GcoGSA-BA: A GLOBAL CORE GENOME SNP ANALYSIS FOR *BACILLUS ANTHRACIS*

Akifumi Yamashita, Tsuyoshi Sekizuka, and Makoto Kuroda

As an issue of biosecurity, it is important to identify the origin of a suspected sample to distinguish whether it originated from the release of a bioterrorism agent or from environmental contamination with a virulent agent. Here we have developed an analytical pipeline that can infer the phylogenetic position of *Bacillus cereus* group species, including *B. anthracis*, from next-generation sequencing reads without extensive genomics skills. GcoGSA-BA can also detect the existence of anthrax plasmid pXO1 carrying 3 anthrax toxins (lethal factor, edema factor, and protective antigen). This pipeline can also be used to correctly infer the phylogenetic position and to detect the suspected isolate carrying anthrax toxins among *B. cereus* group.

AS AN ISSUE OF BIOSECURITY, it is important to identify whether a suspected sample originates from the intentional release of a bioterrorism agent or from environmental contamination with a virulent agent. Recently developed next-generation sequencing (NGS) techniques have allowed the acquisition of genome-wide sequence reads in a short time, and at a much lower cost than conventional methods, as well as the inference of phylogenetic positions for *Bacillus cereus* group species (including *B. anthracis* species) from NGS reads.¹ Currently, similar approaches to infer single-nucleotide polymorphisms (SNPs) and/or phylogenetic positions from sample sequences are available through websites; examples of this approach are “SNPs finder,” hosted by the Los Alamos National Laboratory,² and “snpTree 1.1”³ and “CSI Phylogeny 1.0a,”⁴ hosted by the Center for Genomic Epidemiology. SNPs finder can accept only complete sequences or contigs for input data and does not infer phylogenetic positions, as it was designed only for detecting SNPs. By contrast, the other 2 services can accept various NGS reads as well as contigs and can infer phylogenetic positions for various bacterial species.

To conduct phylogenetic analysis of *B. cereus* group species (including *B. anthracis*) more conveniently, we implemented a web service designated as global core genome SNP analysis for *B. anthracis* (GcoGSA-BA), which shows the phylogenetic position of the sample among 29 *B. anthracis* and 104 *B. cereus* strains. This web service accepts gzip-compressed, FASTQ-formatted sequences, which is the most popular format for the output of NGS reads. GcoGSA-BA also detects the existence of anthrax plasmids pXO1 and pXO2^{5,6} and 3 plasmid-coded toxin proteins, including lethal factor (LF), edema factor (EF), and protective antigen (PA) in the pXO1.

MATERIALS AND METHODS

The GcoGSA-BA web tool consists of simple pipelines that include read quality trimming, reference genome and toxin gene mapping, and phylogeny extraction based on the core genome sequence. Read trimming is performed using the skewer program⁷ with a quality threshold of 15 and a minimum remaining sequence length of at least 40.

Akifumi Yamashita, PhD, is a Senior Researcher; Tsuyoshi Sekizuka, PhD, is a Senior Researcher; and Makoto Kuroda, PhD, is the Director; all at the Pathogen Genomics Center, National Institute of Infectious Diseases, Tokyo, Japan.

Figure 1. Screenshot of the results page on the GcoGSA-BA web tool. Less than 80% region coverage for the respective target sequences is shown in gray, and less than 5 × read depth is shown in dark blue. All results can be retrieved from the “Download All” icon. (Color graphics available online at www.liebertonline.com/hs)

(<https://genome.niid.go.jp/gcogsa/images/bacillus/SNP-allele.table.gz>). The SimSeq program was employed to generate the simulated reads from the complete or draft genome sequence of the above *Bacillus* strains (SimSeq available at <https://github.com/jstjohn/SimSeq>). A neighbor-joining (NJ) phylogenetic tree⁹ could then be constructed from the concatenated 657,183 SNP alleles using the APE package of R.¹⁰ A zoomed-in version of the phylogenetic tree surrounding the query strain(s) is also created for user convenience. The results can then be downloaded from the website, including the SNP table in tabular format and multifasta format, NJ phylogenetic tree in Newick format and PDF format, and mapped ratios for further analysis.

format and PDF format, and mapped ratios for further analysis.

challenged data sets are shown in red. Branches corresponding to
line at www.liebertonline.com/hs)

The source code for installing on a local server is available on request. The GcoGSA-BA website is <https://genome.niid.go.jp/gcogsa/BA/index.html>.

RESULTS AND DISCUSSION

In total, 657,183 SNP sites on *B. anthracis* Ames Ancestor chromosome DNA (5,227,419 bp) were extracted using complete or draft genome sequence from 29 *B. anthracis* and 104 *B. cereus* strains, indicating that 12.57% of chromosome regions were comprehensively selected for phylocore-genome analysis. To evaluate the performance of GcoGSA-BA, we challenged raw or simulated data obtained from 4 types of *Bacilli*: 2 strains of *B. anthracis* (Ames Ancestor and A0149) and 2 strains of *B. cereus* (G9241 and VD166). Sequence read archive data for *B. anthracis* A0149 and *B. cereus* VD166 were retrieved from SRR1318540 and SRR398308 in NCBI (<http://www.ncbi.nlm.nih.gov/Traces/sra>), respectively, and simulated short reads were prepared from *B. anthracis* Ames Ancestor (chromosome DNA, NC_007530.2; pXO1, NC_007322.2; pXO2, NC_007323.3) and the draft genome assembly of *B. cereus* G9241 (AAEK01000001-AAEK01000207). All of the tested samples were deeply mapped to the reference Ames Ancestor chromosome with sufficient coverage above 20× to assign the SNP sites (646,749/657183 sites; 98.4% coverage).

Regarding coverage of the chromosome, *B. anthracis* Ames Ancestor and A0149 showed greater than 99% coverage, whereas *B. cereus* G9241 and VD166 showed less than 85% coverage, which is consistent with the fact that *B. anthracis* Ames Ancestor is used as a reference genome sequence. This phylogenetic distance between *B. anthracis* and *B. cereus* can be observed in Figure 1. Despite the lower coverage of *B. cereus* strains, 646,749 out of the 657,183 (98.4%) SNP sites were able to be assigned for construction of the phylogenetic tree (Figure 2), suggesting that the selected SNP sites were sufficient and effective for classification of *B. anthracis* and *B. cereus* species. A0149 is one of the Trans-Eurasian (TEA) group strains of *B. anthracis*, and the obtained phylogenetic tree (Figure 3) suggests that MiSeq raw short reads of A0149 could be correctly assigned as closely related to the Ba4599 strain from the outbreak in 2009 originating from a Scottish heroin user, as well as the report by Price et al.¹¹

GcoGSA-BA could also detect both pXO1 and pXO2 sequences and 3 major anthrax toxins (LF, EF, and PA) with bwa-sw mapping. In addition to the anthrax toxin-positive Ames Ancestor and A0149 strains, the *B. cereus* G9241 strain showed positivity for the 3 toxins evaluated. *B. cereus* G9241 was isolated from the sputum and blood of a patient with life-threatening pneumonia, carrying a plasmid that is fully homologous to the pXO1 plasmid.¹² Thus, toxin-positive *Bacillus* strains can be identified simply using NGS and following the GcoGSA-BA web tool.

Although the NJ phylogenetic method is applied with GcoGSA-BA to facilitate rapid construction of phylocore-

Figure 3. The phylocore-genome analysis for *B. anthracis* strains only. (Color graphics available online at www.liebertonline.com/hs)

genome trees, the maximum likelihood method may better estimate the correct phylogenetic lineage in regard to the molecular clock. The user can further analyze the results with sufficient bootstrapping analysis on another website, called RAXML BlackBox (<http://embnet.vital-it.ch/raxml-bb/>),¹³ using concatenated SNP alleles (“SNPs data in fasta format”) from the download site.

ACKNOWLEDGMENTS

This work was supported by a grant-in-aid from the Ministry of Health, Labour, and Welfare, Japan (H23 Shinko-Ippan-007 and H26-Shinkogyosei-Shitei-002). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. An icon in the web content was provided free of charge from the homepage

of the materials' provider "Airuumu" (this web page is now closed). The authors declare no conflict of interest associated with this manuscript.

REFERENCES

1. Kuroda M, Serizawa M, Okutani A, Sekizuka T, Banno S, Inoue S. Genome-wide single nucleotide polymorphism typing method for identification of *Bacillus anthracis* species and strains among *B. cereus* group species. *J Clin Microbiol* 2010;48(8):2821-2829.
2. Song J, Xu Y, White S, Miller KW, Wolinsky M. SNPsFinder—a web-based application for genome-wide discovery of single nucleotide polymorphisms in microbial genomes. *Bioinformatics* 2005;21(9):2083-2084.
3. Leekitcharoenphon P, Kaas RS, Thomsen MC, Friis C, Rasmussen S, Aarestrup FM. SnpTree—a web-server to identify and construct SNP trees from whole genome sequence data. *BMC Genomics* 2012;13(Suppl 7):S6.
4. Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS One* 2014;9(8):e104984.
5. Okinaka RT, Cloud K, Hampton O, et al. Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. *J Bacteriol* 1999;181(20):6509-6515.
6. Ravel J, Jiang L, Stanley ST, et al. The complete genome sequence of *Bacillus anthracis* Ames "Ancestor." *J Bacteriol* 2009;191(1):445-446.
7. Jiang H, Lei R, Ding SW, Zhu S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* 2014;15:182.
8. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;26(5):589-595.
9. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987;4(4):406-425.
10. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 2004;20(2):289-290.
11. Price EP, Seymour ML, Sarovich DS, et al. Molecular epidemiologic investigation of an anthrax outbreak among heroin users, Europe. *Emerg Infect Dis* 2012;18(8):1307-1313.
12. Hoffmaster AR, Ravel J, Rasko DA, et al. Identification of anthrax toxin genes in a *Bacillus cereus* associated with an illness resembling inhalation anthrax. *Proc Nat Acad Sci U S A* 2004;101(22):8449-8454.
13. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML Web servers. *Syst Biol* 2008;57(5):758-771.

Manuscript received October 22, 2014;
accepted for publication November 12, 2014.

Address correspondence to:
Makoto Kuroda, PhD
Director
Laboratory of Bacterial Genomics, Pathogen Genomics Center
National Institute of Infectious Diseases
1-23-1 Toyama, Shinjuku-ku
Tokyo 162-8640, Japan

E-mail: makokuro@niid.go.jp