SRAnsack : metagenome signatures and assembled genomes for the exploration of SRA's aquatic data

Moritz Buck[1,*]Malhiheh?, Stefan?

[1]Swedish University of Agricultural Sciences, [*]corresponding author

The SRA (short read archive) is full of underused metagenomic data that is hard to put in context due to lacking metadata. Even for well characterised metagenomes, only selected assemblies and metagenome assembled genomes (MAGs) are made public. This makes it extremely difficult to put your own metagenomes in the right context of genomic and metagenomic diversity.

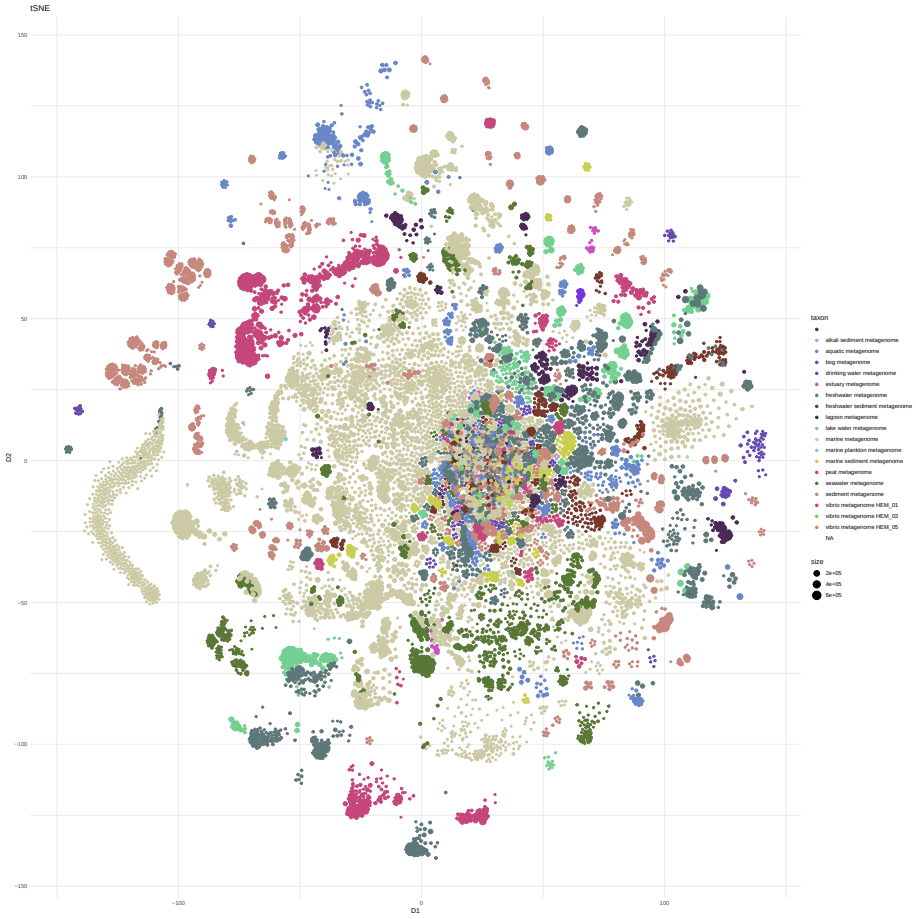
To remedy this, we have developed a simple snakemake-based pipeline that processes all aquatic metagenome[1] from SRA independently, equaly and reasonably efficiently, with the possibility to expand to other biomes at a later point. The pipeline produces data-collection: a collection of assemblies and MAGs, a collection of min-hash signatures and a collection of rarefied libraries.

Some of this data is then processed to some degree to make it more usable and allow for easier comparison with your own data. The MAG collection is clustered into metagenomic Operation Taxonomic Units (mOTUS), and taxonomically annotated. The signature collection is used to make an 'all vs all' comparison of the input SRA-libraries to compute similarity values, which allows for fast identification of highly similar samples. Additionally the rarefied libraries are screened for rRNA genes as many amplicon datasets are wrongly annotated as metagenome in SRA. We processed a total of 20.653 libraries, ~370Gb of assemblies, and 144.222 MAGs which cluster in ~27.000 metagenomic Operation Taxonomic Units (mOTUs).

This is but the first step of this growing dataset. The similarities in samples will be used in the future to compute a large collection of co-assemblies to expand the MAG collection to the more rare fraction of these samples and additionally the

[1]We call aquatic metagenomes any metagenomic dataset with their ncbi taxonomy as any of : peat metagenome, seawater metagenome, marine metagenome, marine plankton metagenome, marine sediment metagenome, karst metagenome, lagoon metagenome, lake water metagenome, glacier lake metagenome, freshwater metagenome, freshwater sediment metagenome, aquatic metagenome, bog metagenome, drinking water metagenome, estuary metagenome, Winogradsky column metagenome, alkali sediment metagenome, sediment metagenome or water metagenome.

rarefied libraries will be mapped to representative MAGs of the MAG collection to identify global distributions. This short paper is here to make this first iteration of the data available as the authors think it is valuable for it to reach the community. Additionally some controversy is currently rife about the 'misuse' of some the data submitted to SRA. I hereby present the community with my project and some preliminary results, out of which any actor that is unhappy to have his public data involved is invited to contact the corresponding author, who will remove the data-set from any future peer-reviewed publication, and down-stream analysis, and remark in the supplemental data on the removal of the data.



tSNE transformation of the similarity matrix of the library MinHashes

Methods

The libraries of interest are identified with the `Entrez`-module[1] of the `biopython` python-package[2], and then downloaded with `parallel-fastq-dump`[3]

a conveniently parallelized wrapper for `fastq-dump`[4]. The reads were quality filtered and trimmed with `fastp`[5](version 0.20.1) due to it's agnosticism to adapters, and speed. `MinHash`-signatures were computed for all libraries with `sourmash`[6](version 3.5.0 and with the `--track-abundance -k31 --scaled 1000` options) and all of the obtained signatures were compaired pairwise to obtain similarities. The libraries were rarefied to 100.000 read-pairs (or reads if library is unpaired) with the `reformat.sh` program from the `bbtools` software suite[7](version 38.18), to provide a handy exploration dataset of reads. The rRNA content of all libraries was obtained from the rarefied libraries with `SortMeRNA`[8] (version 4.2.0, in future version it will be obtained from whole library, it was an after though in this case). The full QCed libraries are assembled with megahit [9](version 1.2.9), only the contigs larger than 2.5kb were kept to reduce size and increase average quality of the assembly data. Each library is mapped to its assembly with bowtie2[10](version 2.4.1), and the mapping are post-processed with samtools[11](version 1.10). The mapping combined with the assembly is then used to bin the contigs with metabat2[12](version 2.15), the obtained bins are then quality checked with CheckM[13](version 1.1.3), and filtered based on a 30% completeness and 5% redundancy threashold. These are pretty lax filters based on [14] and personal experience, but are selected for capturing a wide rather then particularly accurate biodiversity. All bins were taxonomically annotated with `GTDBtk` [15](version 1.4.0, with release 95 of the database), and clustered into metagenomic Operational Taxonomic Units (mOTUs) with `mOTUlizer`[16](version 0.2.2), to speed up the process, only bins classified within the same family were clustered together. This whole workflow is intended to be wrapped in a single `snakemake`-pipeline[17], as of now it is not entirely so, only the processing of the single libraries is, and the rest is a number of separate scripts still to be properly integrated, nevertheless, all is available at `github.com/moritzbuck/SRAnsack`.

Usage cases

This data can be used for a number of purposes. You can look for mOTUs of your favorite microorganism and back track to samples rich in these, or use the read library against your favorite genome to find samples where they are highly abundant. Conversely you can compute a MinHash-signature of your freshly sequenced metagenome and compare it to the available signatures to find related samples even if the contextual data is missing or misleading. An other similar use would be to find your favorite metagenome from the SRA and check the library similarity file for related samples. Finally, what we intend to do it, first we will use the similarities between libraries to partition the data-set to generate coassemblies, and bins from these which will probably multiply around ~5 fold the number of MAGs, and we will use a combination of metadata and similarity between samples to generate a number of reference sets for certain biomes, so as to give a better genomic context for any metagenomic study done in them.

Available data

- MAG collection: ~100-zipped Gb
- MAG metadata:
- MAG ANIs: ~1 Gb
- Assembly collection: ~400-zipped Gb
- Signature collection: ~300-zipped Gb
- Rarefied read collection: ~400-zipped Gb
- Library metadata:
- Library similarities:

Bibliography

[1] "Entrez-module from biopython." https://biopython.org/docs/1.75/api/Bio.Entrez.html.

[2] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: Freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, Jun. 2009.

[3] R. Valieris, "Parallel-fastq-dump." May-2021.

[4] "SRA-Tools." https://ncbi.github.io/sra-tools/fastq-dump.html.

[5] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "Fastp: An ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018.

[6] C. T. Brown and L. Irber, "Sourmash: A library for MinHash sketching of DNA," *Journal of Open Source Software*, vol. 1, no. 5, p. 27, Sep. 2016.

[7] B. Bushnell, "BBMap," *SourceForge*. https://sourceforge.net/projects/bbmap/, 2014.

[8] E. Kopylova, L. Noé, and H. Touzet, "SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data," *Bioinformatics*, vol. 28, no. 24, pp. 3211–3217, Dec. 2012.

[9] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph," *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, May 2015.

[10] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012.

[11] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.

[12] D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang, "MetaBAT 2: An adaptive binning algorithm for robust and efficient genome

reconstruction from metagenome assemblies," *PeerJ*, vol. 7, Jul. 2019.

[13] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Research*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015.

[14] M. Olm, "Why genome completeness and contamination estimates are more complicated than you think," *microBEnet: the microbiology of the Built Environment network*. Dec-2017.

[15] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database," *Bioinformatics*, vol. 36, no. 6, pp. 1925–1927, Mar. 2020.

[16] M. Buck, "mOTUlizer." Jun-2021.

[17] J. Köster and S. Rahmann, "Snakemakea scalable bioinformatics workflow engine," *Bioinformatics*, vol. 28, no. 19, pp. 2520–2522, Oct. 2012.