Title:

Recentrifuge: robust comparative analysis and contamination removal for metagenomics

Running title:

Recentrifuge

Jose Manuel Martí[1,#]

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

January 19, 2019

Abstract

Metagenomic sequencing is becoming widespread in biomedical and environmental 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 specimen and between specimens arises. Reagents, laboratory, and host related contaminants complicate such analysis. Contamination is particularly critical in low microbial biomass body sites and environments, where it can comprise most of a sample if not all. Recentrifuge implements a robust method for the removal of negative-control and crossover taxa from the rest of samples. With Recentrifuge, researchers can analyze results from taxonomic classifiers using interactive charts with emphasis on the confidence level of the classifications. In addition to contamination-subtracted samples, Recentrifuge provides shared and exclusive taxa per sample, thus enabling robust contamination removal and comparative analysis in environmental and clinical metagenomics.

Keywords— metagenomics, robustness, comparative analysis, contamination removal, low microbial biomass

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

Introduction

Studies of microbial communities by metagenomics are becoming more popular in different biological arenas, like environmental, clinical, food and forensic studies (Miller et al. 2013; Ercolini 2013; Fricke et al. 2011). (...). With the development of nanopore sequencing, portable and affordable real-time SMS is a reality (Edwards et al. 2016).

In the case of low microbial biomass samples, there is very little native DNA from microbes; the library preparation and sequencing methods will return sequences whose principal source is contamination (Weiss et al. 2014; Kim, Hofstaedter, et al. 2017). Sequencing of RNA requiring additional steps introduces still further biases and artifacts (Kim, Song, et al. 2016), which in case of low microbial biomass studies translates into a severe problem of contamination and spurious taxa detection (Perlejewski et al. 2016). The clinical metagenomics community is stressing the importance of negative controls in metagenomics workflows and, recently, raised a fundamental concern about how to subtract the contaminants from the results (Ruppé and Schrenzel 2018). (...)

All these tools are performing taxonomic classification and abundance estimation, whereas LMAT (Ames, Hysom, et al. 2013; Ames, Gardner, et al. 2015) is also able to annotate genes. For the taxonomic classification, both LMAT and Kraken (Wood and Salzberg 2014) use an exact k-mer matching algorithm with large databases (∼100 GiB) while Centrifuge (Kim, Song, et al. 2016) use compression algorithms to reduce the databases size (∼10 GiB) but at some speed expense. CLARK-S (Ounit and Lonardi 2016) use discriminative spaced k-mers to improve the sensitivity but with a toll on the performance.

Methods

Robust contamination removal

For a taxonomic rank k, after the 'tree folding' procedure detailed above, the contamination removal algorithm retrieves the set of candidates $\bar{T}_s^{\to k}$ to contaminant taxa from the $N < S$ control samples. Depending on the relative frequency $(f_i = n_i / \sum_i n_i)$ ...

Other classes of BibLaTeX references are *inproceedings* with COMMET (Maillet et al. 2014), *online* with entrez (Burguet-Castell and Martí 2015), and *misc* with PEP-484 (Rosum et al. 2015).

Derived samples

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

$$
\begin{aligned}
{}^{\texttt{CTRL}}T_s^k &= T_s^{\to k} \setminus \bigcup_n^N T_n^{\to k} \\
{}^{\texttt{EXCLUSIVE}}T_s^k &= T_s^{\to k} \setminus \bigcup_{m \neq s}^S T_m^{\to k} \\
{}^{\texttt{SHARED}}T^k &= \bigcap_m^S T_m^{\to k} \\
{}^{\texttt{SHARED\_CONTROL}}T^k &= \bigcap_{m>N}^S T_m^{\to k} \setminus \bigcup_n^N T_n^{\to k}
\end{aligned}
$$

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 (Ondov et al. 2011) JavaScript library to take full advantage of the classification confidence level. (...)

Discussion

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

Data access

Recentrifuge's main website is www.recentrifuge.org. The data and source code are anonymously and freely available on GitHub at https://github.com/khyox/recentrifuge and PyPI at https://pypi.org/project/recentrifuge. The Recentrifuge computing code is licensed under the

GNU Affero General Public License Version 3 (www.gnu.org/licenses/agpl.html).

Recentrifuge's continuous integration (CI) information is public on Travis CI at the website https://travis-ci.org/khyox/recentrifuge.

The wiki (https://github.com/khyox/recentrifuge/wiki) is the most extensive and updated source of documentation for Recentrifuge, including installation, testing, quick-start, and comprehensive use cases for the different taxonomic classification engines supported.

Acknowledgments

I would like to thank ...

References

Ames SK, Gardner SN, Marti JM, Slezak TR, Gokhale MB, and Allen JE. 2015. Using populations of human and microbial genomes for organism detection in metagenomes. *Genome research*. **25**: 1056–1067.

Ames SK, Hysom DA, Gardner SN, Lloyd GS, Gokhale MB, and Allen JE. 2013. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics (Oxford, England)*. **29**: 2253–2260.

Burguet-Castell J and Martí JM 2015. Entrez: Call the NCBI E-utilities from Python. Version 2018. URL: https://github.com/jordibc/entrez.

Edwards A, Debbonaire AR, Sattler B, Mur LA, and Hodson AJ. 2016. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. *BioRxiv*. DOI: 10.1101/073965.

Ercolini D. 2013. High-Throughput Sequencing and Metagenomics: Moving Forward in the Culture-Independent Analysis of Food Microbial Ecology. *Applied and Environmental Microbiology*. **79**: 3148–3155.

Fricke WF, Cebula TA, and Ravel J. 2011. Chapter 28 (Genomics). In *Microbial Forensics*. (Ed. by B Budowle, SE Schutzer, RG Breeze, PS Keim, and SA Morse), pp. 479–492. Academic Press, San Diego.

Kim D, Song L, Breitwieser FP, and Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*. **26**: 1721–1729.

Kim D, Hofstaedter CE, Zhao C, Mattei L, Tanes C, Clarke E, Lauder A, Sherrill-Mix S, Chehoud C, and Kelsen J. 2017. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome*. **5**: 52.

Maillet N, Collet G, Vannier T, Lavenier D, and Peterlongo P 2014. Commet: Comparing and combining multiple metagenomic datasets. English. In: *2014 IEEE Int Conf on BIBM*. IEEE, pp. 94–98.

Miller RR, Montoya V, Gardy JL, Patrick DM, and Tang P. 2013. Metagenomics for pathogen detection in public health. *Genome medicine*. **5**: 81.

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

Ounit R and Lonardi S. 2016. Higher classification sensitivity of short metagenomic reads with CLARK-S. *Bioinformatics*. **32**: 3823–3825.

Perlejewski K et al. 2016. Metagenomic Analysis of Cerebrospinal Fluid from Patients with Multiple Sclerosis. In *Pulmonary Infection and Inflammation*. (Ed. by M Pokorski). Chap. 25, pp. 89–98. Springer International Publishing, Switzerland.

Rosum GV, Lehtosalo J, and Langa L 2015. PEP-484: Type Hints. [Internet]. Python Software Foundation.

Ruppé E and Schrenzel J. 2018. Messages from the second International Conference on Clinical Metagenomics (ICCMg2). *Microbes and Infection*. **2018**: 1–6.

Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, and Knight R. 2014. Tracking down the sources of experimental contamination in microbiome studies. *Genome biology*. **15**: 564.

Wood DE and Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*. **15**: R46.