BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Server

Volume 8(11)

Bacterial genome mapper: A comparative bacterial genome mapping tool

Kang Seon Lee[1,2,†], Ryong Nam Kim[1,†], Byoung Ha Yoon[1,2,†], Dae Soo Kim[1], Sang Haeng Choi[1], Dong Wook Kim[1], Seong Hyeuk Nam[1], Aeri Kim[1], Aram Kang[1,2], Kun Hyang Park[1], Jae Eun Jung[1], Sung Hwa Chae[3] & Hong Seog Park[1, 2,*]

[1]Genome Resource Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 305-806, Korea ; [2]University of Science and Technology (UST), Daejeon 305-333, Korea; [3]Research Institute of GnC Bio Co. Ltd, Daejeon 305-150, South Korea; Hong Seog Park – Email: hspark@kribb.re.kr; *Corresponding author
†-These authors contributed equally to this work.

Received May 31, 2012; Accepted June 3, 2012; Published June 16, 2012

Abstract:
Recently, next generation sequencing (NGS) technologies have led to a revolutionary increase in sequencing speed and cost-efficacy. Consequently, a vast number of contigs from many recently sequenced bacterial genomes remain to be accurately mapped and annotated, requiring the development of more convenient bioinformatics programs. In this paper, we present a newly developed web-based bioinformatics program, Bacterial Genome Mapper, which is suitable for mapping and annotating contigs that have been assembled from bacterial genome sequence raw data. By constructing a multiple alignment map between target contig sequences and two reference bacterial genome sequences, this program also provides very useful comparative genomics analysis of draft bacterial genomes.

Availability: Bacterial Genome Mapper is freely accessible at http://mbgm.kribb.re.kr

Key words: bacterial genome, contig mapping program

Background:
Recently, with the application of NGS technologies, the number of sequenced prokaryotic genomes has increased dramatically. According to published reports (GOLD database **[1]**, February 2012) the current number of completed prokaryotic genome sequences is 2,802, and the number of incomplete prokaryotic genome sequences is 7,001. In addition, a vast number of prokaryotic genomes are predicted to be sequenced in the upcoming years. At present, one of the most challenging problems for microbiologists is how to quickly analyze and annotate the huge amount of raw bacterial genome sequence data that is rapidly being deposited into public databases. One way to solve this difficulty could be to develop more user-friendly genome analysis programs that can be used by microbiologists with no professional informatics training. In particular, accurate contig mapping is an important step in constructing the final bacterial genome sequences. Although several contig-mapping programs are currently available, they are still insufficient to guarantee both the accuracy of the final bacterial genome sequence and user-friendliness. In this paper, we present a newly developed web-based program, Bacterial Genome Mapper, which is used to map bacterial genome contigs. This program maps contigs using the information that is obtained from multiple alignments between target contigs from bacterial genome sequence raw data and one or two corresponding reference genome sequences. The unique mapping process used by this program also enables the

comparative genomics analysis of a draft bacterial genome sequence.

Methodology:

The pipeline process of the Bacterial Genome Mapper is explained in (**Figure 1A**). This web-based program constructs contig maps by performing multiple alignments between target genome contigs and reference bacterial genome sequences **(Figure 1B and C)**. The Bacterial Genome Mapper is composed of two parts, Mapper 1 and Mapper 2 **(Figures 1A, B & C).** Mapper 1 aligns target contigs with a reference bacterial genome sequence, thereby mapping the target contigs and constructing a comparative genome map simultaneously. The user must provide target contig sequences and one reference genome sequence. Target contigs are aligned with the reference genome sequence by BLAST **[2]** and then ordered and merged based on BLAST outputs. Unmerged separate contigs are merged by adding 100bp of N sequence between them. The merged contig is aligned to the reference genome sequence by BLAST, generating a comparative genomic map. Optionally, the Genbank file of the reference genome sequence and the paired end sequence FASTA files for the target contigs can be uploaded to provide detailed annotation of the gene regions (arrowheads indicate the transcriptional direction in (**Figures 1B, C, D & E)**. Users should also provide a list file of the paired end sequences, based on which Mapper1 can retrieve correctly each pair of paired end sequences from an input paired end sequence FASTA file. The positions of the paired end sequences are indicated in the merged contig map.

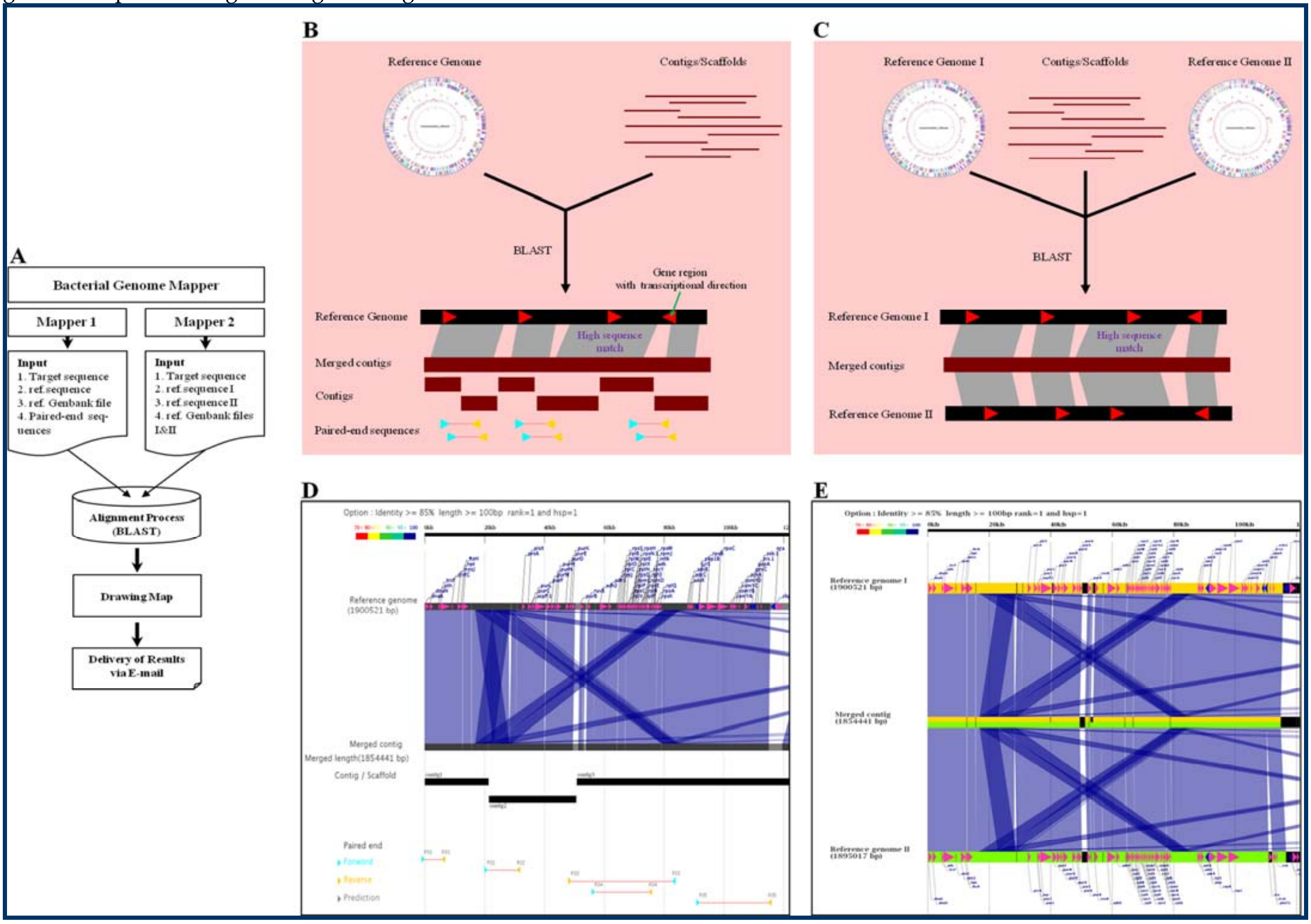

Figure 1: Bacterial Genome Mapper: **(A)** Pipeline process of Bacterial Genome Mapper; **(B)** Schematic outline of the Mapper 1 workflow; **(C)** Schematic outline of the Mapper 2 workflow; **(D)** A local region in a result image file generated by Mapper 1. The color of each beam connecting a region in the merged contig and a region in the reference genome indicates sequence identity, as shown by the multicolored bar symbol with numbered sequence identity intervals. The red and blue arrowheads in the thick black line (reference genome) represent transcriptional direction within gene regions. The gray regions in the reference genome and the merged contig represent unmatched regions; **(E)** a local region in a result image file generated by Mapper 2. The black regions in reference genome I (thick orange line), reference genome II (thick green line) and the merged contig represent the unmatched sequence regions between reference genome I and the merged contig and between reference genome II and the merged contig. The orange and green portions in the thick line showing the merged contig represent the comparisons between reference genome I and the merged contig and between reference genome II and the merged contig, respectively.

Mapper 2 aligns target contigs with two reference bacterial genome sequences, I and II, and can thereby draw a comparative genome map. The user must provide target contig sequences and two reference bacterial genome sequences. Target contigs are aligned with reference bacterial genome sequence I by BLAST and then merged based on the BLAST output. The merged contig is aligned with reference bacterial genome sequence II by BLAST, generating a detailed comparative genome map. Optionally, the Genbank files of the two reference bacterial genome sequences can be uploaded to indicate each reference gene region on the map. The servlets associated with the Bacterial Genome Mapper platform are hosted on an Apache-Tomcat 5.5 server running the Linux operating system. JSP is employed as the web engine of the web server. The web interface of Bacterial Genome Mapper is implemented in an operating system-independent manner and has been tested in Internet Explorer 8.0 and Chrome.

Discussion:

Bacterial Genome Mapper has been tested using three bacterial genome sequences from *Streptococcus pyogenes MGAS315* **[3]**, *S. pyogenes M1 GAS* **[4]** and *S. pyogenes MGAS8232* **[5],** with genome sizes of ~1.9 Mb, ~1.85 Mb and ~1.89 Mb, respectively. First, Mapper 1 was tested using the genome sequences from *S. pyogenes* MGAS315 and *S. pyogenes* M1 GAS as the reference genome and target genome, respectively. The target genome was randomly divided into 20 contigs, and 20 pairs of paired end sequences were created. The results are shown in **(Figure 1D)**. The detailed locations of individual contigs in the merged contig relative to the reference genome sequence are shown in the result analysis file (blastout file). Second, Mapper 2 was tested using the genome sequences of *S. pyogenes* MGAS315 and *S. pyogenes* MGAS8232 as reference genomes I and II, respectively. The target contigs were the same as those used in the test for the Mapper 1. The results are shown in (**Figure 1E**). Also, the detailed locations of individual contigs in the merged contig relative to the two reference genome sequences are shown in the result analysis file (blastout file).

Conclusion:

Bacterial Genome Mapper is used to map contigs and construct comparative genome maps by wide alignments between target bacterial genome contigs and one or two reference bacterial genome sequences. In addition, using the paired end sequences and the Genbank files of the reference genomes, the Bacterial Genome Mapper helps in finishing and annotating a draft bacterial genome sequence. It is a user-friendly web-based program that can be used even by microbiologists with little or no professional informatics training. Due to its user-friendliness, free accessibility and effective applications toward finishing a draft genome, Bacterial Genome Mapper is likely to become one of the most popular bioinformatics programs for bacterial genome analysis in the future.

Reference:

[1] Liolios K *et al. Nucleic Acids Res.* 2006 34: D332 [PMID: 16381880]

[2] Altschul SF *et al. J Mol Biol.*1990 **215**: 403 [PMID: 2231712]

[3] Beres SB *et al. Proc Natl Acad Sci U S A.* 2002 **99**: 10078 [PMID: 12122206]

[4] Ferretti JJ *et al. Proc Natl Acad Sci U S A.* 2001 **98**: 4658 [PMID: 11296296]

[5] Smoot JC *et al. Proc Natl Acad Sci U S A.* 2002 **99**: 4668 [PMID: 11917108]

Edited by P Kangueane
Citation: **Lee *et al.*** Bioinformation 8(11): 532-534 (2012)