The Journal of Open Source Software

ReferenceSeeker: rapid determination of appropriate reference genomes

Oliver Schwengers[1, 2, 3], **Torsten Hain**[2, 3], **Trinad Chakraborty**[2, 3], **and Alexander Goesmann**[1, 3]

1 Bioinformatics and Systems Biology, Justus Liebig University Giessen, Giessen, 35392, Germany **2** Institute of Medical Microbiology, Justus Liebig University Giessen, Giessen, 35392, Germany **3** German Centre for Infection Research (DZIF), partner site Giessen-Marburg-Langen, Giessen, Germany

DOI: 10.21105/joss.01994

Software

- Review
- Repository
- Archive

Editor: William Rowe

Reviewers:

- @standage
- @luizirber

Submitted: 19 December 2019
Published: 04 February 2020

Summary

The enormous success and ubiquitous application of next and third generation sequencing has led to a large number of available high-quality draft and complete microbial genomes in the public databases. Today, the NCBI RefSeq database release 90 alone contains 11,060 complete bacterial genomes (Haft et al., 2018). Concurrently, selection of appropriate reference genomes (RGs) is increasingly important as it has enormous implications for routine in-silico analyses, as for example in detection of single nucleotide polymorphisms, scaffolding of draft assemblies, comparative genomics and metagenomic tasks. Therefore, a rigorously selected RG is a prerequisite for the accurate and successful application of the aforementioned bioinformatic analyses. In order to address this issue several new databases, methods and tools have been published in recent years *e.g.* RefSeq, DNA-DNA hybridization (Meier-Kolthoff, Auch, Klenk, & Göker, 2013), average nucleotide identity (ANI) as well as percentage of conserved DNA (conDNA) values (Goris et al., 2007) and Mash (Ondov et al., 2016). Nevertheless, the sheer amount of currently available databases and potential RGs contained therein, together with the plethora of tools available, often requires manual selection of the most suitable RGs. To the best of the authors' knowledge, there is currently no such tool providing both an integrated, highly specific workflow and scalable and rapid implementation. ReferenceSeeker was designed to overcome this bottleneck. As a novel command line tool, it combines a fast kmer profile-based lookup of candidate reference genomes (CRGs) from high quality databases with rapid computation of (mutual) highly specific ANI and conserved DNA values.

Implementation

ReferenceSeeker is a linux command line tool implemented in Python 3. All necessary external binaries are bundled with the software. The tool itself requires no external dependencies other than Biopython for file input and output.

Databases

ReferenceSeeker takes advantage of taxon-specific custom databases in order to reduce data size and overall runtime. Pre-built databases for the taxonomic groups bacteria, archaea, fungi, protozoa and viruses are provided. Each database integrates genomic as well as taxonomic information comprising genome sequences of all RefSeq genomes with an assembly level 'complete' or whose RefSeq category is either denoted as 'reference genome' or 'representative genome', as well as kmer profiles, related species names, NCBI Taxonomy identifiers and

RefSeq assembly identifiers. For convenient and fully automatic updates, we provide locally executable scripts implemented in bash and Nextflow (Di Tommaso et al., 2017). Non-public genomes can be imported into existing or newly created databases by an auxiliary command line interface.

Database Lookup of CRGs

To reduce the number of necessary ANI calculations a kmer profile-based lookup of CRGs against custom databases is carried out. This step is implemented via Mash parameterized with a Mash distance of 0.1, which was shown to correlate well with an ANI of roughly 90% (Ondov et al., 2016) and thereby establishing a lower limit for reasonably related genomes. The resulting set of CRGs is subsequently reduced to a configurable number of CRGs (default=100) with the lowest Mash distances.

Determination of RG

Mash distances used for the preliminary selection of CRGs were shown to correlate well with ANI values capturing nucleotide-level sequence similarities. However, Mash distances do not correlate well with the conDNA statistic, which captures the query sequence coverage within a certain reference sequence (Figure 1). In order to precisely calculate sequence similarities beyond the capability of kmer fingerprints and to assure that RGs share an adequate portion of the query genome, ReferenceSeeker calculates both ANI and conDNA to derive a highly specific measure of microbial genome relationships (Goris et al., 2007).

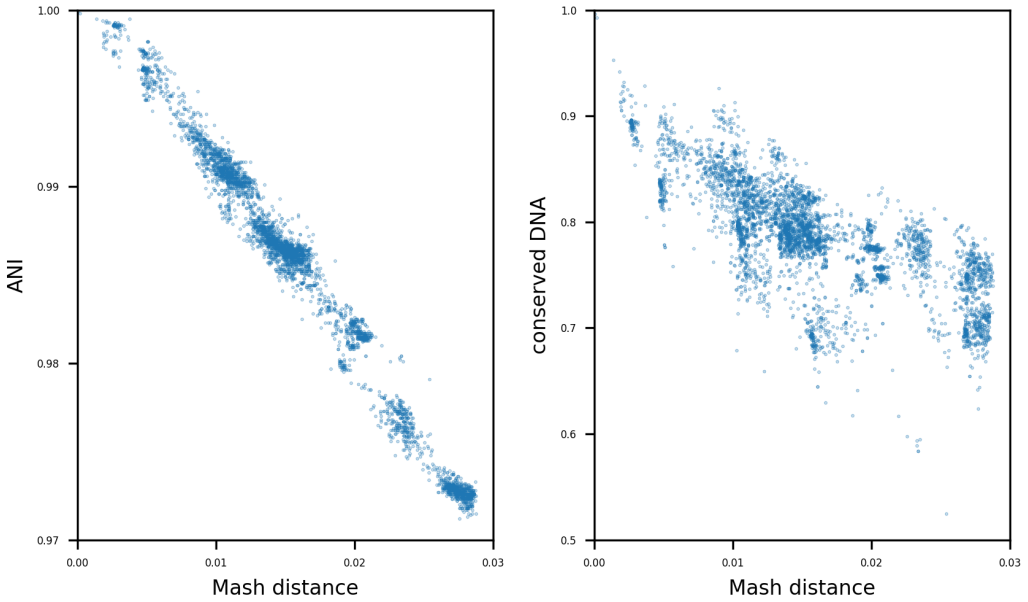

Figure 1: Figure 1: Scatter plots showing the correlation between Mash distance, ANI and conDNA values. ANI and conserved DNA values are plotted against Mash distance values for 500 candidate reference genomes with the lowest Mash distance within the bacterial database for 10 randomly selected *Escherichia coli* genomes from the RefSeq database, each.

Therefore, required sequence alignments are conducted via Nucmer of the MUMmer package (Marçais et al., 2018) as it was recently shown that Nucmer based implementations (ANIn) compare favourably to BLAST+ based implementations (ANIb) in terms of runtime. Exact calculations of ANI and conDNA values were adopted from (Goris et al., 2007) and are carried

out as follows. Each query genome is split into consecutive 1,020 bp nucleotide fragments which are aligned to a reference genome via Nucmer. The conDNA value is then calculated as the ratio between the sum of all aligned nucleotides within nucleotide fragments with an alignment with a sequence identity above 90% and the sum of nucleotides of all nucleotide fragments. The ANI value is calculated as the mean sequence identity of all nucleotide fragments with a sequence identity above 30% and an alignment length of at least 70% along the entire fragment length.

Given that compared genomes are closely related, *i.e.* they share an ANI of above 90%, it was also shown that ANIn correlates well with ANIb (Yoon, Ha, Lim, Kwon, & Chun, 2017). This requirement is ensured by the prior Mash-based selection of CRGs. As an established threshold for species boundaries (Goris et al., 2007), results are subsequently filtered by configurable ANI and conDNA values with a default of 95% and 69%, respectively. Finally, CRGs are sorted according to the harmonic mean of ANI and conDNA values in order to incorporate both the nucleotide identity and the genome coverage between the query genome and resulting CRGs. In this manner, ReferenceSeeker ensures that the resulting RGs sufficiently reflect the genomic landscape of a query genome. If desired by the user, this approach can be extended to a bidirectional computation of aforementioned ANI and conDNA values.

Application

ReferenceSeeker takes as input a microbial genome assembly in fasta format and the path to a taxonomic database of choice. Results are returned as a tabular separated list comprising the following information: RefSeq assembly identifier, ANI, conDNA, NCBI taxonomy identifier, assembly status and organism name. To illustrate the broad applicability at different scales we tested ReferenceSeeker with 12 microbial genomes from different taxonomic groups and measured overall runtimes on a common consumer laptop providing 4 cores and a server providing 64 cores (Table 1). For the tested bacterial genomes, ReferenceSeeker limited the number of resulting RGs to a default maximum of 100 genomes. Runtimes of archaeal and viral genomes are significantly shorter due to a small number of available RGs in the database and overall smaller genome sizes, respectively.

Table 1: Runtimes and numbers of resulting RG executed locally on a quad-core moderate consumer laptop and a 64 core server machine.

Genome	Genome Size [kb]	Laptop [mm:ss]	Server [mm:ss]	# RG
Escherichia coli str. K-12 substr. MG1665 (GCF_000005845.2)	4,641	3:24	0:30	100*
Pseudomonas aeruginosa PAO1 (GCF_000006765.1)	6,264	5:20	0:44	100*
Listeria monocytogenes EGD-e (GCF_000196035.1)	2,944	2:52	0:24	100*
Staphylococcus aureus subsp aureus NCTC 8325 (GCF_000013425.1)	2,821	2:31	0:21	100*
Halobacterium salinarum NRC-1 (GCF_000006805.1)	2,571	0:04	0:03	2
Methanococcus maripaludis X1 (GCF_000220645.1)	1,746	0:22	0:09	5
Aspergillus fumigatus Af293 (GCF_000002655.1)	29,384	3:11	2:07	1
Candida albicans SC5314 (GCF_000182965.3)	14,282	0:21	0:19	1

Genome	Genome Size [kb]	Laptop [mm:ss]	Server [mm:ss]	# RG
Entamoeba histolytica HM-1:IMSS (GCF_000208925.1)	20,835	6:04	4:41	1
Plasmodium falciparum 3D7 (GCF_000002765.4)	23,326	2:52	1:49	1
Influenza A virus (GCF_001343785.1)	13	0:03	0:02	1
Human coronavirus NL63 (GCF_000853865.1)	27	0:04	0:02	1

Availability

The source code is available on GitHub under a GPL3 license: https://github.com/oschwengers/referenceseeker. The software is packaged and publicly available via BioConda. Pre-built databases for bacteria, archaea, fungi, protozoa and viruses are hosted at Zenodo: https://doi.org/10.5281/zenodo.3562005. All installation instructions, examples and download links are provided on GitHub.

Funding

This work was supported by the German Center of Infection Research (DZIF) [DZIF grant 8000 701–3 (HZI), TI06.001 and 8032808811 to T.C.]; the German Network for Bioinformatics Infrastructure (de.NBI) [BMBF grant FKZ 031A533B to A.G.]; and the German Research Foundation (DFG) [SFB-TR84 project A04 (TRR84/3 2018) to T.C., KFO309 Z1 (GO 2037/5-1) to A.G., SFB-TR84 project B08 (TRR84/3 2018) to T.H., SFB1021 Z02 (SFB 1021/2 2017) to T.H., KFO309 Z1 (HA 5225/1-1) to T.H.].

Authors declare that there are no conflicts of interest.

Acknowledgement

The authors thank Karina Brinkrolf for valuable discussions, testing and bug reports.

References

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, *35*(4), 316–319. doi:10.1038/nbt.3820

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology*, *57*(1), 81–91. doi:10.1099/ijs.0.64483-0

Haft, D. H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., et al. (2018). RefSeq: An update on prokaryotic genome annotation and curation. *Nucleic Acids Research*, *46*(D1), D851–D860. doi:10.1093/nar/gkx1068

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), e1005944. doi:10.1371/journal.pcbi.1005944

Meier-Kolthoff, J. P., Auch, A. F., Klenk, H.-P., & Göker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, *14*, 60. doi:10.1186/1471-2105-14-60

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: Fast genome and metagenome distance estimation using minhash. *Genome Biology*, *17*(1), 132. doi:10.1186/s13059-016-0997-x

Yoon, S. H., Ha, S. M., Lim, J., Kwon, S., & Chun, J. (2017). A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, *110*(10), 1281–1286. doi:10.1007/s10482-017-0844-4