Article

Simka: large scale *de novo* comparative metagenomics

Gaëtan BENOIT[1], Pierre PETERLONGO[1], Mahendra MARIADASSOU[2], Erwan DREZEN[1,3], Sophie SCHBATH[2], Dominique LAVENIER[1] and Claire LEMAITRE[1]

[1] INRIA/IRISA, Genscale team, UMR6074 IRISA CNRS/INRIA/Université de Rennes 1, Campus de Beaulieu, 35042, Rennes, France

[2] MaIAGE, INRA, Université Paris-Saclay, 78350, Jouy-en-Josas, France

[3] CHU Pontchaillou, 35000, Rennes, France

Corresponding author: gaetan.benoit@inria.fr

Reference paper: **Benoit *et al.* (2016) Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.94**

Large metagenomic projects, such as Human Microbiome Project [1] or Tara Ocean Project [2], are becoming increasingly important for understanding life processes at the individual scale or at more global scales. Therefore, serious efforts are put in funding projects, collecting DNA, sequencing, and analyzing the resulting sequences. Comparative metagenomics is one of the most ubiquitous and informative of those analyzes. The purpose is mainly to estimate proximity (or distance) between two or more environmental sites at the genomic level. The comparisons are often based on identified species content. However, this approach is limited to sequences correctly assigned to known species documented in public biobanks, and this may correspond to small fractions of the datasets, in particular for environmental samples such as sea water. This limitation motivated the development of *de novo* comparison tools, such as Compareads [3] or Mash [4], based only on the non assembled read set comparisons. Compareads was for instance successfully used for the first analyses of the Tara Ocean data [5].

These reference-free methods share the use of k-mers as the fundamental unit used for comparing samples. Actually, k-mers are a natural unit for comparing communities: (1) sufficiently long k-mers are usually specific of a genome [6], (2) k-mer frequency is linearly related to genome's abundance [7], (3) k-mer aggregates organisms with very similar k-mer composition (*e.g.* related strains from the same bacterial species) without need for a classification of those organisms [8]. However, even if Compareads approach was designed to scale-up to large metagenomic read sets, its use on data generated by large scale projects is turning into a bottleneck in terms of time requirements. By contrast, Mash outperforms by far all other methods in terms of computational resource usage. However, this frugality comes at the expense of result quality and precision: the output distances and Jaccard indexes do not take into account relative abundance information and are not computed exactly due to k-mer sub-sampling. This is what motivated this work in which we propose a new *de novo* comparative metagenomic method, called Simka. Simka compares N metagenomic datasets based on their exact k-mers counts. It computes a large collection of distances classically used in ecology to compare communities, by replacing species counts by k-mer counts, for a large range of k-mer sizes, including large ones (up to 30). Simka is, to our knowledge, the first method able to rapidly compute a full range of distances enabling the comparison of any number of datasets. Simka outperforms state-of-the-art read comparison methods in terms of computational needs and result quality. For instance, Simka ran on 690 samples from the Human Microbiome Project (HMP) (totalling 32 billion reads) in less than 10 hours and using no more than 70 GB RAM.

Simka works as follows. Firstly, the k-mer spectrum of each dataset is computed. The k-mer spectrum of a dataset is the set of all its distinct k-mers associated with their abundance in the dataset. Secondly, k-mer spectra are compared in a pairwise manner to compute their distance. This comparison process basically aims at identifying which k-mers are shared by both spectra and which ones are not. It can be computationally very expensive because each k-mer spectrum can contains millions to billions of distinct k-mers when k is large (> 15). Moreover, the number of comparisons grows quadratically with the number of input datasets. To tackle this issue, we have designed a new k-mer counting strategy of numerous datasets, called Multiset k-mer Counting (MKC). MKC takes N datasets as input and provides an abundance vector for each distinct k-mer. The abundance vector of a k-mer consists of its N counts in the N datasets. The abundance vector generation by the MKC task is divided into two phases: (1) Sorting Count, (2) Merging Count. During the first step, the k-mers of each dataset are counted independently. This is performed by sorting the k-mers in lexicographical order. Distinct k-mers can thus be identified and their number of occurrences computed. This task can be very efficiently performed by popular disk-based k-mer counting tool such as DSK [9] or KMC2 [10]. The resulting k-mer spectra are written on the disk. During the second step, a Merge-Sort algorithm can be efficiently

applied on the sorted k-mer spectra to directly generate abundance vectors. Given those abundance vectors, the distances between each pair of datasets can be computed simultaneously. Interestingly, most of the ecological distances are additive over the distinct k-mers, meaning that they can be iteratively updated one abundance vector at a time. Once an abundance vector has been processed, there is thus no need to keep it on record, allowing Simka to have a very low memory footprint.

One advantage of the overall Simka workflow is its high parallelism potential. During the sorting count phase of the MKC, a first parallelism level is given by the independent counts of each dataset. N processes can thus be run in parallel, each one dealing with a specific dataset. A second level is given by the fine grained parallelism implemented in software such as DSK or KMC2 that intensively exploit today multicore processor capabilities. As the number of distinct k-mers is generally huge, those tools separate the k-mers in P smaller disjoint sets that can be counted independently and thus result in P k-mer spectrum chunks per dataset. During its second step, the MKC exploit this partitionning to merge up to P k-mer spectrums chucks in parallel. Each of these merge processes generates abundance vectors from which independent contribution to the distances are computed. Since the distance computed by Simka are additive over the distinct k-mers, each contribution is simply accumulated and the final distance is computed. Simka implementation is based on the GATB library [11], a C++ library optimized to handle very large sets of k-mers. Simka is usable on standard computers and has also been entirely parallelized for grid infrastructures made of hundred of nodes, and where each node implements 8 or 16-core systems.

The quality of the distances computed by Simka were evaluated answering two questions. First, are they similar to distances between read sets computed using other *de novo* approaches? Second, do they recover the known biological structure of HMP samples? For the first evaluation, we show that Simka result are perfectly well correlated with Compareads results. We go further in this evaluation by showing that Simka results are highly correlated with costly but extremely accurate *de novo* comparison techniques relying on all-versus-all sequence alignment strategy. For the second evaluation, Simka distances were compared to taxonomic distances that are a traditional way of comparing metagenomic samples. Taxonomic distances are based on sequence assignation to taxons by mapping to reference databases. To compare Simka to such traditional reference-based methods, we used the HMP dataset. One advantage of this dataset is that it has been extensively studied, in particular the microbial communities are relatively well represented in reference databases [1,12]. We show that substituting k-mer counts by species counts gives admittedly different distances but that those distances are biologically relevant as they capture the same underlying biological structure and lead to the same conclusions as those based on taxonomic composition. In particular, Simka was able to retrieve two major biological results. The first one is the segregation of the HMP datasets by body sites. The second one reveals that the organisation of the gut samples is mainly driven by the relative abundances of three bacterial genera, known as enterotypes, and characterized by the relative abundances of a few genera: *Bacteroides*, *Prevotella* and genera from the *Ruminococcaceae* family. In contrast of Simka, Mash performed badly when considering HMP datasets per body site since this tool can only take into account presence/absence information and not relative abundances. As a matter of fact, differences in relative abundances are subtler signals that are often at the heart of interesting biological insights in comparative genomics studies [13,14,15,16,17].

We introduced Simka, a new method for computing a collection of ecological distances, between many large metagenomic datasets, based on their k-mer composition. This was made possible thanks to the Multiset k-mer Counting algorithm (MKC), a new strategy that counts k-mers of numerous datasets with state-of-the-art time, memory and disk performances.

Acknowledgements

This work was supported by the French ANR-14-CE23-0001 Hydrogen Project.

References

[1] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, 2012.

[2] Eric Karsenti, Silvia G. Acinas, Peer Bork, Chris Bowler, Colomban de Vargas, Jeroen Raes, Matthew Sullivan, Detlev Arendt, Francesca Benzoni, Jean Michel Claverie, Mick Follows, Gaby Gorsky, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Stefanie Kandels-Lewis, Uros Krzic, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Em-

manuel Georges Reynaud, Christian Sardet, Michael E. Sieracki, Sabrina Speich, Didier Velayoudon, Jean Weissenbach, and Patrick Wincker. A holistic approach to marine Eco-systems biology. *PLoS Biology*, 9, 2011.

[3] Nicolas Maillet, Claire Lemaitre, Rayan Chikhi, Dominique Lavenier, and Pierre Peterlongo. Compareads: comparing huge metagenomic experiments. *BMC bioinformatics*, 13(Suppl 19):S10, 2012.

[4] Brian D. Ondov, Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol*, 17(1):132, 2016.

[5] E. Villar, G. K. Farrant, M. Follows, L. Garczarek, S. Speich, S. Audic, L. Bittner, B. Blanke, J. R. Brum, C. Brunet, R. Casotti, A. Chase, J. R. Dolan, F. Ortenzio, J.-P. Gattuso, N. Grima, L. Guidi, C. N. Hill, O. Jahn, J.-L. Jamet, H. Le Goff, C. Lepoivre, S. Malviya, E. Pelletier, J.-B. Romagnan, S. Roux, S. Santini, E. Scalco, S. M. Schwenck, A. Tanaka, P. Testor, T. Vannier, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, S. G. Acinas, P. Bork, E. Boss, C. de Vargas, G. Gorsky, H. Ogata, S. Pesant, M. B. Sullivan, S. Sunagawa, P. Wincker, E. Karsenti, C. Bowler, F. Not, P. Hingamp, and D. Iudicone. Environmental characteristics of agulhas rings affect interocean plankton transport. *Science*, 348(6237):1261447–1261447, may 2015.

[6] Y. Fofanov, Y. Luo, C. Katili, J. Wang, Y. Belosludtsev, T. Powdrill, C. Belapurkar, V. Fofanov, T.-B. Li, S. Chumakov, and B. M. Pettitt. How independent are the appearances of n-mers in different genomes? *Bioinformatics*, 20(15):2421–2428, apr 2004.

[7] Yu-Wei Wu and Yuzhen Ye. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, 18(3):523–534, 2011.

[8] Hanno Teeling, Jost Waldmann, Thierry Lombardot, Margarete Bauer, and Frank O Glöckner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC bioinformatics*, 5(1):163, 2004.

[9] Guillaume Rizk, Dominique Lavenier, and Rayan Chikhi. DSK: k-mer counting with very low memory usage. *Bioinformatics*, page btt020, 2013.

[10] Sebastian Deorowicz, Marek Kokot, Szymon Grabowski, and Agnieszka Debudaj-Grabysz. KMC 2: Fast and resource-frugal k-mer counting. *Bioinformatics*, 31(10):1569–1576, 2015.

[11] Erwan Drezen, Guillaume Rizk, Rayan Chikhi, Charles Deltel, Claire Lemaitre, Pierre Peterlongo, and Dominique Lavenier. GATB: Genome assembly & analysis tool box. *Bioinformatics*, 30(20):2959–2961, 2014.

[12] Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, Jun 2012.

[13] Sébastien Boutin, Simon Y. Graeber, Michael Weitnauer, Jessica Panitz, Mirjam Stahl, Diana Clausznitzer, Lars Kaderali, Gisli Einarsson, Michael M. Tunney, J. Stuart Elborn, Marcus A. Mall, and Alexander H. Dalpke. Comparison of microbiomes from different niches of upper and lower airways in children and adolescents with cystic fibrosis. *PLoS ONE*, 10(1):1–19, 01 2015.

[14] A. Shade, S. E. Jones, J. G. Caporaso, J. Handelsman, R. Knight, N. Fierer, and J. A. Gilbert. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio*, 5(4):e01371–14–e01371–14, jul 2014.

[15] S. Genitsaris, S. Monchy, E. Viscogliosi, T. Sime-Ngando, S. Ferreira, and U. Christaki. Seasonal variations of marine protist community structure based on taxon-specific traits using the eastern english channel as a model coastal system. *FEMS Microbiology Ecology*, 91(5):fiv034–fiv034, mar 2015.

[16] Suzanne Coveley, Mostafa S Elshahed, and Noha H Youssef. Response of the rare biosphere to environmental stressors in a highly diverse ecosystem (zodletone spring, ok, usa). *PeerJ*, 3:e1182, 2015.

[17] V. Gomez-Alvarez, S. Pfaller, J. G. Pressman, D. G. Wahman, and R. P. Revetta. Resilience of microbial communities in a simulated drinking water distribution system subjected to disturbances: role of conditionally rare taxa and potential implications for antibiotic-resistant bacteria. *Environ. Sci.: Water Res. Technol.*, 2(4):645–657, 2016.