Bioinformatics, 2020, 1–3
doi: 10.1093/bioinformatics/btaa212
Advance Access Publication Date: 28 March 2020
Applications Note

Genome analysis

Gcluster: a simple-to-use tool for visualizing and comparing genome contexts for numerous genomes

Xiangyang Li [1,2,*,†], Fang Chen[3,†] and Yunpeng Chen[4]

[1]Department of Environmental Science and Engineering, Fudan University, Shanghai 200433, China, [2]College of Environment and Life Sciences, Kaili University, Kaili 556011, China, [3]School of Basic Medical Sciences, Shanxi Medical University, Taiyuan 030001, China and [4]Key Laboratory of Urban Agriculture of Ministry of Agriculture School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai 200240, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Peter Robinson

Received on December 15, 2019; revised on March 12, 2020; editorial decision on March 21, 2020; accepted on March 23, 2020

Abstract

Motivation: Comparing the organization of gene, gene clusters and their flanking genomic contexts is of critical importance to the determination of gene function and evolutionary basis of microbial traits. Currently, user-friendly and flexible tools enabling to visualize and compare genomic contexts for numerous genomes are still missing.

Results: We here present Gcluster, a stand-alone Perl tool that allows researchers to customize and create high-quality linear maps of the genomic region around the genes of interest across large numbers of completed and draft genomes. Importantly, Gcluster integrates homologous gene analysis, in the form of a built-in orthoMCL, and mapping genomes onto a given phylogeny to provide superior comparison of gene contexts.

Availability and implementation: Gcluster is written in Perl and released under GPLv3. The source code is freely available at https://github.com/Xiangyang1984/Gcluster and http://www.microbialgenomic.com/Gcluster_tool.html. Gcluster can also be installed through conda: 'conda install -c bioconda gcluster'.

Contact: lixiangyang@fudan.edu.cn

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

1 Introduction

High-throughput sequencing technology has become a routine practice for microbiology research. Linear visualization and comparison of the genomic region flanking given genes and gene clusters of interest across large numbers of genomes could provide robust clues to investigate its underlying biological properties and expound the genome evolutionary history. However, it is difficult to accomplish this using existing tools (Aziz *et al.*, 2008; Darling *et al.*, 2004; Oberto, 2013; Pejaver *et al.*, 2012; Sullivan *et al.*, 2011). Most of the current available tools allow visualization of only a few genomes. Although GeneSpy supports visualization of a number of genomes, it depends on association of gene annotations among genomes and a third-party tool, iTOL, to map genomes onto a phylogenetic tree for comparison of gene contexts (Garcia *et al.*, 2019). Some tools cannot customize input genomes (e.g. draft and private genomes) or have limited options when it comes to customizing figures [e.g. output scalable vector graphics (SVG) formats, gene label edition]. Finally, some tools are difficult to install and not user-friendly for researchers who do not have profound bioinformatics knowledge. Accordingly, we have developed Gcluster, a stand-alone Perl application and an easy-to-use genome comparison visualizer. It enables users to explore the genomic contexts flanked gene of interest for large numbers of finished and draft genomes.

2 Implementation

Gcluster was written in Perl and requires only Perl Modules, BLAST+ and MCL, which makes for simple installation. Additionally, it is available as a bioconda package with all the dependencies preinstalled (Gruning *et al.*, 2018). Gcluster needs two mandatory inputs: (i) a directory containing annotated genomes in GenBank format and (ii) a list of genes of interest, in which each row contains a locus tag of the gene of interest from each of the genomes being analyzed. The most striking feature is that Gcluster allows visualization of hundreds of genomes with low memory consumption, and it supports both completed and draft genomes (e.g. Fig. 1). Gcluster has a Perl script (interested_gene_generation.pl) that can auto-generate a list of locus tags of interested genes based on a local blastp analysis. Alternatively, the locus tag of the gene of interest can be found directly using keywords in GenBank files or in BLAST outputs from online sources (e.g. NCBI, RAST).

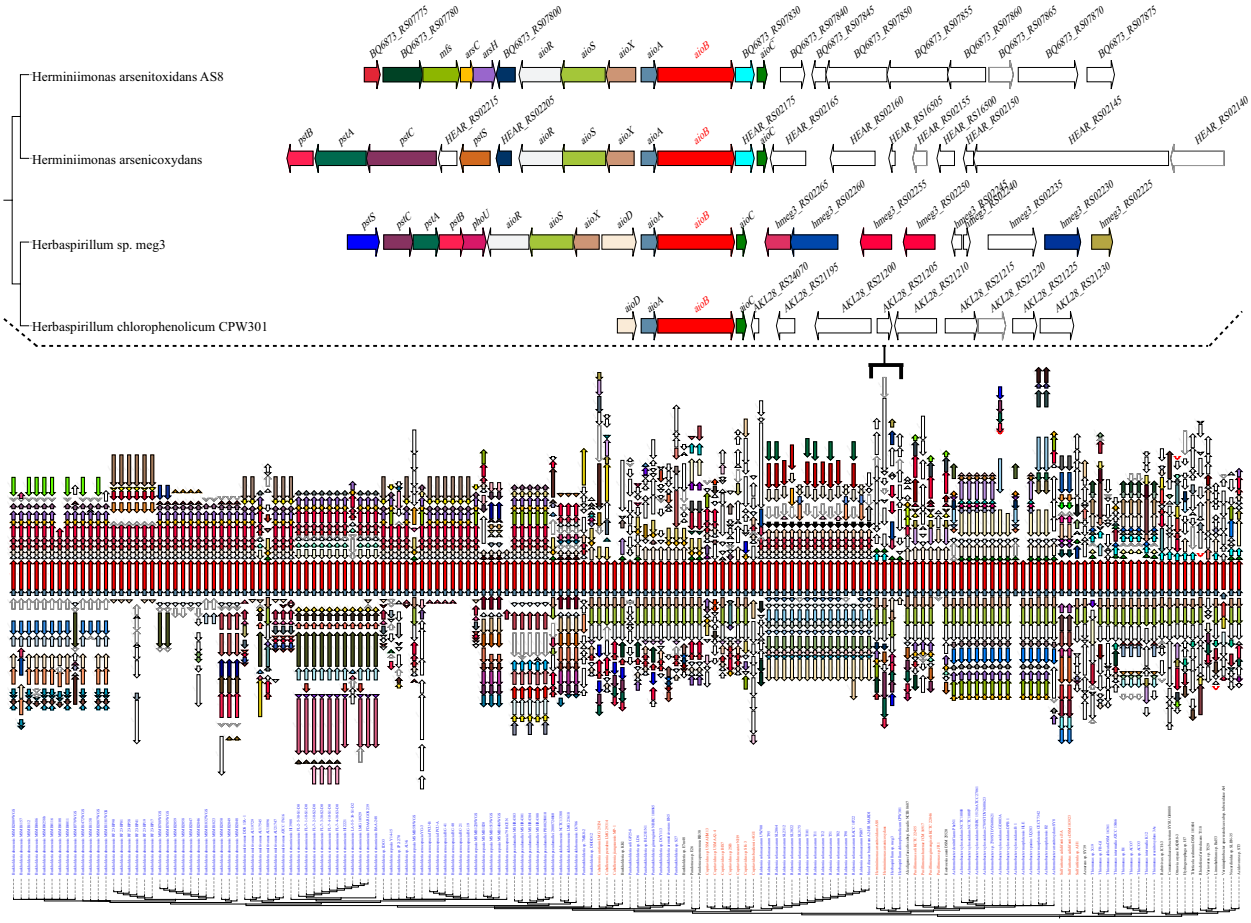

Fig. 1. Example of visualization and mapping of the arsenite oxidase large subunit gene (*aioB*, in red) and its flanking genome contexts (20 genes) to a 16S rRNA gene phylogenetic diagram by Gcluster in 160 β-*Proteobacteria* genomes. This figure consists of two images from two runs through Gcluster. Genome contexts of all strains are shown at the bottom, and a small branch of four genomes is shown magnified at the top. Genomic contexts have been reorientated around *aioB*, and parts of flanking genes were found to be missing from certain draft genomes. Homologous gene clusters are filled in different colors; unique genes and pseudo genes are in white, with black and deep-gray borders, respectively. The nodes of different genera are alternately colored red and blue using Adobe Illustrator. These results show that the aio cluster has two types (distribution of aioXSR in their upstream and non-upstream areas), and the conservation of the flanking gene contents is very low among and within certain genera. (Color version of this figure is available at *Bioinformatics* online.)

Gcluster allows the use of multiple threads and works as follows (Supplementary Fig. S1). A collection of TFT files (a seven-column tab-delimited feature table) is extracted from the annotated genomes. Each TFT holds the gene information for an individual genome, in which each line represents one gene set and consists of its start position, end position, gene type, locus tag, product function and scaffold (e.g. chromosome, contig) accession number; for pseudogene, it is flagged with 'pseudo' qualifier in seventh column. Then, a collection of sub-TFT files is generated according to the genes of interest and their number of flanking customized genes. Finally, Gcluster reads sub-TFT files and produces figures in PNG-/SVG using Perl Modules (GD and GD::SVG). As shown in Figure 1, genomic contexts are reorientated around the genes of interest.

Gcluster employs a built-in OrthoMCL tool (Li *et al.*, 2003) for homologous analysis of genes in sub-TFT files. Each set of homologous gene clusters is filled in a different color (Fig. 1). If enough RGB colors are provided, there is no limit to the number of genomes and the customized number of genes flanking the gene of interest that can be shown. Alternatively, Gcluster also accepts results of homologous groups from other homologous genes detection tools. To explore the evolutionary history of these gene contexts, Gcluster allows auto-mapping of the genomic contexts into a phylogenetic diagram provided by the user (Fig. 1).

Gcluster is sufficiently flexible to allow customization of its figures. The user can adjust the margins, the interval between two neighboring genomes, the text size, the gene length and width, the scale, the rotation angle of gene labels, the number of genes flanking the gene of interest and the order of genomic contexts. Among each set of homologous proteins, if a gene is annotated with a name X, the user has the option of displaying all other genes labeled X. The user can also modify the gene label by directly editing the locus tag in sub-TFT files or in homologous genes analysis result. Gcluster also allows convenient and quick highlighting of specific types of functional genes of interest by treating each type of highlighted gene as homologous gene cluster.

3 Conclusion

To our knowledge, Gcluster is the first genome comparison visualizer that enables linear visualization of numerous genomes and integrates homologous gene analysis and auto-binding genome contexts to a given phylogeny to improve the comparison of gene contexts. Gcluster is also highly customizable and easy to use. It can quickly generate high-quality linear figures showing gene contexts in PNG-/SVG format. These features make Gcluster a more powerful software package than other currently available visualization tools.

Funding

This work was supported by grants from the National Natural Science Foundation of China [No. 21507012 and 21966015], the China Postdoctoral

Science Foundation [No. 2015M570329], the Science and Technology Foundation of Guizhou Province [No. (2019)1287], the Joint foundation of GuiZhou Province Science and Technology Commission of China [No. LH(2017)7176], the National Science Foundation of Guizhou Provincial Department of Education of China [No. KY(2015)395] and Startup Foundation for Doctors of Shanxi Medical University [No. 03201503].

Conflict of Interest: none declared.

References

Aziz,R.K. *et al*. (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

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

Garcia,P.S. *et al*. (2019) GeneSpy, a user-friendly and flexible genomic context visualizer. *Bioinformatics*, **35**, 329–331.

Gruning,B. *et al*. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods*, **15**, 475–476.

Li,L. *et al*. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*., **13**, 2178–2189.

Oberto,J. (2013) SyntTax: a web server linking synteny to prokaryotic taxonomy. *BMC Bioinformatics*, **14**, 4.

Pejaver,V.R. *et al*. (2012) GeneclusterViz: a tool for conserved gene cluster visualization, exploration and analysis. *Bioinformatics*, **28**, 1527–1529.

Sullivan,M.J. *et al*. (2011) Easyfig: a genome comparison visualizer. *Bioinformatics*, **27**, 1009–1010.