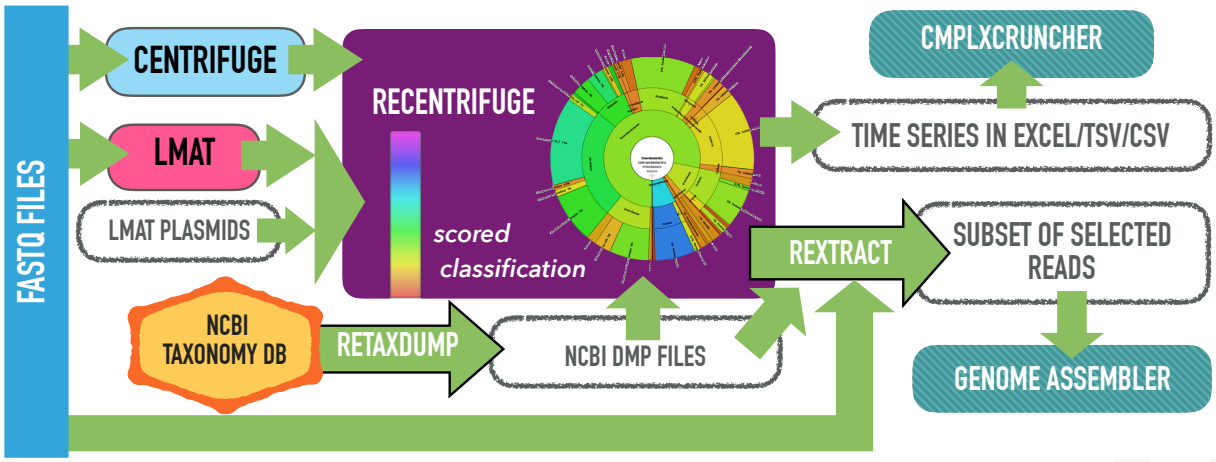

Figure 5. Outline of the Recentrifuge package with its context and main data flows. The green arrows show fully implemented pipelines. Recentrifuge accepts both Centrifuge (7) and LMAT (6) direct output files, thus enabling a scored oriented visualization for any of these codes. Recentrifuge is also supporting LMAT plasmids assignment system (17). The additional output of Recentrifuge to different text field formats enable further longitudinal (time or space) series analysis, for example, using *cmplxCruncher* (in development). The NCBI Taxonomy (23) dump databases are easily retrieved using *Retaxdump* script. *Rextract* utility extracts a subset of reads of interest from single or paired-ends FASTQ input files, which can be used in any downstream application, like genome assembling and visualization.

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. Another Recentrifuge output is an Excel spreadsheet detailing all the classification results in a compact way. This format is useful for further data mining on the data, for example, as input for applications such as longitudinal (time or space) series analyzers like *cmplxCruncher* (in development). In addition, the user can choose between different score visualization algorithms, some of which are more interesting for datasets containing variable length reads, for example, the ones generated by Oxford Nanopore sequencers. Finally, some filters are available, like the minimum score threshold. This filter can be used to generate different output sets from a single run of the classifier with a low minimum hit length (mhl) setting. That would not only save time and computational resources, but also would make

bioRxiv pre-print LaTeX v0.3

compatible, on the one hand, the advantage of a low mhl in low quality reads with, on the other hand, the benefit of a high mhl to avoid false positives.

Further 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 single or paired-ends FASTQ input files. These reads can be used in any downstream application, such as genome visualization and assembling.

Recentrifuge is currently available in Github (https://github.com/khyox/recentrifuge) under different free software licenses. The wiki (https://github.com/khyox/recentrifuge/wiki) is the most extensive and updated source of documentation for Recentrifuge.

Discussion

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

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

From the beginning of the application of SMS to environmental samples (*30*, *31*) in order to provide biologists with insight of microbial communities not obtainable from BAC clone–sequencing surveys or 16S rRNA, the need and challenges of comparative metagenomics were underlined (*32*, *33*). MEGAN (*34*), one of the first metagenomic data analysis tools, provided in its initial release a very basic comparative of samples that has improved with an interactive approach in more recent versions (*35*). In general, metagenomic classification and assembly software is more intra- than inter-sample oriented (*36*). Several tools have tried to fill the gap, starting with CoMet (*37*) in 2011, a web based tool for compara-

bioRxiv pre-print LaTeX v0.3

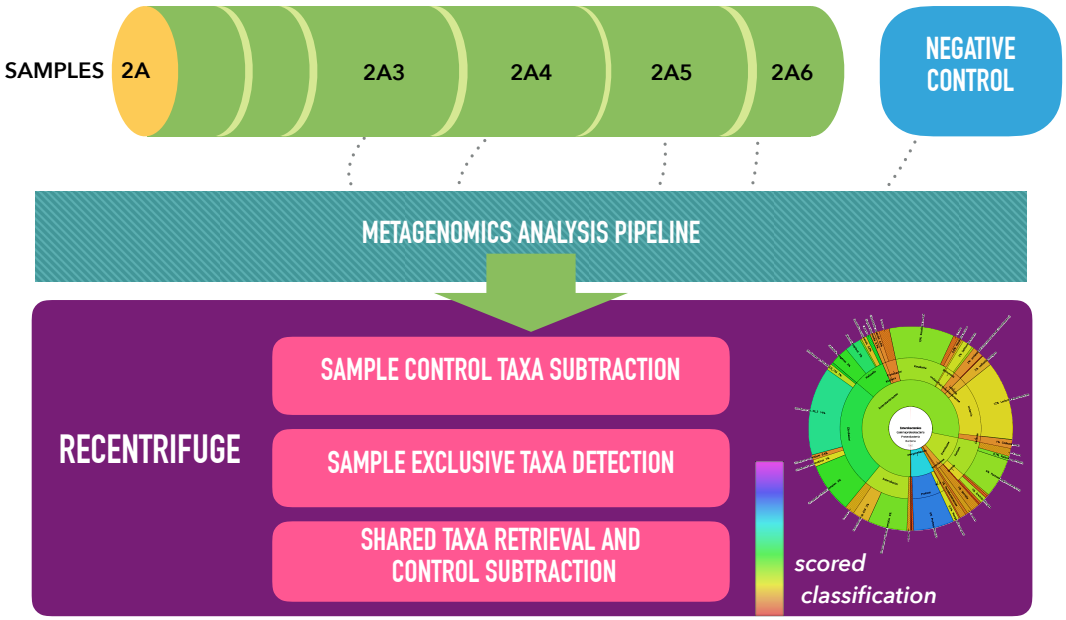

Figure 6. Summary of immediate advantages of applying Recentrifuge to the example of Figure 1, a study involving SMS of different but related samples, including a negative control. Four different sets of scored charts are generated for each taxonomic level of interest in addition to the scored plots 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 metagenomic studies.

tive functional profiling of metagenomes. In 2012, a different approach was presented with the discovery of the crAssphage thanks to the crAss software (*38*), which provides reference-independent comparative metagenomics using cross-assembly. The following year, a new tool for visually comparing microbial community structure across microbiomes, Community-analyzer (*39*) appeared. In 2014 it is released COMMET (*40*), a tool that goes a step further by enabling the combination of multiple metagenomic datasets. A package for direct comparative metagenomics by using k-mers (*41*) is published in 2016.

In 2015, a highly publicized report on the metagenomics of the New York subway suggested that the plague and anthrax pathogens were part of the normal subway microbiome. Soon afterwards, it was widely criticized (*42*) and, later, it was proved that reanalysis of the New York subway data with appropriate methods did not detect the pathogens (*43*). As a conse-

bioRxiv pre-print LaTeX v0.3

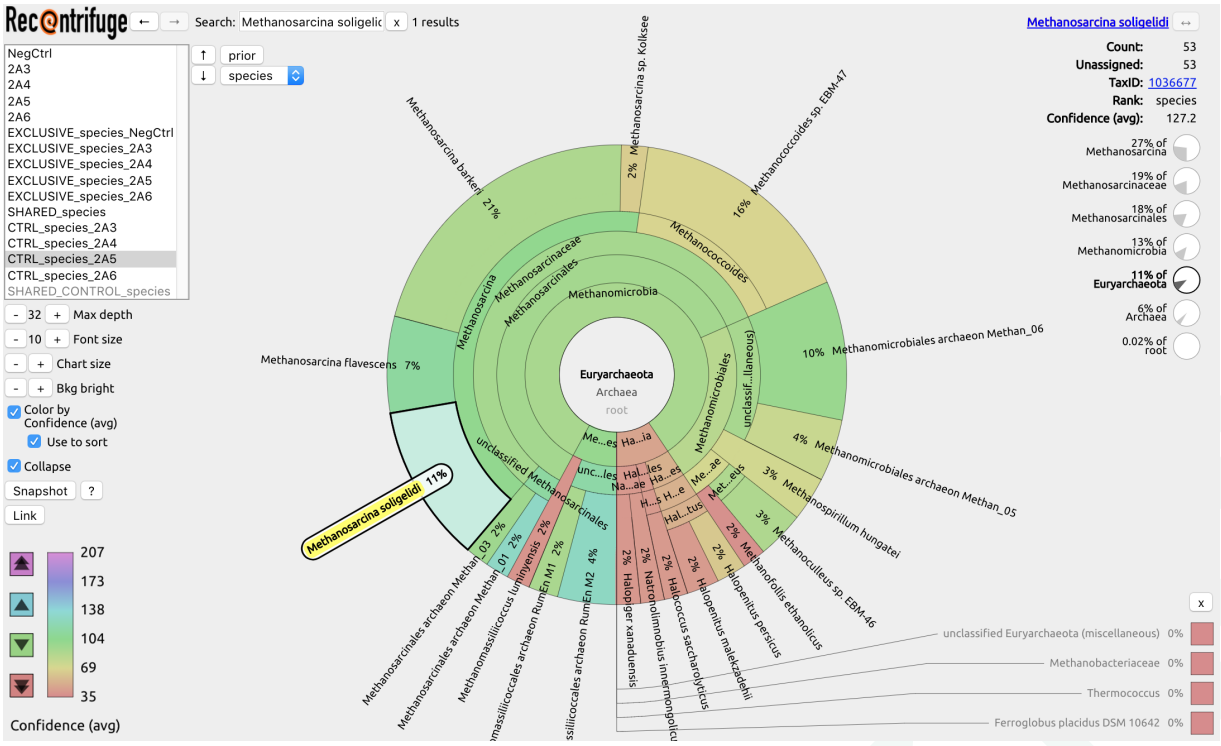

Figure 7. Screenshot of the Recentrifuge HTML interface in the SMS example study of Figure 1 with results summarized in Figure 6. The sample 2A6 after control removal at species level is selected (see the sample selection box in the top left under the Recentrifuge logo). In the center, the corresponding hierarchical pie chart is drawn with zoom in the phylum Euryarchaeota. 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, where the buttons for the taxa score navigation are also found. The sort of the taxa is also done attending to the average confidence level, as that option is enabled in the plot. In this particular case, the taxon *Methanosarcina soligelidi* 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 (*23*), while the statistics include the number of reads assigned to this or lower taxonomic levels —Counts— and their average confidence —Confidence (avg)—, the number of reads just assigned to this level —Unassigned—, the NCBI taxid —TaxID— and rank —Rank—, and some relative frequencies data.

quence of this and other similar problems involving metagenomic studies, a work directed by Rob Knight (*44*) emphasized the importance of validation in metagenomic results and issued a tool based on BLAST (Platypus Conquistador) that confirms the presence or absence of a taxon of interest within shotgun metagenomic datasets by relying on two reference sequence databases: one for inclusions, with the sequences of interest, and the other for exclusions, with any known sequence background. Another BLAST-based method for validating the as-

bioRxiv pre-print LaTeX v0.3

signments made by less precise sequence classification programs has been just announced (3).

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 or LMAT 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 especially useful in case of notable variations in the length of the reads, like in the datasets generated by nanopore sequencers.

Only a few codes, such as Krona (8) and MetaTreeMap (45), are hitherto able to handle a score assigned to the classification nodes. In Recentrifuge, the calculated score is kept in all the downstream analysis and comparisons. That is an essential advantage of Recentrifuge over other metagenomic dataset analysis tools. It provides with an interactive framework for an easy assessment of the validity of the taxonomic assignments.

By enabling score-oriented robust comparative analysis of multiple samples in metagenomic studies, Recentrifuge could play a key role not only in the analysis of oligotrophic microbes in environmental samples, but also in a more reliable detection of minority organisms in clinical or forensic samples.

Acknowledgments

I would like to thank Jordi Burguet-Castell for critical comments on the manuscript. I would also like to thank Carlos P. Garay and Jose Fco. Martí for valuable conversations about Recentrifuge. Finally, I would like to thank Jonathan E. Allen for his hospitality during my stay in his lab at Lawrence Livermore National Laboratory in 2015, a visit then that has been essential now to be able to integrate LMAT with Recentrifuge.

PREPRINT

bioRxiv pre-print LaTeX v0.3

Funding Information

No funding applies to this work.

Competing interests

The author declares no competing interests.

References

1. **Miller RR, Montoya V, Gardy JL, Patrick DM, Tang P**. 2013. Metagenomics for pathogen detection in public health. Genome medicine **9**:81.
2. **Ercolini D**. 2013. High-throughput sequencing and metagenomics: moving forward in the culture-independent analysis of food microbial ecology. Applied and environmental microbiology **10**:3148-3155.
3. **Fricke WF, Cebula TA, Ravel J**. 2011. Genomics. In *Microbial Forensics* (Second Edition), Academic Press, ISBN 9780123820068 (pp. 479-492).
4. **Edwards et. al**. 2016. Extreme metagenomics using nanopore sequencing: a field report from Svalbard, 78°N, BioRxiv **073965**.
5. **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.
6. **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.

PREPRINT

bioRxiv pre-print LaTeX v0.3

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

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

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

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

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

12. **Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R**. 2014. Tracking down the sources of experimental contamination in microbiome studies. Genome Biology. **12**:564.

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

14. **Skolnik MI**. 2001. Introduction to Radar Systems, 3rd ed (Int'l Ed). McGraw-Hill, Electrical Engineering Series. ISBN 0-07-118189-X.

15. **Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ and Walker AW**. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology **12**:87.

16. **Olm MR, Butterfield CN, Copeland A, Boles TC, Thomas BC, Banfield JF**. 2017. The Source and Evolutionary History of a Microbial Contaminant Identified Through Soil Metagenomic Analysis. mBio **8**:e01969-16.

PREPRINT

bioRxiv pre-print LaTeX v0.3

17. **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.

18. **Lusk RW**. 2014. Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data. PLoS ONE, **9**:10.

19. **Gruber K**. 2015. Here, there, and everywhere. EMBO reports, 16(**8**):898-901.

20. **Nayfach S, Pollard KS**. 2016. Toward Accurate and Quantitative Comparative Metagenomics. Cell. 166(**5**):1103-1116.

21. **Bazinet AL, Ondov BD, Sommer DD, Ratnayake S**. 2017. BLAST-based validation of metagenomic sequence assignments. BioRxiv **181636**.

22. **Doggett NA, Mukundan H, Lefkowitz EJ, Slezak TR, Chain PS, Morse S, Anderson K, Hodge DR, Pillai S**. 2016. Culture-independent diagnostics for health security. Health security, 14(**3**):122-142.

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

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

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

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

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

bioRxiv pre-print LaTeX v0.3

28. **Balvočiūtė M, Huson DH**. 2017. SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare?. BMC genomics **2**:114.

29. 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.

30. **Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF**. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **6978**:37-43.

31. **Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE**. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science, **5667**:66-74.

32. **Schloss PD, Handelsman J**. 2005. Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. Genome biology, **8**:229.

33. **Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P**. 2005. Comparative metagenomics of microbial communities. Science **5721**:554-557.

34. **Huson DH, Auch AF, Qi J, Schuster SC**. 2007. MEGAN analysis of metagenomic data. Genome research **3**:377-386.

35. **Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh HJ, Tappu R**. 2016. MEGAN community edition-interactive exploration and analysis of large-scale microbiome sequencing data. PLoS computational biology, **6**:e1004957.

36. **Breitwieser FP, Lu J, Salzberg SL**. 2017. A review of methods and databases for metagenomic classification and assembly. Briefings in bioinformatics, 2017:1-15.

bioRxiv pre-print LaTeX v0.3

37. **Lingner T, Aßhauer KP, Schreiber F, Meinicke P**. 2011. CoMet—a web server for comparative functional profiling of metagenomes. Nucleic Acids Research, **39**:W518–W523.

38. **Dutilh BE, Schmieder R, Nulton J, Felts B, Salamon P, Edwards RA, Mokili JL**. 2012. Reference-independent comparative metagenomics using cross-assembly: crAss. Bioinformatics, **24**:3225-3231.

39. **Kuntal BK, Ghosh TS, Mande SS**. 2013. Community-analyzer: a platform for visualizing and comparing microbial community structure across microbiomes. Genomics **4**:409-418.

40. **Maillet N, Collet G, Vannier T, Lavenier D, Peterlongo P**. 2014. COMMET: comparing and combining multiple metagenomic datasets. IEEE BIBM 2014, Nov 2014, Belfast, United Kingdom. (pp. 94-98)

41. **Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, Lemaitre C**. 2016. Multiple comparative metagenomics using multiset k-mer counting. PeerJ Computer Science **2**:e94.

42. **Ackelsberg J, Rakeman J, Hughes S, Petersen J, Mead P, Schriefer M, Kingry L, Hoffmaster A, Gee JE**. 2015. Lack of evidence for plague or anthrax on the New York City subway. Cell systems **1**:4-5.

43. **Hsu T, Joice R, Vallarino J, Abu-Ali G, Hartmann EM, Shafquat A, DuLong C, Baranowski C, Gevers D, Green JL, Morgan XC**. 2016. Urban transit system microbial communities differ by surface type and interaction with humans and the environment. mSystems **3**:e00018-16.

44. **Gonzalez A, Vázquez-Baeza Y, Pettengill JB, Ottesen A, McDonald D, Knight R**. 2016. Avoiding Pandemic Fears in the Subway and Conquering the Platypus. mSystems **3**:e00050-16.

PREPRINT

bioRxiv pre-print LaTeX v0.3

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

PREPRINT