bioRxiv preprint doi: https://doi.org/10.1101/863621; this version posted December 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ReferenceSeeker: rapid determination of appropriate reference genomes

Authors:
Schwengers, O.[1,2,3,*], Hain, T.[2,3], Chakraborty, T.[2,3,†], Goesmann, A.[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

* To whom correspondence should be addressed.
† Contributed equally to this work.

Running head: ReferenceSeeker

Abstract

Summary: The large and growing number of microbial genomes available in public databases makes the optimal selection of reference genomes necessary for many in-silico analyses, e.g. single nucleotide polymorphism detection, scaffolding and comparative genomics, increasingly difficult. Here, we present ReferenceSeeker, a novel command line tool combining a fast kmer profile-based database lookup of candidate reference genomes with subsequent calculation of highly specific average nucleotide identity (ANI) values for the rapid determination of appropriate reference genomes. Pre-built databases for bacteria, archaea, fungi, protozoa and viruses based on the RefSeq database are provided for download.
Availability and Implementation: ReferenceSeeker is open source software implemented in Python. Source code and binaries are freely available for download at https://github.com/oschwengers/referenceseeker under the GNU GPL3 license.
Contact: referenceseeker@computational.bio

bioRxiv preprint doi: https://doi.org/10.1101/863621; this version posted December 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

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, 2013**), average nucleotide identity (ANI) 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 require 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 highly specific ANI values.

Implementation

ReferenceSeeker is a linux command line tool implemented in Python. All necessary external binaries are bundled with the software. It 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**).

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

bioRxiv preprint doi: https://doi.org/10.1101/863621; this version posted December 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 maximum of 100 CRG with the lowest Mash distances.

Determination of RG

As a highly specific measure for microbial genome relationships ReferenceSeeker uses ANI and conserved DNA (conDNA) values. Calculation of these metrics is implemented vía Nucmer contained in the MUMmer package (**Kurtz et al., 2004**) as it was recently shown that Nucmer based implementations (ANIn) compare favourably against BLAST+ based implementations (ANIb) in terms of runtime. Given that compared genomes are closely related, i.e. shared ANI is above 90%, it was also shown that ANIn correlates well with ANIb (**Yoon et al., 2017**). This 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 ANI and conDNA values of 95% and 69%, respectively. Finally, CRGs are sorted according to the harmonic mean of ANI and conDNA values in order to incorporate both nucleotide identity and genome coverage of query genomes and resulting CRGs. In this manner, ReferenceSeeker ensures that the resulting RGs sufficiently reflect the genomic landscape of a query genome.

Optional Scaffolding of Draft Assemblies

As recently published multi-reference contig scaffolders highly profit from the inclusion of a reasonable number of related RGs, ReferenceSeeker can optionally use the best matching 20 RGs as an input for a subsequent scaffolding step conducted via MeDuSa (**Bosi et al., 2005**).

Setup

ReferenceSeeker and its dependencies require the Python packages Biopython, NumPy and NetworkX, which are installed via pip. In order to set up ReferenceSeeker, users need to download the public repository from GitHub and set an environment variable pointing to the download directory. Pre-built databases for bacteria, archaea, fungi, protozoa and viruses can be downloaded from a public web server. For further convenience and interoperability, we provide a Docker image hosted at Docker Hub. All installation instructions, examples and download links are provided on GitHub.

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

bioRxiv preprint doi: https://doi.org/10.1101/863621; this version posted December 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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** [m:s]	**Server** [m:s]	**# 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
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

Asterisks denote number of RGs limited by ReferenceSeeker to a default maximum of 100.

bioRxiv preprint doi: https://doi.org/10.1101/863621; this version posted December 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Availability

The source code is available on GitHub under a GPL3 license: https://github.com/oschwengers/referenceseeker. The Docker container image is accessible at Docker Hub: https://hub.docker.com/r/oschwengers/referenceseeker. Download links to pre-built databases are provided in the GitHub repository readme file.

FUNDING

This work was supported by the German Center of Infection Research (DZIF) [DZIF grant 8000 701–3 (HZI) and TI06.001 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.].

Conflict of Interest: none declared.

ACKNOWLEDGEMENT

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

REFERENCES

Di Tommaso, P. et al. (2017) Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4), pp.316–319.

Bosi, E. et al. (2015) MeDuSa: A multi-draft based scaffolder. *Bioinformatics*, 31(15), pp.2443–2451.

Goris, J. et al. (2007) DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.*, 57(1), pp.81–91.

Haft, D.H. et al. (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic acids research*, 46(D1), pp.D851–D860.

Kurtz, S. et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, 5(2), p.R12.

Meier-Kolthoff, J.P. et al. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, 14, p.60.

bioRxiv preprint doi: https://doi.org/10.1101/863621; this version posted December 6, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Ondov, B.D. et al., 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, 17(1), p.132.

Yoon, S.H. et al. (2017) A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol.*, 110(10), pp.1281–1286.