The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes

Luis M Rodriguez-R, Konstantinos T Konstantinidis

Genomic and metagenomic analyses are increasingly becoming commonplace in several areas of biological research, but recurrent specialized analyses are frequently reported as in-house scripts rarely available after publication. We describe the enveomics collection, a growing set of actively maintained scripts for several recurrent and specialized tasks in microbial genomics and metagenomics, and present a graphical user interface and several case studies. Our resource includes previously described as well as new algorithms such as Transformed-space Resampling In Biased Sets (TRIBS), a novel method to evaluate phylogenetic under- or over-dispersion in reference sets with strong phylogenetic bias. The enveomics collection is freely available under the terms of the Artistic License 2.0 at https://github.com/lmrodriguezr/enveomics and for online analysis at http://enve-omics.ce.gatech.edu .

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1900v1 | CC-BY 4.0 Open Access | rec: 27 Mar 2016, publ: 27 Mar 2016

The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes

Luis M Rodriguez-R[1,2,*] and Konstantinos T. Konstantinidis[1,2,3,*]

[1] School of Biology, [2] Center for Bioinformatics and Computational Genomics, and [3] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, 30332, USA
* To whom correspondence should be addressed. Tel: +1 (404) 580-0541; Fax: +1 (404) 894-8266; Email: lmrodriguezr@gmail.com. Correspondence may also be addressed to K.T. Konstantinidis. Tel: +1 (404) 385-4581; Fax: +1 (404) 894-8266; Email: kostas@ce.gatech.edu

ABSTRACT

Genomic and metagenomic analyses are increasingly becoming commonplace in several areas of biological research, but recurrent specialized analyses are frequently reported as in-house scripts rarely available after publication. We describe the enveomics collection, a growing set of actively maintained scripts and bioinformatics algorithms for several recurrent and specialized tasks in microbial genomics and metagenomics, and present a graphical user interface and several case studies. Our resource includes previously described (*e.g.*, Nonpareil, MyTaxa, ANI/AAI calculator) as well as new algorithms such as Transformed-space Resampling In Biased Sets (TRIBS), a novel method to evaluate phylogenetic under- or over-dispersion in reference sets with strong phylogenetic bias. The enveomics collection is freely available under the terms of the Artistic License 2.0 at https://github.com/lmrodriguezr/enveomics and for online analysis at http://enve-omics.ce.gatech.edu.

INTRODUCTION

Microbial genomics and metagenomics have become key components of several areas of modern research including biomedicine, epidemiology, plant and animal pathology, environmental engineering and science, microbial ecology, and evolutionary biology. Specialized computational analyses in these areas are hence becoming commonplace for the non-expert, often resulting in the reimplementation of scripts critical for understanding and reproducing the reported results, with varying levels of quality, reproducibility, and availability. While the literature analysing *ad hoc* scripts is scarce, a 2004 survey on the availability of URLs reported in MEDLINE found that 19% of 1,020 analysed pages were always unavailable, and only 63% were always available (Wren, 2004). Moreover, we searched the manuscripts available as full-text in PubMed Central with the terms “in-house script”, “in-house developed script”, “in-house perl script” (other languages didn’t return additional results), or the same terms in plural, and found 1,929 matching articles (as of January 05, 2016). From these, 1,654 were related to genomics or metagenomics and 449 to microbial genomics or metagenomics. We further explored the latter set of manuscripts, and found that only 6% provided access to the source code (26/449), with an additional 1% reporting websites no longer available (3/449) or not including the reported scripts (3/449). 3% of the manuscripts explicitly indicated that the scripts were available upon request (13/449), but in only 3 cases the authors provided the code within two months of the request. The large majority (90%) did not provide

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1900v1 | CC-BY 4.0 Open Access | rec: 27 Mar 2016, publ: 27 Mar 2016

any reference, provided references to previous publications of the same group not including the source code, or referenced only the programming language in which the scripts were implemented. While this survey is not a systematic analysis of availability of in-house scripts, nor does it provide quality assessments, the results do underscore the prevalence of a phenomenon that undermines reproducibility in studies applying microbial genomics or metagenomics. On one hand, individual tools are the basis for complete and reproducible methods that are reported either in manuscripts, white papers, or standard operating procedures (SOPs), but the abovementioned statistics showed that the tools infrequently become available. On the other hand, a suitable alternative to providing developed tools for results reproducibility is to make data on each step of the analyses publicly available, but this approach is also rarely adopted, with the added issues of larger file sizes and heterogeneity, making the documentation of data even more challenging than documenting code. Here, we present a growing collection of actively maintained scripts for several recurrent and specialized tasks in microbial genomics and metagenomics, together with comprehensive documentation, a graphical user interface, and some cases of use. Our collection may also constitute a reference example for other researchers in the future, and an actively maintained framework that could be collaboratively expanded.

IMPLEMENTATION

The enveomics collection is a multi-language set of over 70 independent scripts that accomplish specialized tasks in genomics and metagenomics, including code in Ruby, Perl, AWK, Bash, and R. In addition, the collection features reusable libraries that automate recurrent sub-tasks. For example, the enveomics_rb Ruby library includes object-oriented representations of trees, read-placement results, sets of orthologous genes, and complex (non-contiguous) sequence coordinates, together with methods for accessing and downloading remote data. The R code is packaged into a single library (enveomics.R) to simplify its distribution.

Preferred file formats

Format incompatibility between data sources and analysis tools is a common problem in bioinformatics, and there are several tools and libraries dedicated to the translation between format specifications (Rice, Longden & Bleasby, 2000). In order to mitigate the impact of this problem, the enveomics collection has been designed to support only a reduced number of formats, with a wide range of alternative variations. For example, sequence files are expected to be in FastA, but the scripts in the collection always support multi-FastA and can parse variations of the definition lines and colon-lead comments (Suppl. Table S1). Supported formats for other data types include tabular BLAST for similarity searches (including variations with additional columns and comments), JPlace for phylogenetic read placement (Matsen et al., 2012), and tables in raw text with tab-delimited columns.

Access to remote servers

Local data sources are often insufficient and commonly out-dated. In response, we have implemented several utilities to simplify the automated access to remote databases using the Representational State Transfer Application Program Interfaces (RESTful APIs) of the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI), the U.S. National Center for Biotechnology Information (NCBI) E-Utilities, the Kyoto Encyclopedia of Genes and Genomes (KEGG), and the M5nr (Kanehisa & Goto, 2000; Sayers et al., 2009; Wilke et al., 2012; Li et al., 2015). All the scripts using these modules are categorized in Annotation/database mapping, and include additional documentation such as informing the user that third-party software or database is used and thus, the latter resources should be cited appropriately in any resulting publications.

Enveomics-GUI

The documentation and parameter descriptions for all the scripts are standardized into a set of JSON files that allow the dynamic creation of Graphical User Interface (GUI) forms though the enveomics-GUI package, including a set of examples and reference files (Fig. 1). The package is a collection of Ruby libraries, including EnveGUI that implements graphical user interaction with Shoes 4 (https://github.com/shoes/shoes4). The JSON files meet the definitions of the ECMA-404 standard (Ecma International, 2013), but their processing (implemented in the EnveJSON library) ignores object entries with "_" key, that are utilized for comments, and implements external file inclusion using the object entries with "_include" key. The package is distributed as source code (requires Shoes 4 and JRuby), as a stand-alone OS-independent Java Archive (JAR), and as a bundled Mac OS X application.

RESULTS

Reimplementations and novel algorithms

The enveomics collection aims to simplify the use of novel and previously described algorithms for the analysis of community (*e.g.*, Chao1.pl, AlphaDiversi-ty.pl, Newick.autoprune.R) and population diversity (*e.g.*, BlastTab.recplot2.R, RecPlot2.find_peaks.R, CharTable.classify.rb), among other tasks in microbial genomics and metagenomics. Here we describe representative modules (see also Suppl. Table S1) including algorithms developed by our group.

Reciprocal Best Match and Average Sequence Identity. The detection of Reciprocal Best Matches (RBMs) is a reliable method for the identification of orthology (Wolf & Koonin, 2012) that has been widely used in genome-aggregate metrics of genetic relatedness (Konstantinidis & Tiedje, 2005; Goris et al., 2007). Although phylogenetic reconstruction remains the gold standard for orthology detection, RBM provides a fast alternative for high-throughput analyses such as genome-wide scanning. The enveomics collection contains utilities for the detection of RBMs (rbm.rb) and the compilation of Orthology Groups (OGs; ogs.mcl.rb), as well as the estimation of Average Nucleotide Identity (ANI; ani.rb; generally suitable for comparisons of genomes assigned to the same genus) and Average Amino acid Identity (AAI; aai.rb; suitable for comparisons of genomes assigned to different species).

Transformed-space Resampling In Biased Sets (TRIBS). Environmental analyses often rely on pre-existing reference databases as a proxy to the presence of features in query datasets. However, databases seldom represent the source of the query sets uniformly, introducing sampling biases. TRIBS is a novel algorithm that reduces the impact of biased sampling by uniformly resampling reference objects in a transformed space generated by Multidimensional Scaling (MDS). This enables the testing of differences between a dataset and a given subset for the detection of under- or over-dispersion of traits (TRIBS.test.R, TRIBS.plot-test.R). The method was originally designed for the detection of phylogenetic under-dispersion of traits in groups of genomes with strong phylogenetic bias (Suppl. Fig. S1; TRIBS.test.R).

Automated pruning of phylogenetic trees. The enveomics collection also features a utility to automatically prune trees keeping clade representatives (Newick.autoprune.R), a useful tool for the navigation of large trees such as those produced from 16S rRNA gene databases. This script iteratively extracts the cophenetic matrix from a tree and removes terminal nodes with at least one other node closer than a target minimum distance (by default, the first quartile of all the paired distances in the initial tree). In some cases, the complete cophenetic matrix is prohibitively expensive to estimate (in the initial iterations for large trees); in those cases the script takes a random sample of terminal nodes and removes sister nodes (or their children) closer than the target distance. An example of a pruned tree is presented in Fig. 2B-C.

Case studies using the enveomics collection

Core genome phylogenies. Whole-genome phylogenetic reconstruction is a powerful method for the resolution of evolutionary relationships. The enveomics collection includes utilities to download genomes of a given species, detect RBMs between pairs of genomes, identify OGs, and identify the genes shared among all the genomes in the collection –the core genome– (RefSeq.download.bash, rbm.rb, ogs.mcl.rb, ogs.extract.rb). After computing independent alignments of each core OG, a concatenated alignment can be generated with the options of excluding invariable sites and keeping track of coordinates (Aln.cat.rb) to generate robust phylogenies with OG-specific models. In addition, the OGs can be used to estimate several gene-content properties (ogs.stats.rb) and the rarefied core and pan-genomes (ogs.core-pan.rb) of the species. As a less expensive alternative to the entire core genome phylogeny, one could also identify and analyse only the collection of 111 single-copy genes typically present in archaeal (often ~26 genes) and bacterial (often ~106 genes) genomes (HMM.essential.rb). We implemented a workflow using the enveomics collection, together with the multiple alignment tool Clustal Omega (Sievers et al., 2011) and the phylogenetic reconstruction tool RAxML (Stamatakis, 2014) and applied it to the 17 publicly available complete genomes of *Xanthomonas oryzae* (Fig. 2). The resulting phylogeny identifies known pathovars and the overall structure is consistent with a previous phylogenomic reconstruction (Rodriguez-

R et al., 2012). The complete analysis is fully automated, and the code is deposited in the enveomics collection at Examples/essential-phylogeny.bash. The execution took 31.2 minutes using two 2.9 GHz processors.

Gene variants in a metagenome. Characterizing the allelic diversity of genes in metagenomes allows targeted analyses of specific traits and the exploration of population discreteness and intra-population variations, independent of cultivation and amplification (Caro-Quintero & Konstantinidis, 2012; Rodriguez-R & Konstantinidis, 2014a). We explored the intra-population diversity of a metagenomic-recovered bin (LL-70.1) using the mapping of metagenomic reads (LL_1101B; SRR948448 (Tsementzi et al., 2014)) from a water sample in January 2011 at Lake Lanier (GA, USA). Read mapping was performed with BLAST (Altschul et al., 1990), and results were analysed and visualized using BlastTab.catsbj.pl and BlastTab.recplot.R (Fig. 3), revealing small gene-content variations (panels 2 and 4), but a large allelic variation and the presence of closely related organisms at about 90% ANI (panels 1 and 3). However, a clear genetic discontinuity exists separating this species, as evidenced by the gap around 95% identity, a phenomenon further discussed in (Caro-Quintero & Konstantinidis, 2012; Rodriguez-R & Konstantinidis, 2014a). The enveomics collection also includes utilities for the normalization (BlastTab.topHits_sorted.rb, BlastTab.sumPerHit.pl, BlastTab.seqdepth_ZIP.pl), characterization (Chao1.pl, AlphaDiversity.pl, TRIBS.test.R), and visualization (Table.barplot.R, TRIBS.plot-test.R, BlastTab.recplot2.R) of reference allele distributions in a metagenome using read mapping. Additionally, the allelic diversity of a particular gene of interest can be explored beyond known variants using phylogenetic read placement (Matsen, Kodner & Armbrust, 2010; Berger, Krompass & Stamatakis, 2011), that can be visualized in the interactive Tree of Life (iToL) (Letunic & Bork, 2007), and further explored to characterize distances to known variants or ancestral nodes (JPlace.to_iToL.rb, JPlace.distances.rb) as in (Rodriguez-R et al., 2015).

Availability

The source code for all the scripts and additional documentation are deposited and maintained at https://github.com/lmrodriguezr/enveomics. The enveomics-GUI is maintained at https://github.com/lmrodriguezr/enveomics-gui. In addition, we have made available a server with online interfaces for select tools at http://enve-omics.ce.gatech.edu/, including the ANI and AAI calculators, and previously reported tools like Nonpareil (Rodriguez-R & Konstantinidis, 2014b), a tool to estimate the level of coverage in metagenomic samples, and MyTaxa (Luo, Rodriguez-R & Konstantinidis, 2014), a taxonomic classification tool for sequence fragments.

DISCUSSION

The enveomics collection offers a wide array of tools implementing specialized recurrent tasks in microbial genomics and metagenomics and is aimed for users with or without expertise in bioinformatics. The collection features **(i)** a web-based interface for select tools and the complete documentation of all the tools, **(ii)** a comprehensive graphical user interface (GUI), **(iii)** a command-line interface (CLI) that allows integration with development platforms and automation, and **(iv)** Ruby and R application interfaces (API) for developers. In addition, the collection has a language-agnostic design, allowing the implementation of different tools in the most convenient language depending on available libraries or other considerations. To allow this heterogeneity, all the tools are integrated using a standardized JSON-based documentation scheme, allowing the incorporation of additional tools into the collection for the different interfaces. Finally, examples of input data and parameters are provided to encourage the quick use of the tools without dauntingly extensive user manuals.

A few of the scripts in our collection, in particular those implementing the most simple tasks, are overlapping with those developed by others (*e.g.*, (Rice, Longden & Bleasby, 2000; Stajich et al., 2002; Cock et al., 2009)). Our goal here was not to perform exhaustive comparisons to previously published scripts. As explained above, these scripts were frequently not available for comparisons. Rather, the goal was to put together a resource that offers easy-to-use tools for the non-bioinformatician and is comprehensive with respect to recurrent tasks in microbiome research. As such, we hope that the scientific community will find this resource useful, and will provide feedback on the scripts and algorithms, and suggestions for further improvements.

SUPPLEMENTARY DATA

Supplementary data are available at NAR Online.

ACKNOWLEDGEMENT

We would like to thank Luis H. (Coto) Orellana, Despina Tsementzi, and Michael R. Weigand, whose active involvement in testing and valuable discussions made this collection possible, and Dr. James R. Cole for his commentary on the tools described and the present manuscript. We would also like to thank early users for their active feedback.

FUNDING

This work was supported by the United States Department of Energy, Biological and Environmental Research Division (BER), Genomic Science Program [DE-SC0006662, DE-SC0004601]; and the United States National Science Foundation [1241046, 1356288].

REFERENCES

Altschul SF., Gish W., Miller W., Myers EW., Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215:403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Berger SA., Krompass D., Stamatakis A. 2011. Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology* 60:291–302. DOI: 10.1093/sysbio/syr010.

Caro-Quintero A., Konstantinidis KT. 2012. Bacterial species may exist, metagenomics reveal. *Environmental microbiology* 14:347–355. DOI: 10.1111/j.1462-2920.2011.02668.x.

Cock PJA., Antao T., Chang JT., Chapman BA., Cox CJ., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., Hoon MJL de. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423. DOI: 10.1093/bioinformatics/btp163.

Ecma International 2013. The JSON Data Interchange Format.

Goris J., Konstantinidis KT., Klappenbach JA., Coenye T., Vandamme P., Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology* 57:81–91. DOI: 10.1099/ijs.0.64483-0.

Kanehisa M., Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* 28:27–30. DOI: 10.1093/nar/28.1.27.

Konstantinidis KT., Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* 102:2567–2572. DOI: 10.1073/pnas.0409727102.

Letunic I., Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23:127–128. DOI: 10.1093/bioinformatics/btl529.

Li W., Cowley A., Uludag M., Gur T., McWilliam H., Squizzato S., Park YM., Buso N., Lopez R. 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Research* 43:W580–W584. DOI: 10.1093/nar/gkv279.

Luo C., Rodriguez-R LM., Konstantinidis KT. 2014. MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research* 42:e73–e73. DOI: 10.1093/nar/gku169.

Matsen FA., Hoffman NG., Gallagher A., Stamatakis A. 2012. A Format for Phylogenetic Placements. *PLoS ONE* 7:e31009. DOI: 10.1371/journal.pone.0031009.

Matsen FA., Kodner RB., Armbrust EV. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 11:538. DOI: 10.1186/1471-2105-11-538.

Rice P., Longden I., Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16:276–277. DOI: 10.1016/S0168-9525(00)02024-2.

Rodriguez-R LM., Grajales A., Arrieta-Ortiz M., Salazar C., Restrepo S., Bernal A. 2012. Genomes-based phylogeny of the genus Xanthomonas. *BMC Microbiology* 12:43. DOI: 10.1186/1471-2180-12-43.

Rodriguez-R LM., Overholt WA., Hagan C., Huettel M., Kostka JE., Konstantinidis KT. 2015. Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. *The ISME Journal*. DOI: 10.1038/ismej.2015.5.

Rodriguez-R LM., Konstantinidis KT. 2014a. Bypassing Cultivation To Identify Bacterial Species. *Microbe* 9:111–118.

Rodriguez-R LM., Konstantinidis KT. 2014b. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 30:629–635. DOI: 10.1093/bioinformatics/btt584.

Sayers EW., Barrett T., Benson DA., Bryant SH., Canese K., Chetvernin V., Church DM., DiCuccio M., Edgar R., Federhen S., Feolo M., Geer LY., Helmberg W., Kapustin Y., Landsman D., Lipman DJ., Madden TL., Maglott DR., Miller V., Mizrachi I., Ostell J., Pruitt KD., Schuler GD., Sequeira E., Sherry ST., Shumway M., Sirotkin K., Souvorov A., Starchenko G., Tatusova TA., Wagner L., Yaschenko E., Ye

J. 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 37:D5–15. DOI: 10.1093/nar/gkn741.

Sievers F., Wilm A., Dineen D., Gibson TJ., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson JD., Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7. DOI: 10.1038/msb.2011.75.

Stajich JE., Block D., Boulez K., Brenner SE., Chervitz SA., Dagdigian C., Fuellen G., Gilbert JGR., Korf I., Lapp H., Lehväslaiho H., Matsalla C., Mungall CJ., Osborne BI., Pocock MR., Schattner P., Senger M., Stein LD., Stupka E., Wilkinson MD., Birney E. 2002. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research* 12:1611–1618. DOI: 10.1101/gr.361602.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. DOI: 10.1093/bioinformatics/btu033.

Tsementzi D., Poretsky R., Rodriguez-R LM., Luo C., Konstantinidis KT. 2014. Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environmental Microbiology Reports* 6:640–655. DOI: 10.1111/1758-2229.12180.

Wilke A., Harrison T., Wilkening J., Field D., Glass EM., Kyrpides N., Mavrommatis K., Meyer F. 2012. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics* 13:141. DOI: 10.1186/1471-2105-13-141.

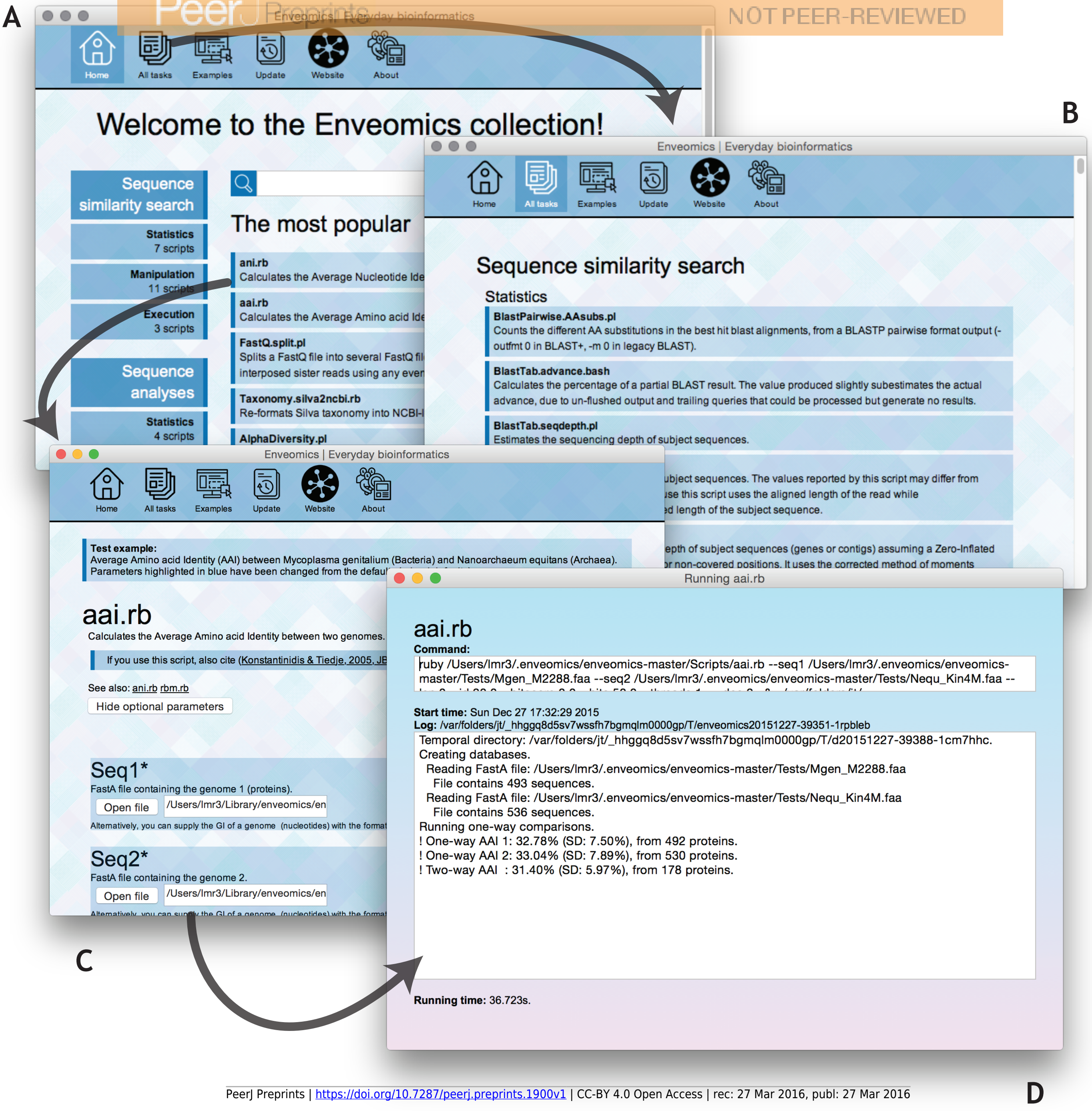
Wolf YI., Koonin EV. 2012. A Tight Link between Orthologs and Bidirectional Best Hits in Bacterial and Archaeal Genomes. *Genome Biology and Evolution* 4:1286–1294. DOI: 10.1093/gbe/evs100.

Wren JD. 2004. 404 not found: the stability and persistence of URLs published in MEDLINE. *Bioinformatics* 20:668–672. DOI: 10.1093/bioinformatics/btg465.

Figure 1(on next page)

Screen captures of the enveomics GUI in Mac OS X.

(A) Initial (home) screen with search bar, listing all categories and subcategories, and highlighting a few selected and randomly picked scripts. (B) Complete list of scripts per category. (C) Task form for aai.rb pre-filled with an example. (D) Result of the aai.rb analysis. All screen captures correspond to v0.1.2. Future versions may differ.



Welcome to the Enveomics collection!

Sequence similarity search

Statistics
7 scripts

Manipulation
11 scripts

Execution
3 scripts

Sequence analyses

Statistics
4 scripts

The most popular

ani.rb
Calculates the Average Nucleotide Ide

aai.rb
Calculates the Average Amino acid Ide

FastQ.split.pl
Splits a FastQ file into several FastQ fil
interposed sister reads using any ever

Taxonomy.silva2ncbi.rb
Re-formats Silva taxonomy into NCBI-l

AlphaDiversity.pl

Sequence similarity search

Statistics

BlastPairwise.AAsubs.pl
Counts the different AA substitutions in the best hit blast alignments, from a BLASTP pairwise format output (-outfmt 0 in BLAST+, -m 0 in legacy BLAST).

BlastTab.advance.bash
Calculates the percentage of a partial BLAST result. The value produced slightly subestimates the actual advance, due to un-flushed output and trailing queries that could be processed but generate no results.

BlastTab.seqdepth.pl
Estimates the sequencing depth of subject sequences.

ubject sequences. The values reported by this script may differ from
ise this script uses the aligned length of the read while
ed length of the subject sequence.

epth of subject sequences (genes or contigs) assuming a Zero-Inflated
or non-covered positions. It uses the corrected method of moments

Test example:
Average Amino acid Identity (AAI) between Mycoplasma genitalium (Bacteria) and Nanoarchaeum equitans (Archaea).
Parameters highlighted in blue have been changed from the defaul

aai.rb
Calculates the Average Amino acid Identity between two genomes.

If you use this script, also cite (Konstantinidis & Tiedje, 2005, JB

See also: ani.rb rbm.rb

Hide optional parameters

Seq1*
FastA file containing the genome 1 (proteins).
Open file | /Users/lmr3/Library/enveomics/en
Alternatively, you can supply the GI of a genome (nucleotides) with the format

Seq2*
FastA file containing the genome 2.
Open file | /Users/lmr3/Library/enveomics/en
Alternatively, you can supply the GI of a genome (nucleotides) with the format

aai.rb

Command:

```
ruby /Users/lmr3/.enveomics/enveomics-master/Scripts/aai.rb --seq1 /Users/lmr3/.enveomics/enveomics-
master/Tests/Mgen_M2288.faa --seq2 /Users/lmr3/.enveomics/enveomics-master/Tests/Nequ_Kin4M.faa --
```

Start time: Sun Dec 27 17:32:29 2015
Log: /var/folders/jt/_hhggq8d5sv7wssfh7bgmqlm0000gp/T/enveomics20151227-39351-1rpbleb

```
Temporal directory: /var/folders/jt/_hhggq8d5sv7wssfh7bgmqlm0000gp/T/d20151227-39388-1cm7hhc.
Creating databases.
  Reading FastA file: /Users/lmr3/.enveomics/enveomics-master/Tests/Mgen_M2288.faa
    File contains 493 sequences.
  Reading FastA file: /Users/lmr3/.enveomics/enveomics-master/Tests/Nequ_Kin4M.faa
    File contains 536 sequences.
Running one-way comparisons.
! One-way AAI 1: 32.78% (SD: 7.50%), from 492 proteins.
! One-way AAI 2: 33.04% (SD: 7.89%), from 530 proteins.
! Two-way AAI  : 31.40% (SD: 5.97%), from 178 proteins.
```

Running time: 36.723s.

Figure 2(on next page)

Example of a complete workflow primarily using tools from the enveomics collection applied to *Xanthomonas oryzae* genomes.

(A) The workflow uses the enveomics collection, Clustal Omega, and RAxML, to generate a phylogenetic tree based on the concatenated alignment of 105 single-copy essential genes. (B) In the resulting phylogeny, two clades form naturally corresponding to the pathovars oryzae (*Xoo*, left) and oryzicola (*Xoc*, right). Note that the tree is un-rooted, but the rooting point is suggested (vertex) based on phylogenomics of the genus (Rodriguez-R et al., 2012). The invariable sites were removed using Aln.cat.rb (35,386 sites removed) and the phylogeny was reconstructed using the remaining 552 informative sites. From the 105 detected essential genes, 22 were identical across all genomes and were excluded from the analysis. (C) A simplified version of the tree was produced by automatically pruning terminal nodes at a distance lesser than 0.01, resulting in a tree with 7 genomes (out of 17) with similar structure.

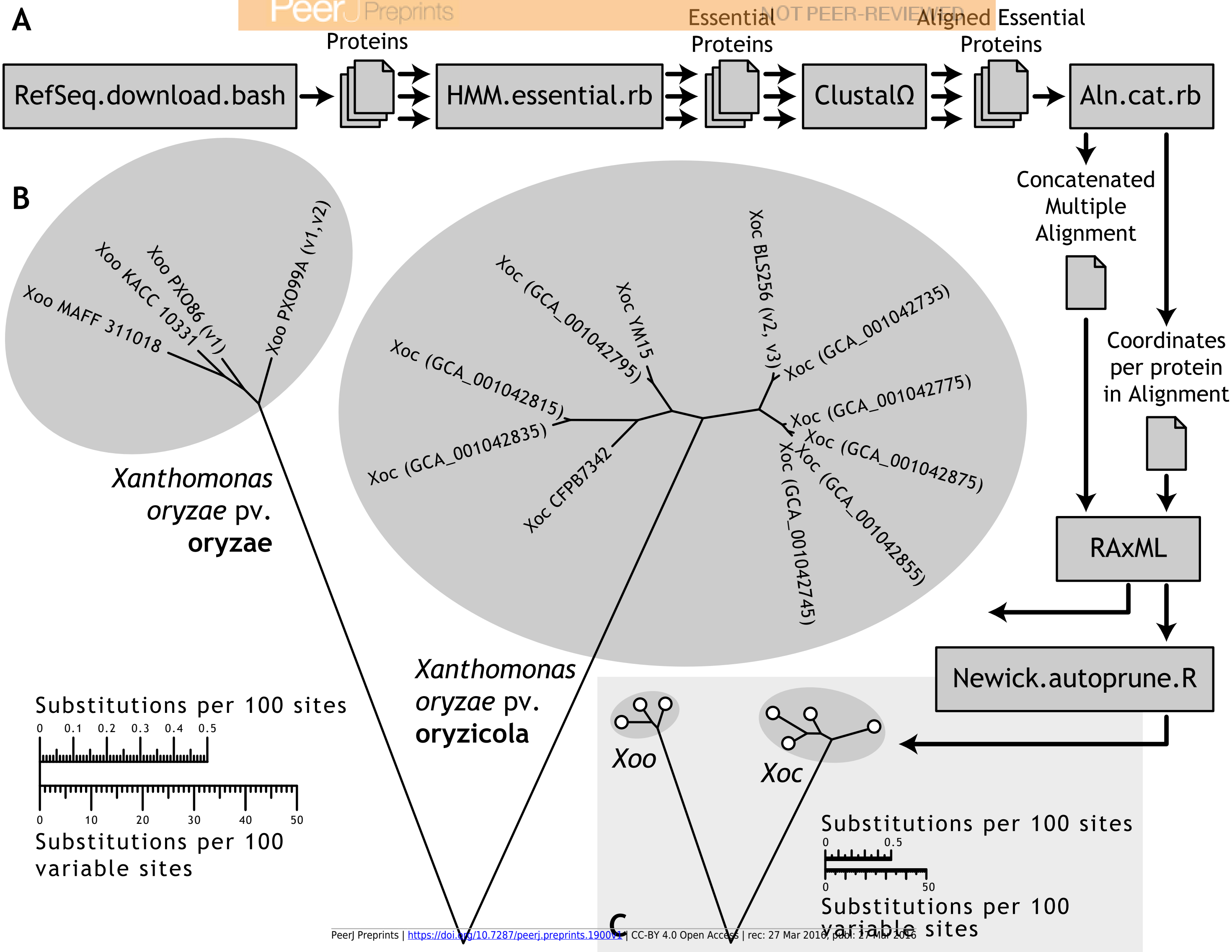
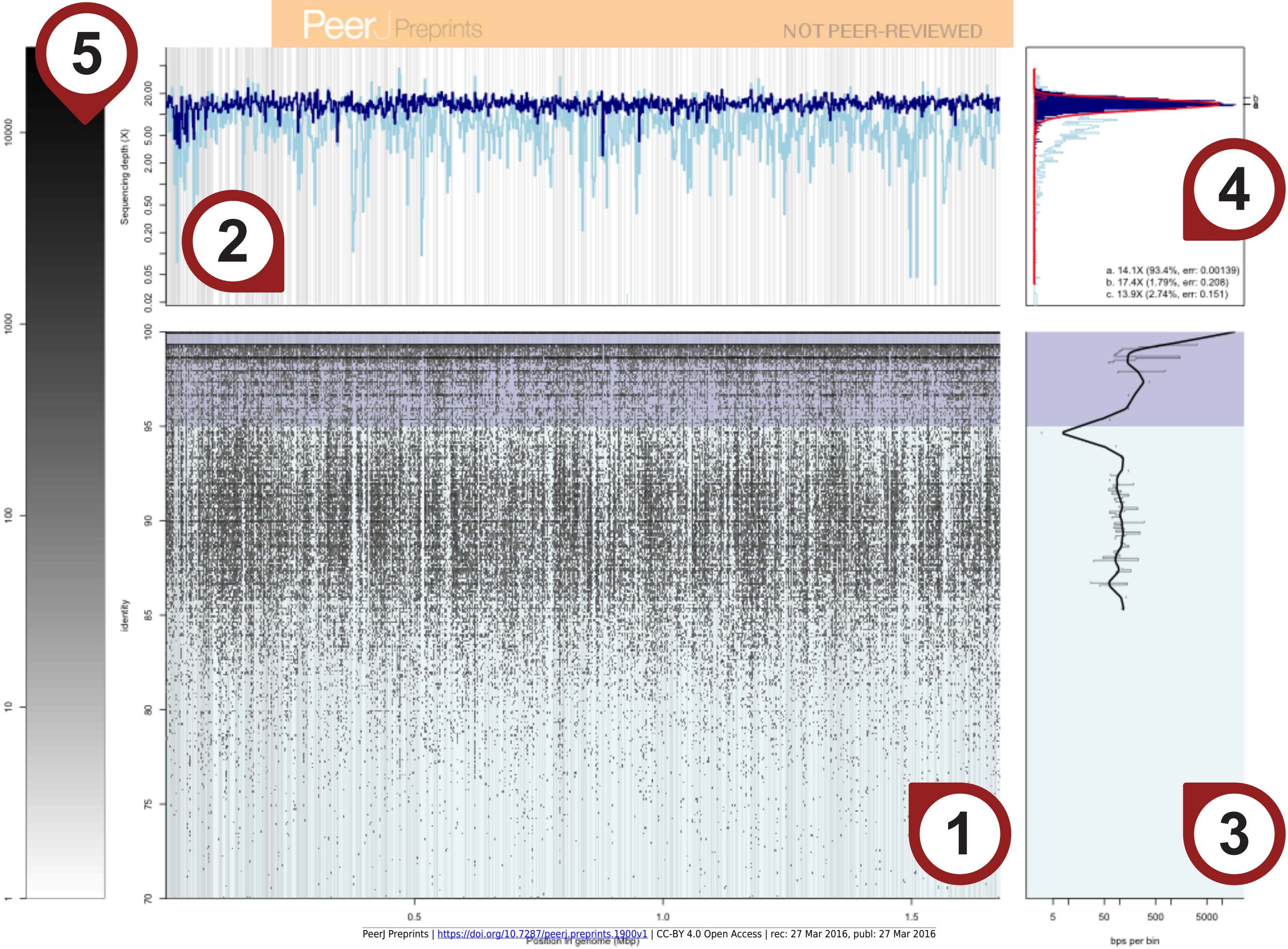

A
Proteins
Essential
Proteins
Aligned Essential
Proteins
RefSeq.download.bash
HMM.essential.rb
ClustalΩ
Aln.cat.rb
Concatenated
Multiple
Alignment
Coordinates
per protein
in Alignment
RAxML
Newick.autoprune.R
B
Xoo MAFF 311018
Xoo KACC 10331
Xoo PXO86 (v1)
Xoo PXO99A (v1,v2)
Xanthomonas
oryzae pv.
oryzae
Xoc (GCA_001042795)
Xoc YM15
Xoc BLS256 (v2, v3)
Xoc (GCA_001042735)
Xoc (GCA_001042815)
Xoc (GCA_001042835)
Xoc CFPB7342
Xoc (GCA_001042775)
Xoc (GCA_001042875)
Xoc (GCA_001042855)
Xoc (GCA_001042745)
Xanthomonas
oryzae pv.
oryzicola
Substitutions per 100 sites
0 0.1 0.2 0.3 0.4 0.5
0 10 20 30 40 50
Substitutions per 100
variable sites
C
Xoo
Xoc
Substitutions per 100 sites
0 0.5
0 50
Substitutions per 100
variable sites

Figure 3(on next page)

Example of a fragment recruitment plot.

This figure showcases the result of processing a BLAST search of metagenomic short sequencing reads (150 bp long in this case; each matching read is represented by a dot in main panel 1) against a population genome sequence assembled/binned from the same metagenome (X-axis). The tabular BLAST result was parsed using BlastTab.catsbj.pl, and graphical representation was generated with the BlastTab.recplot2.R. The circled numbers 1 through 5 denote the distinct panels of the layout: **(1)** Main panel representing the reads recruited, placed by location (X-axis) and identity (Y-axis). **(2)** Sequencing depth across the reference, in logarithmic scale. Bars at the bottom represent regions without mapping reads (sequencing depth of zero). **(3)** Identity histogram of mapping reads (light gray) and smoothed spline (black), in logarithmic scale. **(4)** Sequencing depth histogram. Peaks from values above 95% identity are automatically identified as skewed normal distributions (red), with centrality measures, percentage of the reference length, and fit error (bottom-right legend) reported for each peak (marked in the right edge). **(5)** Color scale for the number of stacked reads per 2-dimensional bin in panel 1. The background of panels 1 and 3, and the line colors in panels 2 and 4, correspond to matches with identity above (dark blue) and below (light blue) a user-defined cutoff. By default, the identity cutoff is set to 95%, corresponding to the species boundary (Konstantinidis & Tiedje, 2005). See also (Rodriguez-R & Konstantinidis, 2014a) for additional discussion.


5
4
2
1
3
10000
1000
100
10
1
Sequencing depth (X)
20.00
5.00
2.00
0.50
0.20
0.05
0.02
identity
100
95
90
85
80
75
70
0.5
1.0
1.5
Position in genome (Mbp)
b
a
a. 14.1X (93.4%, err: 0.00139)
b. 17.4X (1.79%, err: 0.208)
c. 13.9X (2.74%, err: 0.151)
5
50
500
5000
bps per bin