WEB TOOL

C-Sibelia: an easy-to-use and highly accurate tool for bacterial genome comparison [version 1; referees: 2 approved]

Ilya Minkin[1], Hoa Pham[2], Ekaterina Starostina[1], Nikolay Vyahhi[1], Son Pham[3]

[1]Bioinformatics Institute, St. Petersburg, Russian Federation
[2]GNT Incorporation, HoChiMinh City, Vietnam
[3]Computer Science and Engineering, University of California San Diego, La Jolla, 92092, USA

v1 **First published:** 25 Nov 2013, **2**:258 (doi: 10.12688/f1000research.2-258.v1)
Latest published: 25 Nov 2013, **2**:258 (doi: 10.12688/f1000research.2-258.v1)

Abstract

We present C-Sibelia, a highly accurate and easy-to-use software tool for comparing two closely related bacterial genomes, which can be presented as either finished sequences or fragmented assemblies. C-Sibelia takes as input two FASTA files and produces: (1) a VCF file containing all identified single nucleotide variations and indels; (2) an XMFA file containing alignment information. The software also produces Circos diagrams visualizing high level genomic architecture for rearrangement analyses. C-Sibelia is a part of the Sibelia comparative genomics suite, which is freely available under the GNU GPL v.2 license at http://sourceforge.net/projects/sibelia-bio. C-Sibelia is compatible with Unix-like operating systems. A web-based version of the software is available at http://etool.me/software/csibelia.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 25 Nov 2013	✓ report	✓ report

1 **Jeffrey McLean**, J. Craig Venter Institute USA

2 **Loren J Hauser**, Oak Ridge National Laboratory USA

Discuss this article

Comments (1)

Corresponding author: Son Pham (kspham@cs.ucsd.edu)

How to cite this article: Minkin I, Pham H, Starostina E *et al.* **C-Sibelia: an easy-to-use and highly accurate tool for bacterial genome comparison [version 1; referees: 2 approved]** *F1000Research* 2013, **2**:258 (doi: 10.12688/f1000research.2-258.v1)

Grant information: This work was partially supported by Russian Foundation for Basic Research RFBR (grant 14-04-31926) and V-P foundation.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: No competing interests were disclosed.

First published: 25 Nov 2013, **2**:258 (doi: 10.12688/f1000research.2-258.v1)

Introduction

The development of inexpensive genome sequencing technologies and efficient assembly methods has revolutionized the study of bacterial genomes, which are being sequenced and assembled on a daily basis. When an assembly is available, the most common first task is to compare it against a reference genome (or another assembly, if no such genome is available) in order to find genetic differences between the newly assembled and reference genomes. This analysis is critical to understand genetic factors that determine certain phenotypes of the isolates.

We present Comparative Sibelia software (C-Sibelia) for the comparison of two bacterial genomes in the form of complete sequences or draft assemblies. C-Sibelia is able to compare genomes in the presence of rearrangements and duplications. C-Sibelia takes as input two FASTA files (the assembly and reference files; if the reference genome is not available, it can be substituted by another draft assembly) and produces: (1) a VCF file containing all identified single nucleotide variations (SNVs) and indels; (2) annotation of these variants by SnpEff; (3) an XMFA[1] file containing alignment information. The web-based version also produces a circular diagram visualizing the rearrangement of synteny blocks in two genomes.

The performance of C-Sibelia in detecting SNVs and indels is comparable to MUMmer and outperforms Mauve in terms of the false-positive rate. C-Sibelia is a part of the Sibelia comparative genomics suite, which is freely available under the GNU GPL v.2 license at http://sourceforge.net/projects/sibelia-bio. Users are encouraged to use the web-based version of C-Sibelia at http://etool.me/software/csibelia.

Methods

From synteny blocks to alignment

The task of finding SNVs and indels connects closely to the problem of whole-genome alignment. Unlike aligning two short DNA segments, aligning two genomes is more challenging because of the presence of rearrangements and repetitive elements. C-Sibelia addresses this problem by first decomposing genomes into synteny blocks, using the *iterative de Bruijn graph algorithm* described in Minkin *et al.*[2]. This step separates linear operations (indels, substitutions) from non-linear operations (rearrangements) and thus allows us to apply global alignment to multiple instances of each synteny block. C-Sibelia incorporates LAGAN[3], a global alignment tool, for aligning different instances of the same synteny block.

From alignment to variant calling. C-Sibelia then finds differences between two genomes (indels, SNVs, rearrangements) by analyzing the resulting synteny and alignment blocks. Regions in one genome not covered by synteny blocks are treated as indels. SNVs and small indels that lie within the regions covered by synteny blocks are reported by analyzing the alignment information produced by LAGAN. Identified variants are annotated by using snpEff[4]. The pipeline of C-Sibelia is described in the following seudocode.

Input: An assembly and a reference genome (in FASTA format).

Algorithm:

- Decompose the sequences into synteny blocks using Sibelia.
- Align instances of synteny blocks using LAGAN.
- Analyze the synteny block decomposition and alignment information.
  - Find indels in non-syntenic regions.
  - Find small indels and SNVs in aligned regions (using the alignment information produced by LAGAN).
  - Annotate the identified variants using SnpEff.
  - Select contigs containing multiple synteny blocks (i.e., rearranged contigs).

Output:

- All SNVs and indel variants, in a VCF file.
- Annotation of these variants produced by SnpEff[4].
- A picture in Circos format[5] for rearranged contigs and the reference genome.

Results

A simulated dataset

To evaluate the variant calling feature, we benchmarked C-Sibelia against Mauve[6] and MUMmer[7] on a simulated dataset, designed as follows.

From the complete genome of *Staphylococcus aureus* (*S. aureus*) NCTC 8325, we performed 10 deletions of random segments of size 2000 bp, and futher introduced 1000 SNVs in the resulting genome. We then generated five reversals and five translocations of random segments in the genome with size 10,000 bp each to evaluate the capability of these tools to perform an alignment in the presence of rearrangements. We obtained a *simulated assembly* of this newly *simulated genome* of 180 contigs; the distribution of contig length was similar to that of the RN4220 assembly reported in Dhanalakshmi *et al.*[8]. We further used C-Sibelia, Mauve and MUMmer to find variants in this simulated assembly and the original reference genome (NCTC 8325). Table 1 and Table 2 demonstrate that the performance of C-Sibelia in detecting variants is comparable to MUMmer and improves upon Mauve in terms of the false-positive rate. Figure 1 shows the Circos diagram of the rearranged contigs and the reference genome. The scripts and commands used for this benchmark are available in the Supplementary material.

A real dataset

The most common approach for comparing an assembly against a reference genome is to first align the assembly against the reference and then write in-house scripts to extract variants. C-Sibelia can achieve this task automatically and with high accuracy. We used C-Sibelia to reproduce the comparison of the *S. aureus* RN4220 assembly and the reference genome NCTC 8325, reported in Dhanalakshmi *et al.*[8] (the authors used MUMmer and in-house scripts for this comparison). Among 132 single nucleotide variants and four large deletions reported in Dhanalakshmi *et al.*[8], C-Sibelia

Table 1. **SNV calling on simulated data.**

Tool	True Positive	False Positive	False Negative
C-Sibelia	976	0	24
MUMmer	977	0	23
Mauve	991	78	9

Table 2. **Indel calling on simulated data.**

Tool	True Positive	False Positive	False Negative
C-Sibelia	9	0	1
MUMmer	9	0	1
Mauve	10	1	0

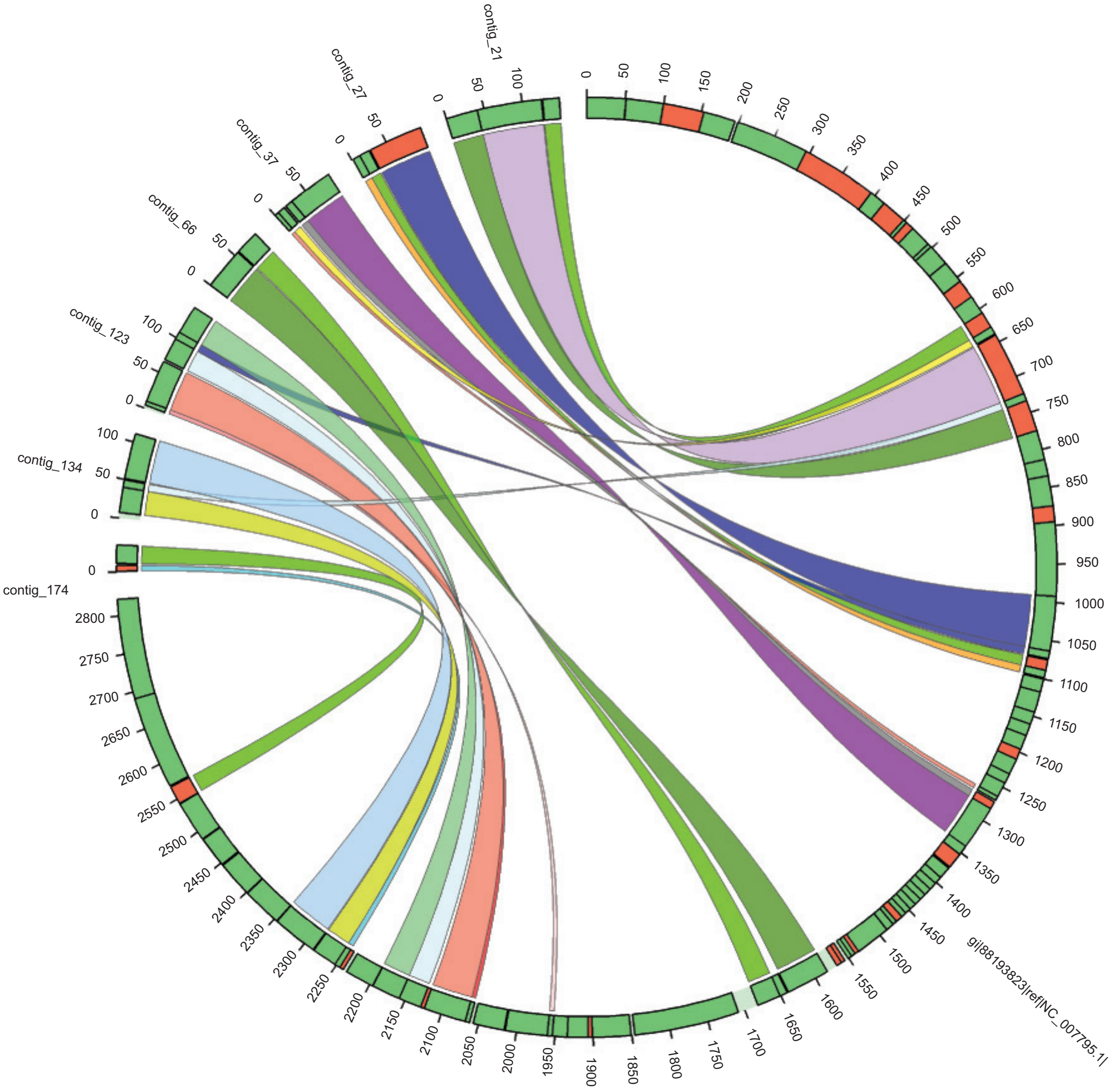

Figure 1. **A picture in Circos format for assembly sequences and the reference genome.** Only contigs with multiple synteny blocks rearranged differently in the genome are shown. Green and red bars depict the direction of synteny blocks on the positive and negative strands, respectively.

confirmed 121 SNVs and all four large deletions. C-Sibelia also reported six additional variants, which are also confirmed by BLAST[9]. The input data as well as the commands for generating these results are available in the Supplementary material.

The Etool Web-Server

The online version of C-Sibelia is available at http://etool.me/software/csibelia. The web form takes as input two FASTA files (one for the assembly and the other for the reference). The web form's parameters allow users to choose whether or not to annotate variants and display the Circos[5] picture for rearrangement analysis (see Figure 1). Results are delivered to registered users by a real time push notification mechanism[10,11].

Discussion

In this application note, we introduced C-Sibelia, a novel software for comparing two closely-related bacterial strains. Performance of C-Sibelia is comparable to MUMmer, and better than Mauve in terms of false positives rate. The web interface of C-Sibelia makes the task of comparing assemblies against a reference genome convenient for microbiologists, who do not want to go to the trouble of downloading and compiling the software. In the future, we plan to extend C-Sibelia to compare multiple genomes or draft assemblies as well as scale the software to larger genomes.

Author contributions

SP designed the project. IM, HP, ES, NV performed experiments and developed the computational tool. IM, SP wrote the manuscripts with critical input from NV, ES, HP. All authors read and approved the final manuscript.

Competing interests

No competing interests were disclosed.

Grant information

This work was partially supported by Russian Foundation for Basic Research RFBR (grant 14-04-31926) and V-P foundation.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We would like to thank Alla Lapidus, Ekaterina Chernyayeva, Ryan Tewhey, Xiao Li, Harish Nagarajan, Mikhail Kolmogorov, Dhanalakshmi Nair, and David Hernandez for many helpful discussions. We are indebted to Phillip Compeau for for many helpful editorial suggestions.

Supplementary material

Supplemenary material can be found online at http://goo.gl/jtLsPl and is permanently available at doi: 10.5281/zenodo.7577.

References

1. Brudno M, Poliakov A, Salamov A, *et al.*: **Automated whole-genome multiple alignment of rat mouse, and human.** *Genome Res.* 2004; **14**(4): 685–692.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Minkin I, Patel A, Kolmogorov M, *et al.*: **A scalable and comprehensive synteny block generation tool for closely related microbial genomes.** *arXiv preprint arXiv: 1307.7941*, 2013.
Publisher Full Text
3. Brudno M, Do CB, Cooper GM, *et al.*: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res.* 2003; **13**(4): 721–731.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Cingolani P, Platts A, Coon M, *et al.*: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3.** *Fly (Austin).* 2012; **6**(2): 80–92.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Krzywinski M, Schein J, Birol I, *et al.*: **Circos: an information aesthetic for comparative genomics.** *Genome Res.* 2009; **19**(9): 1639–1645.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Darling AE, Mau B, Perna NT: **progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement.** *PloS One.* 2010; **5**(6): e11147.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Kurtz S, Phillippy A, Delcher AL, *et al.*: **Versatile and open software for comparing large genomes.** *Genome Biol.* 2004; **5**(2): R12.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Nair D, Memmi G, Hernandez D, *et al.*: **Wholegenome sequencing of *Staphylococcus aureus* strain rn4220, a key laboratory strain used in virulence research, identifies mutations that affect not only virulence factors but also the fitness of the strain.** *J Bacteriol.* 2011; **193**(9): 2332–2335.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Tatusova TA, Madden TL: **BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences.** *FEMS Microbiol Lett.* 1999; **174**(2): 247–250.
PubMed Abstract | Publisher Full Text
10. Brandt S, Kristensen A: **Web push as an internet notification service.** In *W3C Workshop on Push Technology, Boston, Massachusetts*, 1997.
Reference Source
11. Fette I, Melnikov A: **The websocket protocol.** *IETF Internet draft.* 2011.
Reference Source

Open Peer Review

Current Referee Status: ☑ ☑

Version 1

Referee Report 17 July 2014

doi:10.5256/f1000research.2867.r4953

☑

Loren J Hauser
Computational Biology and Bioinformatics Group, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

C-Sibelia was created to be a user friendly tool for pairwise genome comparison. The web tool is relatively easy to use for the non-bioinformatics trained and will therefore make these kinds of analysis easier for small groups to perform. This is a big need. I can see the possibility of additional analyses to be added which will enhance the site. However, the output page has some real deficiencies that need to be fixed prior to public use.

- The web site can’t recognize a fasta file unless named xxx.fasta. This is annoying since all genbank fasta files end in .fna and therefore will have to be renamed before use. This is an easy fix.
- I tested 2 bacterial genomes over 6 Mb and it took about 5 minutes for the results to come back. This is OK but could be a problem if the site eventually gets a lot of use.
- The output needs a lot more definitions or explanations: for example 1) the section titled “*number of effects by impact*”, I have no idea what high, low, moderate or modifier means; 2) there are 10,853 modifiers but there are only 1037 listed in the section “*changes by type*”; 3) in the section “ *number of effects by functional class”* there are 36 missense and 75 silent mutation listed, but there are 160 non-synonymous_coding and 136 synonymous_coding in the section “*number of effect by type and region*”; 4) since these were bacterial genomes how are exons defined?
- There are additional inconsistencies and lack of definitions that I have not pointed out.

The best way to fix this is to get somebody who has never used the system and has not been trained in any way to test it in order to point out all of these inconsistencies and lack of definitions.

Some of the output was unavailable since in ran off the bottom of the allotted space. This will be a real problem when comparing draft genomes with more than 10 contigs. I did not get to see the circular map or the distribution of changes on the main chromosome because of this.

I did not try to download and use the executable, but if it as poorly described as the output it will be hard for novices to implement. Again, get a complete novice to test downloading and installing it to make sure it is easy to use.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 27 November 2013

doi:10.5256/f1000research.2867.r2566

☑ **Jeffrey McLean**
Microbial and Environmental Genomics, J. Craig Venter Institute, San Diego, CA, USA

In general, the rationale behind the decision to develop a user friendly tool to compare finished and draft assemblies, to reference genomes, is highly justified. The accurate calling of single nucleotide variations and indels using the comprehensive SnpEff tool will allow a wide variety of users to make use of C-Sibelia. For non-informatic inclined users, the output formats and the optional Circos diagrams visualizing high level genomic architecture for rearrangement analyses, is an excellent addition. The framework for the software is also an interesting choice, making it accessible to social networking style discussions with other users.

On the technical side, the authors do demonstrate that the performance of C-Sibelia in detecting variants is comparable to MUMmer and improves upon Mauve in terms of the false-positive rate. These tools are currently highly used and the research community would benefit from a user friendly tool such as C-Sibelia. The approach to decompose the genomes into synteny blocks, using the iterative de Bruijn graph algorithm is novel. In order to validate this tool I have signed in to the web-based version (signing in increases the size of the files you can upload) and tested an assembly against a reference genome and obtained results comparable to MAUVE. An explanation of the common use of the VCF format would help readers further. Visualizations of the SNPs would also be a nice addition in the future. I look forward to further enhancements of this tool.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Discuss this Article

Version 1

Author Response 17 Sep 2014

Son Pham, UCSD, USA

Dear Dr. Hauser and Dr. McLean,
Thank you very much for your very helpful comments.
1. About the fasta and fna file, the system now can recognize both file extensions.
2. About the lack of definitions in the output (number of effects by impact, by functional classes..) we now have added a link to snpEff for a full description.
3. For the response time of the server, we are now applying the queuing approach for running tasks submitted by users. When the results are ready, it will notify users and also all runs' results are kept in users' profiles. We plan to add more computational nodes to the server as it scales up.

We greatly appreciate all reviewers comments! They are all very beneficial for C-Sibelia and etools.
Thank you,
Son Pham,

Competing Interests: No competing interests were disclosed.