bioRxiv preprint doi: https://doi.org/10.1101/800920; this version posted October 17, 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 The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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-NC 4.0 International license.

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Application note

Chromatiblock: scalable whole-genome visualization of structural differences in prokaryotes

Mitchell John Sullivan[1] and Harm van Bakel[1*]

[1]Department of Genetics and Genomic Sciences, Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, United States of America

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Chromatiblock is a Python application for visualizing the presence, absence and arrangement of syntenic blocks across large numbers of complete bacterial genomes.

Availability: Chromatiblock is freely available under a GPL license, for macOS, GNU/Linux and and Microsoft Windows from https://github.com/mjsull/chromatiblock/

Contact: harm.vanbakel@mssm.edu

1 Introduction

Visualizing structural variation between complete prokaryotic genomes is important for identifying the genetic basis of strain differences. This is generally accomplished by displaying the results of serial pairwise comparisons or multiple alignments in linear or circular layouts. Serial pairwise comparisons can be created using tools such as Easyfig (Sullivan, et al., 2011) or GenoplotR (Guy, et al., 2010) that display linear pairwise comparisons between two or more genomes. However, genomic loss, gain and structural variation can only be directly inferred for adjacent genomes. Multiple alignment visualization tools such as Mauve (Darling, et al., 2004) solve this issue by representing syntenic regions as linear blocks and using lines to connect blocks across genomes. In large figures this can result in crisscrossing lines that are often difficult to interpret. Alternatively, ring plots, such as those created by the BLAST ring image generator (BRIG) (Alikhan, et al., 2011) or the CGView Comparison Tool (CCT) (Grant, et al., 2012) use a series of concentric circles to display the presence or absence of genomic regions across multiple genomes. These regions are ordered according to a reference, and as such they convey no information about their arrangement in each non-reference genome. Alternatively, Circos (Krzywinski, et al., 2009) plots show genomes around the outside edge of a circle and represents regions of similarity as arcs, but this approach scales poorly as the number of arcs increases exponentially with each genome. Representing many genomes as circles can also result in large size differences between inner and outer rings, further complicating interpretation.

Here we present Chomatiblock, an application for visualizing syntenic blocks in multiple genome alignments. Chromatiblock was designed to create a linear visual representation of structural variation, including the presence and absence of genomic regions in an easy-to-comprehend and scalable manner, adding to the visualization options available for alignments of large numbers of complete genomes.

2 Implementation

Chromatiblock is a Python script available under a GPL license and runs on macOS, GNU/Linux and Microsoft Windows operating systems. Chromatiblock can be used to create publication-quality images displaying arrangement and presence of syntenic blocks. The results can also be viewed as an interactive webpage that allows the user to zoom, pan and highlight shared regions across genomes.

Chromatiblock takes an extended multi-fasta alignment (MAF) files as input, which can be generated by a variety of multi-genome alignment programs (Angiuoli and Salzberg, 2011; Minkin and Medvedev, 2019). Alternatively, when provided with FASTA-formatted files for a set of genomes of interest, Chromatiblock can run Sibelia (Minkin, et al., 2013) to automatically generate the required input. Once syntenic blocks have been identified in the MAF file, Chromatiblock will generate a dual-panel layout consisting of a global alignment view and a detailed view of regions that differ between genomes. The global alignment view shows the arrangement of core blocks (i.e. syntenic regions found once in all genomes) in the alignment and how non-core blocks (i.e found in 2 or more genomes) and unique sequences (i.e. found in a single genome) are arranged relative to the core blocks. Core blocks are aligned according to their arrangement in the first genome. The color the core blocks for each genome is determined by its position. Between each core block there exists a

bioRxiv preprint doi: https://doi.org/10.1101/800920; this version posted October 17, 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-NC 4.0 International license.

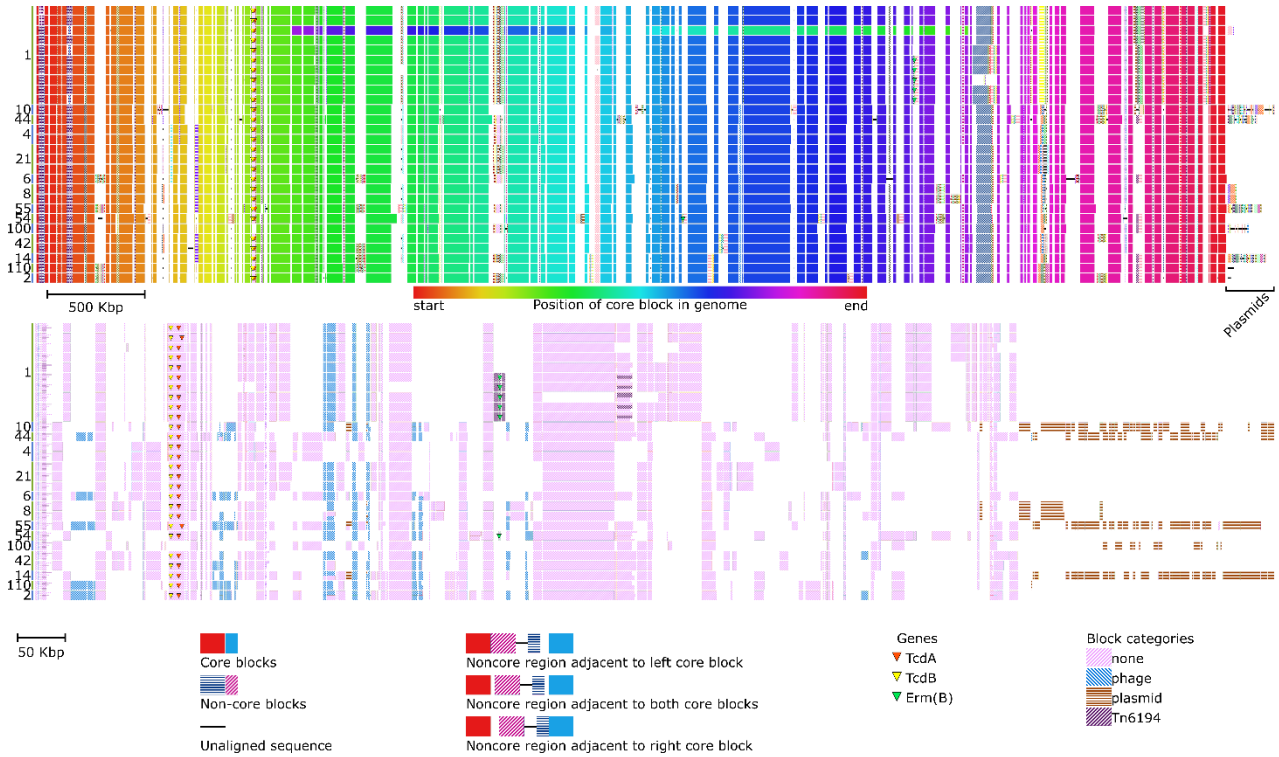

Fig. 1. Chromatiblock visualization of 28 *Clostridium difficile* genomes. MLST of each isolate is indicated on the left. **A)** Global alignment view. Core blocks across genomes (rows) are visualized as vertically aligned solid rectangles that are colored according to their position in the genome. Non-core blocks are visualized as patterned rectangles, with each block represented by a unique combination of pattern and color. Finally, sequences unique to a single genome are depicted as solid black lines. **B)** Alignment difference view. Each genome is represented as a row and each non-core block is assigned a column in the order they are most commonly found in the genome. Presence of each non-core block is shown as a patterned rectangle in the genomes row. As non-core blocks may be present more than once, duplicates are shown by splitting the blocks according to repeat number.

combination of non-core blocks and unique sequence. This combination is grouped and positioned between the two core blocks to which they are adjacent. In instances where the group cannot be placed between its two adjacent core blocks it is placed arbitrarily next to one of the core blocks to which it is adjacent. This is indicated by removing the gap between core and non-core blocks. An example of a global alignment of 28 complete *C. difficile* genomes is shown in **Fig. 1A**. A large inversion can be observed in the third isolate from the top, indicated by a difference in ordering of core block colors relative to the reference. Plasmids, found in 9 genomes, consist entirely of non-core and unique blocks. They are positioned on the right side of the figure. The presence or absence of specific user-provided gene sequences can also be indicated by distinct gene symbols and are automatically annotated using BLASTx. In the example, six isolates contain a transposon carrying the *erm(B)* gene, encoding a 23S rRNA methyltransferase that confers resistance to erythromycin. The *erm(B)*gene is also present in an ST54 isolate but located on a novel transposon and inserted elsewhere in the genome (**Fig. 1A**).

The alignment difference view shows the presence and absence of all non-core blocks. Chromatiblock can use BLAST+ to categorize and color each non-core block based on a user-provided reference database of nucleotide or amino acid FASTA files. Categories can also be assigned based on the size of the contig in which the non-core block is found. The example in **Fig 1B** shows that the main *C. difficile* pathogenicity locus (PaLoc) that contains the genes encoding the TcdA enterotoxin and TcdB cytotoxin, has been lost in the ST100 isolate. Plasmids carried by *C. difficile* are very chimeric, with large regions being shared, but with only the two MLST8 isolates carrying identical plasmids.

In conclusion, Chromatiblock allows users to quickly and easily create publication-quality figures showing structural changes and genetic diversity at the whole genome level.

Funding

This work was supported by a grant from the National Institute of Allergy and Infectious Diseases (NIAD R01 AI119145).

Conflict of Interest: none declared.

References

Alikhan, N.-F., *et al.* BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 2011;12(1):402.

Angiuoli, S.V. and Salzberg, S.L. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics (Oxford, England)* 2011;27(3):334-342.

Darling, A.C.E., *et al.* Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research* 2004;14(7):1394-1403.

Grant, J.R., Arantes, A.S. and Stothard, P. Comparing thousands of circular genomes using the CGView Comparison Tool. *BMC Genomics* 2012;13(1):202.

Guy, L., Roat Kultima, J. and Andersson, S.G.E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 2010;26(18):2334-2335.

Krzywinski, M., *et al.* Circos: An information aesthetic for comparative genomics. *Genome Research* 2009;19(9):1639-1645.

Minkin, I. and Medvedev, P. Scalable multiple whole-genome alignment and locally collinear block construction with SibeliaZ. *bioRxiv* 2019:548123.

bioRxiv preprint doi: https://doi.org/10.1101/800920; this version posted October 17, 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-NC 4.0 International license.

Minkin, I., *et al.* Sibelia: A Scalable and Comprehensive Synteny Block Generation Tool for Closely Related Microbial Genomes. In. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 215-229.

Sullivan, M.J., Petty, N.K. and Beatson, S.A. Easyfig: a genome comparison visualizer. *Bioinformatics* 2011;27(7):1009-1010.