GEnView: a gene-centric, phylogeny-based comparative genomics pipeline for bacterial genomes and plasmids

(article starts on next page)

OXFORD

Genome analysis

GEnView: a gene-centric, phylogeny-based comparative genomics pipeline for bacterial genomes and plasmids

Stefan Ebmeyer[1,2], Roelof Dirk Coertze[1,2], Fanny Berglund [1,2], Erik Kristiansson[1,3] and D. G. Joakim Larsson [1,2,*]

[1]Center for Antibiotic Resistance Research, University of Gothenburg, 41346 Gothenburg, Sweden, [2]Department of Infectious Diseases, Institute of Biomedicine, University of Gothenburg, 41346 Gothenburg, Sweden and [3]Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, 41258 Gothenburg, Sweden

*To whom correspondence should be addressed.

Associate Editor: Can Alkan

Abstract

Summary: Comparing genomic loci of a given bacterial gene across strains and species can provide insights into their evolution, including information on e.g. acquired mobility, the degree of conservation between different taxa or indications of horizontal gene transfer events. While thousands of bacterial genomes are available to date, there is no software that facilitates comparisons of individual gene loci for a large number of genomes. GEnView (Genetic Environment View) is a Python-based pipeline for the comparative analysis of gene–loci in a large number of bacterial genomes, providing users with automated, taxon-selective access to the >800.000 genomes and plasmids currently available in the NCBI Assembly and RefSeq databases, and is able to process local genomes that are not deposited at NCBI, enabling searches for genomic sequences and to analyze their genetic environments through the interactive visualization and extensive metadata files created by GEnView.

Availability and implementation: GEnView is implemented in Python 3. Instructions for download and usage can be found at https://github.com/EbmeyerSt/GEnView under GLP3.

Contact: joakim.larsson@fysiologi.gu.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

With the advancement of sequencing technologies and their increasing efficiency, the number of sequenced bacterial genomes in publicly available databases has increased rapidly during the last two decades (Land *et al.*, 2015). The NCBI Assembly database alone currently contains (November 2021) more than 800.000 bacterial assemblies (Kitts *et al.*, 2016) , which are used by researchers around the world to investigate questions across a variety of fields, from bacterial evolution to public health, including pathogen virulence and antibiotic resistance (Ebmeyer *et al.*, 2021).

Though many comparative genomics tools are available to date for the analysis of bacterial replicons, such as GeneCO (Jung *et al.*, 2019), BRIG (Alikhan *et al.*, 2011) or MAUVE (Darling *et al.*, 2004), most are developed for comparing full genomes and detecting evolutionary events such as insertions, deletions or rearrangements across those genomes, and are therefore limited in the number of genomes that are feasible to compare. Furthermore, these tools are not developed for comparative analysis of particular gene loci, which may be more suitable when researching e.g.

the evolutionary history or taxonomic distribution of single bacterial genes. Such comparative analysis may require retrieval and extensive pre-processing of data, which is often a non-trivial undertaking.

Here, we present GEnView (Genetic Environment View), a fully automated, gene-centric pipeline for comparing genomic regions of several kilobasepairs that combines multiple bioinformatics tools and resources into a workflow that enables visual comparison of gene loci across hundreds of genomes stored locally or in the NCBI Assembly/RefSeq databases. GEnView identifies and processes user specified target genes from either user-provided nucleotide sequences or the genome and plasmid sequences stored in the NCBI Assembly/RefSeq database, creating an interactive visualization of all identified genes, including their phylogeny and their genetic environment. The main computational steps are parallelized which greatly reduces the run-time, making it possible to annotate several gene–loci at once. GEnView is therefore highly suitable for addressing questions about the evolutionary history of prokaryotic genes, such as the horizontal transfer and origins of mobile genes, or their potential (future) spread to other taxa.
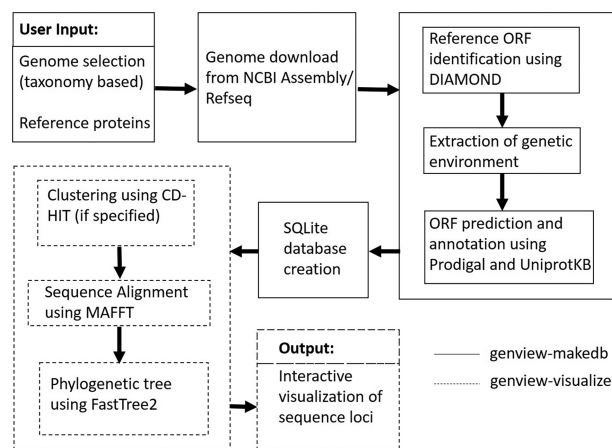
Fig. 1. Flow chart of GEnView

2 Implementation

GEnView consists of two main steps: genome processing and database creation, followed by phylogenetic analysis and visualization of the sequences.

In the first step, plasmids and genome assemblies (or only one of the two, based on user input) of a user defined taxon (e.g. *Enterobacter*) or user provided accession numbers are retrieved from the NCBI Assembly/NCBI RefSeq database. The selection of taxa above genus level is made possible through integration of information from the NCBI Taxonomy database (Federhen, 2012). The downloaded genomes are then searched for the user-provided reference genes at user-defined identity and coverage threshold using DIAMOND blastx v0.9.24.125 (Buchfink *et al.*, 2015). Once a reference gene is found, the gene and its environment up to 20 kb upstream and downstream are extracted. Prodigal v2.6.3 (Hyatt *et al.*, 2010) is then used to identify ORFs, which are searched against a custom version of the Uniprot knowledge base (clustered at 95% using CD-HIT v4.7 (Li and Godzik, 2006), hypothetical proteins removed, downloaded January 2019) using DIAMOND at a user-defined identity threshold. The user can also choose to search all ORFs against a custom database containing non-redundant sequences of insertion sequences and insertion sequence common region elements. All results are then saved to a database using SQLite3.

In the second step, the user simply provides the name of the gene/gene group to analyze, which is extracted from the database and complementary files. The respective genes and their genetic environment are saved as file in FASTA format. In order to reduce the amount of sequences to visualize, unique sequences are extracted from the previously created FASTA file based on their annotation profile (though the user may choose to display all sequences). These sequences are aligned using MAFFT v7.3.10 (Katoh *et al.*, 2002) and a phylogenetic tree based on this alignment is calculated using FastTree v2.1.9 (Price *et al.*, 2009) (gtr model, CAT approximation). Each sequence is then visualized together with the respective annotations from the previously created database and linked to its respective node in the phylogeny in an interactive visualization, which can be viewed in a browser and be exported as high-resolution image. For each visualized sequence, the gene names, taxon and the sequences unique id are also shown in the phylogeny, enabling their identification in the metadata files (containing genome id, annotated ORFs with name, start and end positions on the sequence, as well as the respective ORFs' sequence) for in-detail manual analysis. The workflow is illustrated in Figure 1.

As a demonstration, we used GEnView to extract and visualize the genetic environment of the antibiotic resistance gene FOX-1 in *Aeromonas* species genomes (*n* = 881). By using 1,2,4,8,16 and 32 parallel running processes (on a 48 core 2.20 GHz Intel(R) Xeon(R) CPU E5-2650 v4 with 252 GB RAM), GEnView was able to download all sequences and finish all computations and visualizations

within 5.9, 3.1, 1.8, 1.2, 1 and 0.6 h respectively, using an identity cutoff toward the FOX-1 reference protein of 80% (Supplementary Figure S1). In addition to the visualization, the exact sequences, annotations and more meta-information are provided for further in-depth analysis (e.g. co-occurrence with different antibiotic resistance genes or transposases).

3 Conclusion

GEnView provides automated, selective access to thousands of genomes and plasmids from the NCBI Assembly and RefSeq databases. Parallelized, automated annotation and phylogeny-based visualization of genome sequences allow for visual comparison of gene–loci from several hundreds of genomes at the same time and greatly reduces the amount of manual work that previously was necessary to perform such tasks. The interactive visualization together with the generated metadata, alignments and phylogenetic trees allow for in depth analysis of the gene locus of interest.

Data availability

The FOX-1 reference is available at https://card.mcmaster.ca/ontology/38555; the 881 *Aromonas* genomes are available at https://www.ncbi.nlm.nih.gov/assembly/?term=Aeromonas and easily downloadable using GEnView by following the GEnView tutorial at https://github.com/EbmeyerSt/GEnView/wiki.

References

Alikhan,N.F. *et al.* (2011) BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, **12**, 402.

Buchfink,B. *et al.* (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

Darling,A.C.E. *et al.* (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.

Ebmeyer,S. *et al.* (2021) A framework for identifying the recent origins of mobile antibiotic resistance genes. *Commun. Biol.*, **4**, 1–10.

Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res*, **40**, D136–D143.

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Jung,J. *et al.* (2019) GeneCo: a visualized comparative genomic method to analyze multiple genome structures. *Bioinformatics*, **35**, 5303–5305.

Katoh,K. *et al.* (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.

Kitts,P.A. *et al.* (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.*, **44**, D73–D80.

Land,M. *et al.* (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics*, **15**, 141–161.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Price,M.N. *et al.* (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, **26**, 1641–1650.