W88–W92 *Nucleic Acids Research, 2019, Vol. 47, Web Server issue* *Published online 22 May 2019*
doi: 10.1093/nar/gkz423

M1CR0B1AL1Z3R—a user-friendly web server for the analysis of large-scale microbial genomics data

Oren Avram, Dana Rapoport, Shir Portugez and Tal Pupko*

The School of Molecular Cell Biology & Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel

Received March 12, 2019; Revised April 29, 2019; Editorial Decision May 05, 2019; Accepted May 06, 2019

ABSTRACT

Large-scale mining and analysis of bacterial datasets contribute to the comprehensive characterization of complex microbial dynamics within a microbiome and among different bacterial strains, e.g., during disease outbreaks. The study of large-scale bacterial evolutionary dynamics poses many challenges. These include data-mining steps, such as gene annotation, ortholog detection, sequence alignment and phylogeny reconstruction. These steps require the use of multiple bioinformatics tools and ad-hoc programming scripts, making the entire process cumbersome, tedious and error-prone due to manual handling. This motivated us to develop the M1CR0B1AL1Z3R web server, a 'one-stop shop' for conducting microbial genomics data analyses via a simple graphical user interface. Some of the features implemented in M1CR0B1AL1Z3R are: (i) extracting putative open reading frames and comparative genomics analysis of gene content; (ii) extracting orthologous sets and analyzing their size distribution; (iii) analyzing gene presence–absence patterns; (iv) reconstructing a phylogenetic tree based on the extracted orthologous set; (v) inferring GC-content variation among lineages. M1CR0B1AL1Z3R facilitates the mining and analysis of dozens of bacterial genomes using advanced techniques, with the click of a button. M1CR0B1AL1Z3R is freely available at https://microbializer.tau.ac.il/.

INTRODUCTION

In a typical microbial genomics study, a few dozen bacterial samples are sequenced using next generation sequencing technologies, with each sample representing a different bacterial species, strain or isolate. The obtained reads are assembled, generating a set of contigs for each sample. This set of partially assembled genomes is then analyzed using bioinformatics tools to gain insights into the bacterial evolutionary dynamics and genomic composition of these samples. Typical research challenges are: (i) inferring the core genome and pangenome (the set of genes shared by all members of the analyzed clade and the set of genes shared by at least one member of the analyzed clade, respectively) (1); (ii) reconstructing the evolutionary history of the analyzed samples as a phylogenetic tree (2); (iii) analyzing the variation in GC content among samples (3); (iv) analyzing the gene gain and loss dynamics, which is often an indication of the intensity of horizontal gene transfer (4); (v) detecting genes that are likely to have experienced positive selection (5–7).

The above computations require the use of multiple bioinformatics tools and ad-hoc programming scripts to handle information flow among the various programs, which in turn necessitates a dedicated bioinformatician to conduct such analyses. As a result, research laboratories began implementing their own in-house analysis pipelines, and later, different analysis applications began to emerge (8–10). These applications require specific working environments (i.e., operating systems), computation power (multicore machines), and more than basic technological skills (e.g., installation and running). Previously developed web tools to analyze sequenced microbial genomes are the MG-RAST, Pan-X and PGAweb web servers. MG-RAST allows finding and annotating gene functions or pathways by comparing genes to other databases (11). It differs from M1CR0B1AL1Z3R in that the latter focuses on comparing genomes rather than on their annotation and does not rely on external databases. The Pan-X web server provides ready-made examples of different microbial datasets (8). However, this web server does not allow providing unpublished genomic sequences as input. PGAweb provides several outputs, such as an analysis of the orthologous groups and reconstruction of the phylogenetic relationships among the sequences (12). However, in contrast to the M1CR0B1AL1Z3R web server, described below, it can only handle up to 50 genomic samples. In addition, phylogenetic relationships are reconstructed using neighbor joining or UPGMA, which are known to be less accurate than state-of-the-art methodologies for tree reconstruction such as maximum-likelihood and Bayesian approaches (13).

*To whom correspondence should be addressed. Tel: +972 3 640 7693; Fax: +972 3 642 2046; Email: talp@tauex.tau.ac.il

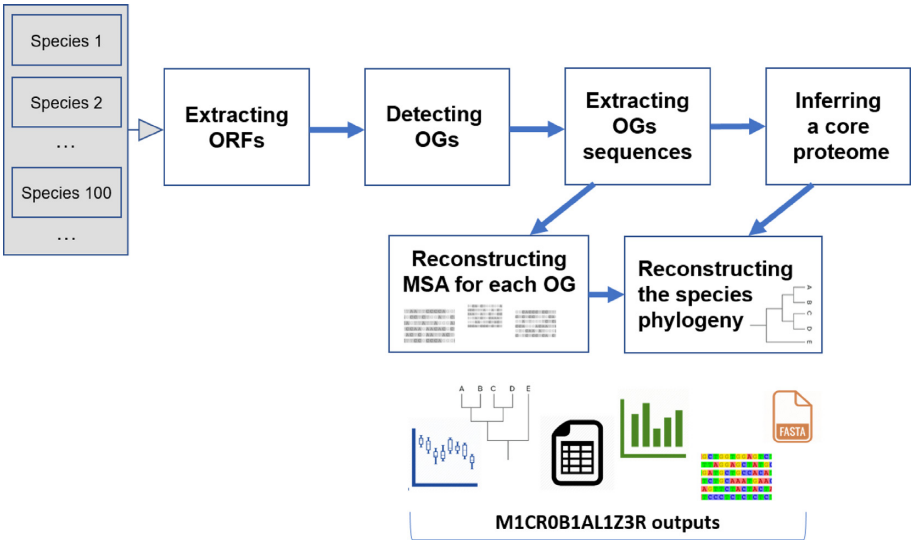

Figure 1. M1CR0B1AL1Z3R web server workflow. MSA, multiple sequence alignment. OG, orthologous group.

Here we present the M1CR0B1AL1Z3R (pronounced: microbializer) web server. M1CR0B1AL1Z3R was developed to facilitate microbial analyses and make them more accessible to the scientific community. M1CR0B1AL1Z3R utilizes a versatile computational pipeline that runs on the cloud and provides quick and easy analyses of bacterial genomics data for all users (Figure 1). No installation and no other prerequisites are needed. Visual and textual results that are ready for publication or further analysis are given as output.

MATERIALS AND METHODS

Input

The M1CR0B1AL1Z3R web server requires assembled genomic sequences (fully assembled or as contigs) from several clades. Each clade can represent a bacterial (or archaeal) isolate, strain or species. Each clade should be in a separate Fasta format file (such files are generated using assembly programs such as Velvet (14) or Canu (15)). Notably, in many metagenomic studies, the assignment of the various contigs to separate clades is unknown, and in this case, the data should be binned prior to running M1CR0B1AL1Z3R (16). To upload the files to M1CR0B1AL1Z3R, we ask the user to put them in a zipped folder (zip or tar.gz). Upon completion of the analyses, a link to the results is sent to the user if they choose to provide their email address. The results remain available on the web server for at least 3 months.

Extracting putative open reading frames (ORFs)

We extract ORFs from each genome using Prodigal (17) in 'normal' mode. Prodigal uses an unsupervised machine learning approach to extract protein-coding ORFs.

Extracting orthologous sets

A homology search is conducted in which each ORF is queried against all other ORFs in the database (all-against-all). Homology searches are executed using the equivalent of tBlastX in the MMSEQS2 program, which is ∼400 times faster than BLAST with similar accuracy (18). For each ORF, we record the top hit in each other genome. If ORF x in genome i is the top hit for ORF y in genome j and vice versa, these two ORFs are considered putative orthologs (best reciprocal hit, as in (19)). This pairwise analysis induces a graph in which each node is an ORF, and two nodes are connected if they are best reciprocal hits. An orthologous group is a set of nodes that are highly connected to each other and are separated from the rest of the nodes. We use the Markov Cluster (MCL) algorithm (as done in the OrthoMCL pipeline (20)) with default parameters (inflation parameter = 2.0) to detect these high-confidence orthologous groups.

Multiple sequence alignments (MSAs) and phylogenetic tree reconstruction

For each orthologous group, all sequences are first translated and the resulting protein sequences are then aligned using MAFFT, with the '–auto' flag, which automatically selects an appropriate MAFFT algorithm (L-INS-i, FFT-NS-i or FFT-NS-2) according to the size of the analyzed dataset (21). Sequences are then reverse-translated so that codon-level alignments can also be computed (as in (22)). A maximum-likelihood phylogenetic tree is reconstructed based on the concatenated protein MSA of all core genes, i.e., genes shared among all strains (see below), using RAxML (23) with default parameters, the LG replacement matrix (24), and a discrete gamma distribution with four categories and an invariant category (LG+G+I) to account for among-site-rate variation (of note, we have recently shown that when searching for the maximum-likelihood tree topology, using LG+G+I provides results that are as accurate as when a model selection step is introduced, and the latter is therefore not mandatory (25)). The tree is visualized using PhyD3 (26).

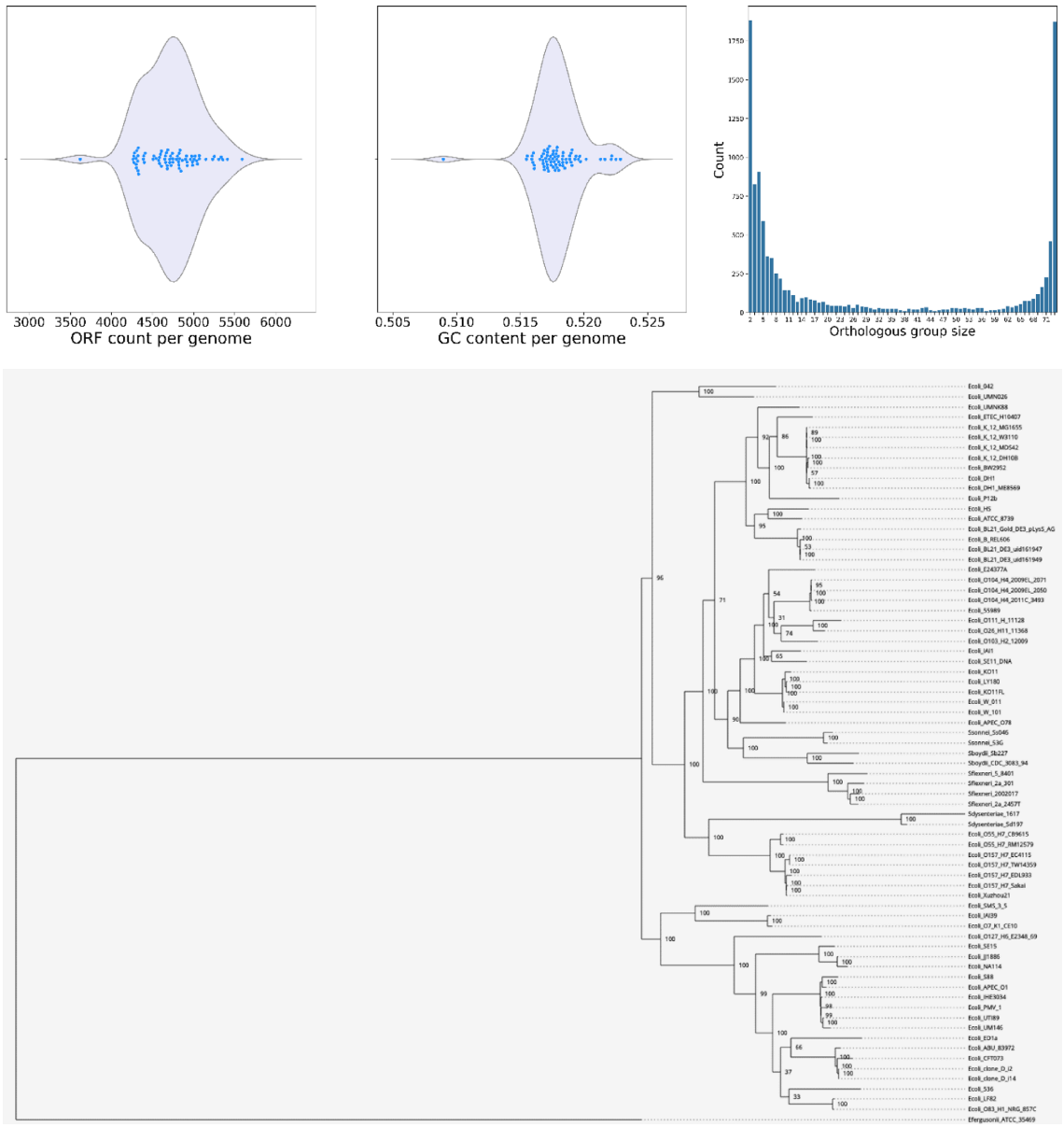

Figure 2. Selected visual outputs of the M1CR0B1AL1Z3R web server. Top panel (left to right): distribution of the number of ORFs in each genome; distribution of %GC in each genome; distribution of the sizes of the various orthologous groups. Bottom panel: phylogenetic tree representing the evolutionary relationships among all samples. The maximum-likelihood-based tree was reconstructed according to the core proteome as inferred from the orthologous group data.

GC content

For each genome, the GC content is computed from the set of ORFs using an in-house Python script.

Output

The following results are provided: (i) a text file with ORF counts per genome and its graphical representation as a violin plot; (ii) a curated file listing the orthologous sets and a histogram providing the distribution of set sizes; (iii) the unaligned sequences, the multiple sequence alignment at the protein level and the multiple sequence alignment at the codon level for each orthologous set. Both protein alignments and codon alignments are often used in downstream analyses, e.g., to find protein motifs (27) and to search for positive Darwinian selection (6), respectively. The unaligned sequences are also available if the user wishes, for example, to realign the sequences using another alignment program; (iv) a table in which each row is an orthologous group and each column is the set of genes of a specific sample (genome). The *i,j* entry contains the corresponding gene name of the i^{th} group and j^{th} sample, if such an entry exists (this is especially useful if the input includes at least one annotated genome). In addition, we provide a Fasta file with the phyletic pattern of all ortholog groups. Each record contains a sequence of '1's and '0's in the i^{th} place, depending whether it has a member gene in the i^{th} ortholo-

gous group or not, respectively (28). The generated phyletic pattern data (together with the species tree) can be further analyzed by the GLOOME web server (4), which allows inference of gene gain and loss rates, and ancestral reconstruction of these events along the species tree. In addition, we specifically provide a file with the list of ORFs shared by all samples, i.e., the orthologous group comprising the core proteome, and the concatenated protein alignment of this core proteome in Fasta format. The web server also provides means to extract the proteome shared by x% of the analyzed strains (where $x = 100$ is the default core proteome); (v) the phylogenetic species tree representing the evolutionary relationships between all samples, both as a text file in Newick format and using an online interactive visualizer; (vi) a text file with the GC content of each genome and a graphical representation using a violin plot.

Implementation

M1CR0B1AL1Z3R is implemented in Python 3.6. The source code is available at: https://github.com/orenavram/microbializer. The web server jobs are processed on ProLiant XL170r Gen9 servers, equipped with 128 GB RAM and 28 CPU cores per node. The Gallery, Overview, and Frequently Asked Questions (FAQ) sections of the web server should help users get the most out of the web server. A running example (different from the case studies analyzed in the Gallery) is also provided.

CASE STUDIES

The various analyses and outputs of M1CR0B1AL1Z3R are demonstrated using three datasets: (i) a set of 50 pathogenic *Escherichia coli* lineage ST131 genomes (29). This dataset represents highly similar clinical isolates of a specific bacterial species. We added an outgroup sequence to this dataset, the genomic sequence of *Escherichia fergusonii*; (ii) a collection of 73 different *Escherichia* genomes (72 of which are *E. coli* and one *E. fergusonii*). The 72 genomes are all fully sequenced *E. coli* genomes available as of December 2018 in the NCBI repository, and the *E. fergusonii* genome is used as an outgroup; (iii) a collection of 29 different Gammaproteobacteria genomes, taken from Pérez *et al.* (30). Together, these datasets demonstrate the applicability of M1CR0B1AL1Z3R for the analysis of a range of phylogenetic diversity, from different isolates to different species belonging to different bacterial orders. The complete results for these three examples are available in the Gallery section of the web server. For example, for dataset (ii), the number of ORFs varies from 3,621 to 5,592, with the smallest genome being 3,976,195 bp and the largest 5,697,240 bp. The entire set is comprised of 8,811 orthologous groups, 1,863 of which comprise the core genome. The multiple sequence alignment of the core proteome (618,921 amino acid sites) was used to reconstruct the maximum-likelihood phylogenetic tree, which is consistent with previously established *E. coli* phylogeny (31). The GC content of the analyzed genomes varies from 50.9 to 52.3%. The graphical outputs describing the ORF counts, orthologous group size dispersion, GC-content variation and phylogenetic relationships are shown in Figure 2.

FUNDING

Israel Science Foundation (ISF) [802/16 to T.P.]; Edmond J. Safra Center for Bioinformatics at Tel Aviv University Fellowship (in part). Funding for open access charge: ISF.
Conflict of interest statement. None declared.

REFERENCES

1. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.
2. Daubin,V., Gouy,M. and Perrière,G. (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.*, **12**, 1080–1090.
3. Hildebrand,F., Meyer,A. and Eyre-Walker,A. (2010) Evidence of selection upon genomic GC-Content in bacteria. *PLoS Genet.*, **6**, e1001107.
4. Cohen,O. and Pupko,T. (2011) Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum Parsimony—a simulation study. *Genome Biol. Evol.*, **3**, 1265–1275.
5. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
6. Stern,A., Doron-Faigenboim,A., Erez,E., Martz,E., Bacharach,E. and Pupko,T. (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Res.*, **35**, W506–W511.
7. Pond,S.L.K., Frost,S.D.W. and Muse,S. V. (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, **21**, 676–679.
8. Ding,W., Baumdicker,F. and Neher,R.A. (2018) panX: pan-genome analysis and exploration. *Nucleic Acids Res.*, **46**, e5.
9. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
10. Huson,D.H., Auch,A.F., Qi,J. and Schuster,S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res.*, **17**, 377–386.
11. Keegan,K.P., Glass,E.M. and Meyer,F. (2016) MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol. Biol.*, **1399**, 207–233.
12. Chen,X., Zhang,Y., Zhang,Z., Zhao,Y., Sun,C., Yang,M., Wang,J., Liu,Q., Zhang,B., Chen,M. *et al.* (2018) PGAweb: A web server for bacterial pan-genome analysis. *Front. Microbiol.*, **9**, 1910.
13. Anisimova,M., Liberles,D.A., Philippe,H., Provan,J., Pupko,T. and von Haeseler,A. (2013) State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC Evol. Biol.*, **13**, 161.
14. Zerbino,D.R. and Birney,E. (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
15. Koren,S., Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
16. Liu,Y., Hou,T., Kang,B. and Liu,F. (2017) Unsupervised binning of metagenomic assembled contigs using improved fuzzy C-Means method. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **14**, 1459–1467.
17. Hyatt,D., Chen,G.-L., LoCascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
18. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
19. Dagan,T. and Martin,W. (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 870–875.
20. Li,L., Stoeckert,C.J. and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
21. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
22. Wernersson,R. and Pedersen,A.G. (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.*, **31**, 3537–3539.

23. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
24. Le,S.Q. and Gascuel,O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.
25. Abadi,S., Azouri,D., Pupko,T. and Mayrose,I. (2019) Model selection may not be a mandatory step for phylogeny reconstruction. *Nat. Commun.*, **10**, 934.
26. Kreft,Ł., Botzki,A., Coppens,F., Vandepoele,K. and Van Bel,M. (2017) PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*, **33**, 2946–2947.
27. Krystkowiak,I., Manguy,J. and Davey,N.E. (2018) PSSMSearch: a server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic Acids Res.*, **46**, W235–W241.
28. Hao,W. and Golding,B. (2008) Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics*, **9**, 235.
29. McNally,A., Oren,Y., Kelly,D., Pascoe,B., Dunn,S., Sreecharan,T., Vehkala,M., Välimäki,N., Prentice,M.B., Ashour,A. *et al.* (2016) Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLOS Genet.*, **12**, e1006280.
30. Pérez,A.G., Angarica,V.E., Vasconcelos,A.T.R. and Collado-Vides,J. (2007) Tractor_DB (version 2.0): a database of regulatory interactions in gamma-proteobacterial genomes. *Nucleic Acids Res.*, **35**, D132–D136.
31. Oren,Y., Smith,M.B., Johns,N.I., Kaplan Zeevi,M., Biran,D., Ron,E.Z., Corander,J., Wang,H.H., Alm,E.J. and Pupko,T. (2014) Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 16112–16117.