SAT - 559

Pan-genome and phylogeny of *Bacillus cereus sensu lato*

Adam L. Bazinet National Biodefense Analysis and Countermeasures Center, Fort Detrick, MD, 21702, USA adam.bazinet@nbacc.dhs.gov

ABSTRACT

Background: *Bacillus cereus sensu lato* (*s. l.*) is an ecologically diverse bacterial group of medical and agricultural significance. In this study, I used publicly available genomes to characterize the *B. cereus s. l.* pan-genome and performed the largest phylogenetic analyses of this group to date in terms of the number of genes and taxa included. With these fundamental data in hand, it became possible to identify genes associated with particular phenotypic traits (i.e., "pan-GWAS" analysis), and to quantify the degree to which taxa sharing common attributes were phylogenetically clustered.

Methods: A rapid *k*-mer based approach (Mash) was used to create reduced representations of selected *Bacillus* genomes, and a fast distance-based phylogenetic analysis of this data (FastME) was performed to decide which species should be included in *B. cereus s. l.* The complete genomes of eight *B. cereus s. l.* species were annotated de novo with Prokka, and these annotations were used by Roary to produce the *B. cereus s. l.* pan-genome. Scoary was used to associate gene presence and absence patterns with various phenotypes. The orthologous protein sequence clusters produced by Roary were filtered and used to build HaMStR databases of gene models that were used in turn to construct phylogenetic data matrices. Phylogenetic analyses used RAxML, DendroPy, ClonalFrameML, Gubbins, PAUP*, and SplitsTree. The genealogical sorting index was used to assess the tree-based clustering of taxa sharing common attributes.

Results: The *B. cereus s. l.* pan-genome currently consists of ≈60,000 genes, ≈600 of which are "core" (common to at least 99% of taxa sampled). Pan-GWAS analysis revealed genes that were associated with phenotypes such as isolation source, oxygen requirement, and ability to cause diseases such as anthrax or food poisoning. Extensive phylogenetic analyses using an unprecedented amount of data produced phylogenies that were largely concordant with each other and with previous studies. Phylogenetic support as measured by bootstrap probabilities increased markedly when all suitable pan-genome data was included in phylogenetic analyses, as opposed to when only core genes were used. *B. cereus s. l.* taxa sharing common traits and species designations exhibited varying degrees of phylogenetic clustering.

INTRODUCTION

- *Bacillus cereus sensu lato* (*s. l.*) is an ecologically diverse bacterial group that comprises a growing number of species, many of which are medically or agriculturally important.
- Well-known species include *B. anthracis* (the causative agent of anthrax), *B. cereus sensu stricto* (causes food poisoning and other ailments), and *B. thuringiensis* (used to control insect pests); several other species have also been described.
- A typical *B. cereus s. l.* genome contains ≈5,500 protein-coding genes[5,6]. Due to rampant horizontal gene transfer in bacterial ecosystems, however, the genomes of closely related taxa vary in their gene content. Thus, it is now common practice to seek to characterize the full gene complement of a closely related group of bacterial strains or species, otherwise known as a "pan-genome"[7].
- Published phylogenies of *B. cereus s. l.* tend to agree with and reinforce one another, although different classification systems have been developed with attendant implications for species designations. One popular classification system divides the *B. cereus s. l.* phylogeny into three broad clades[9,12,16]: (Clade 1 = *B. anthracis*, *B. cereus*, and *B. thuringiensis*; Clade 2 = *B. cereus* and *B. thuringiensis*; and Clade 3 = *B. cereus*, *B. cytotoxicus*, *B. mycoides*, *B. thuringiensis*, *B. toyonensis*, and *B. weihenstephanensis*). A somewhat more fine-grained classification system divides the phylogeny into seven major groups[10,11,14], each with its own thermotolerance profile[10] and propensity to cause food poisoning[17].

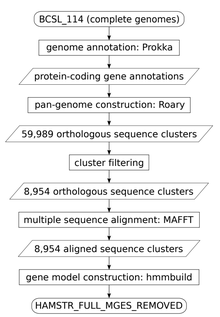
GOAL

To produce the most accurate and comprehensive estimate of the *B. cereus s. l.* pan-genome and phylogeny to date by analyzing all publicly available *B. cereus s. l.* genome data with a novel bioinformatic workflow for pan-genome characterization and pan-genome-based phylogenetic analysis.

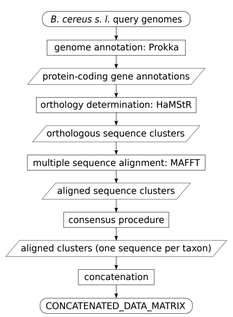
METHODS

- A distance-based phylogeny of 146 genomes of the genus *Bacillus* was computed with Mash[18] and FastME[19].
- 114 complete genomes of eight *B. cereus s. l.* species ("BCSL_114") were used for the majority of the analyses.
- 498 genomes of 13 *B. cereus s. l.* species ("BCSL_498") were used for the final and largest analyses.
- All *B. cereus s. l.* genomes were annotated de novo with Prokka[21].
- The pan-genome of *B. cereus s. l.* was inferred with Roary[22] using the BCSL_114 taxon set.
- Orthologous protein sequence clusters output by Roary were filtered to produce a set of 9,070 gene models for use with HaMStR[20].
- A reduced HaMStR database was created that consisted of only the 594 gene models corresponding to core genes.
- Variants of the HaMStR databases were created with gene models corresponding to putative mobile genetic elements removed (Additional file 5).
- HaMStR was used to search the protein-coding gene annotations of BCSL_498 taxa not included in BCSL_114 for sequences matching HaMStR database gene models.
- Amino acid sequences assigned to orthologous sequence clusters were aligned using MAFFT[24].
- Amino acid alignments were converted to nucleotide alignments by substituting for each aa the proper codon from the original coding sequence.
- A "consensus"[29] procedure was used to collapse all sequence variants into a single sequence by replacing multi-state positions with nucleotide ambiguity codes.
- Individual sequence cluster alignments were concatenated, adding gaps for missing data as necessary (Additional file 6).
- Concatenated nucleotide data matrices were analyzed under the maximum likelihood criterion using RAxML[26].
- Data were analyzed either with all nucleotides included in a single data subset, or with sites partitioned by codon position.
- BCSL_114 taxon set analyses included an adaptive best tree search[27] and a procedure that used the `autoMRE` RAxML bootstopping criterion[28].
- PhiPack[29], Gubbins[30], and ClonalFrameML[31] were evaluated for their ability to improve phylogenetic reconstruction by accounting for recombination.
- Concatenated nucleotide data matrices were analyzed under the maximum parsimony criterion using PAUP*[32].
- Standard and normalized Robinson-Foulds distances were calculated using RAxML to quantify the difference between pairs of tree topologies[33].
- Visualizations of phylogenetic trees were produced with FigTree[34].

WORKFLOWS



Additional file 5. Construction of a HaMStR database.



Additional file 6. Construction of a data matrix.

RESULTS

- The Mash-distance-based phylogeny of the genus *Bacillus* indicated a *B. cereus s. l.* clade comprised of 13 distinct species.
- The first taxon to split from the remainder of *B. cereus s. l.* was *B. manliponensis*[4], followed by *B. cytotoxicus*[1] (previously recognized as an outlier[8,15]).
- Roary produced a total of 59,989 protein-coding gene sequence clusters. The *B. cereus s. l.* "core genome", consisting of genes present in at least 99% of taxa sampled, was represented by 598 genes (≈1% of all genes).
- The 59,391 non-core genes were divided into 32,324 "accessory genes" (i.e., non-core genes present in at least two taxa; ≈54% of all genes), and 27,067 "unique genes" (i.e., genes present in only one taxon; ≈45% of all genes).
- Seven different concatenated nucleotide data matrices were constructed and analyzed. The majority of the data matrices used the BCSL_114 taxon set; only one matrix used the BCSL_498 taxon set. Various gene sets were used, including 1) all core genes identified by Roary; 2) core genes used to build the HaMStR database; 3) HaMStR core genes with mobile genetic elements (MGEs) removed, and variants of this gene set with either PhiPack sites removed or Gubbins sites removed; and 4) all HaMStR genes with MGEs removed. Aligned data matrices ranged from 96,802 nt to 8,207,628 nt in length. Matrix completeness, defined as the percentage of non-missing data, ranged from 47.4% to 99.5%. The percentage of ambiguous characters present in data matrices ranged from 0.0% to 17.0%.
- Ten different phylogenetic analyses of the seven concatenated data matrices were performed; 9/10 analyses used maximum likelihood, and one analysis used maximum parsimony. All exploratory analyses used the BCSL_114 taxon set; only when the best-performing methods were established was analysis of the BCSL_498 taxon set pursued. During the exploratory phase, several variables were tested for their effect on phylogenetic outcome: 1) use of MAFFT instead of PRANK[23] to align protein sequence clusters; 2) removal of MGEs; 3) use of maximum parsimony in addition to maximum likelihood; 4) partitioning of sites by codon position; 5) masking or removal of sites implicated in recombination; and 6) use of all eligible genes from the pan-genome versus only core genes. Variables 1-5 did not significantly change the phylogenetic results. Using all eligible genes for phylogenetic analysis instead of only core genes caused bootstrap probabilities to increase substantially.
- All phylogenetic analysis results recapitulated the three-clade and seven-group classification systems of previous studies. Taxa were consistently assigned to the same clade and group, independent of the particular phylogenetic analysis performed. Topological differences between analysis results, as measured by the Robinson-Foulds distance, were confined to intra-group relationships. Bootstrap support was fairly consistent for all analyses that used core genes, and increased dramatically when all eligible genes from the pan-genome were used. The best estimate of the phylogenetic relationships among the BCSL_114 taxa is shown in Figure 2.
- Once the exploratory analyses were completed, an analysis of BCSL_498 was executed (Fig. 3). In contrast to analyses of BCSL_114, Group II is now represented on the tree, and is located where expected[10,11,14].

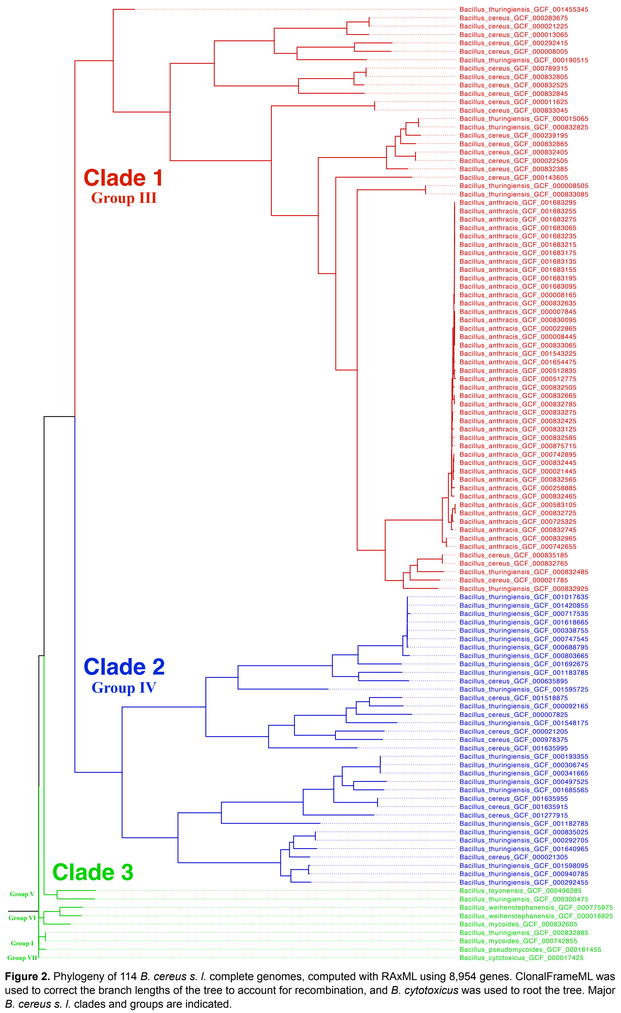
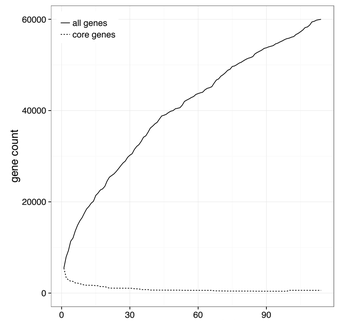
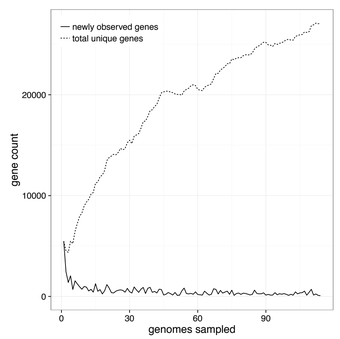


Figure 2. Phylogeny of 114 *B. cereus s. l.* complete genomes, computed with RAxML using 8,954 genes. ClonalFrameML was used to correct the branch lengths of the tree to account for recombination, and *B. cytotoxicus* was used to root the tree. Major *B. cereus s. l.* clades and groups are indicated.



Additional file 8. The rarefaction curve shows that after ≈35 genomes have been sampled (≈31% of all genomes), the number of core genes remains fairly constant at ≈600 genes, while the total number of genes in the pan-genome continues to increase almost linearly.



Additional file 9. The rarefaction curve shows that as genomes are sampled, genes never before observed continue to be found at a fairly steady rate, and the total number of unique genes discovered continues to increase, with no indication of soon approaching an asymptote.

Figure 3. Phylogeny of 498 *B. cereus s. l.* taxa, computed with RAxML using 8,954 genes. *B. manliponensis* was used to root the tree. Major *B. cereus s. l.* clades are colored, and major groups are indicated.

CONCLUSIONS

- The *B. cereus s. l.* pan-genome is "open" (Additional files 8 and 9), implying that continued sampling of the group — especially of underrepresented taxa such as environmental strains[13] — will continue to reveal novel gene content.
- My estimate of the number of protein-coding genes in the *B. cereus s. l.* core and pan-genome (≈600 and ≈60,000, respectively), based on 114 complete genomes, is consistent with previous estimates[8,13], as more extensive sampling of an open pan-genome will necessarily reduce the core genome size while simultaneously increasing the pan-genome size. The diversity and adaptability of *B. cereus s. l.* may be in part attributable to the significant proportion of unique genes in its pan-genome (≈27,000, almost 50% of all genes; Additional file 9).
- All phylogenetic analyses recapitulated the three-clade and seven-group classification systems, and taxa were consistently assigned to the same clade and group (Figs. 2 and 3), irrespective of the data source or analysis methodology used. This strongly suggests that the broad phylogenetic structure of *B. cereus s. l.* has been inferred correctly. I demonstrate that the three-clade and seven-group systems are compatible with each other, as no group has its member taxa assigned to multiple clades. Clades 1 and 2 are much more extensively sampled than Clade 3 due to historical interest in *B. anthracis* and *B. thuringiensis*; a recent study has shown that there is likely to be a tremendous amount of as-yet incompletely characterized diversity in Clade 3 that can be assayed by sampling various natural environments[13]. Indeed, Clade 3 exhibited the greatest degree of species diversity; in particular, Group I contained representatives of seven different species, including two newly characterized species (*B. bingmayongensis*[2] and *B. gaemokensis*[3]). Five of the 498 taxa did not place into one of the seven previously circumscribed groups, which suggests that classification systems will need to be updated and refined as additional isolates are sequenced. Perhaps most interesting among the unplaced taxa is *B. manliponensis*[4], which appears to be even more divergent from other *B. cereus s. l.* taxa than *B. cytotoxicus*[1] (Fig. 3).
- In a bioforensic setting, phylogenies that include well-supported strain-level relationships aid greatly in the identification of unknown isolates. However, the extremely high level of genomic conservation among closely related bacterial strains, especially in the core genome or in commonly typed conserved regions such as housekeeping genes, has limited the ability of previous analyses to make robust strain-level phylogenetic inferences. An important contribution of the current study is to show that bootstrap probabilities increase substantially when accessory genes are included in phylogenetic analyses along with core genes. Thus, I have been able to resolve many strain-level, intra-group relationships of *B. cereus s. l.* with 100% bootstrap support for the first time.

TAKE-HOME MESSAGES

- *B. cereus s. l.* currently has an estimated 600 protein coding genes in its core genome and 60,000 genes in its pan-genome.
- Continued sampling of the pan-genome will continue to reveal novel gene content.
- The phylogenetic structure of *B. cereus s. l.* was consistently recapitulated in each analysis, suggesting it is accurate.
- Bootstrap values increase substantially when accessory genes are included in phylogenetic analyses.
- Many strain-level relationships of *B. cereus s. l.* were resolved with 100% bootstrap support for the first time.

REFERENCES

1. Guinebretière, M.-H., Auger, S., Galleron, N., Contzen, M., De Sarrau, B., De Buyser, M.-L., Lamberet, G., Fagerlund, A., Granum, P.E., Lereclus, D., De Vos, P., Nguyen-The, C., Sorokin, A.: *Bacillus cytotoxicus* sp. nov. is a novel thermotolerant species of the *Bacillus cereus* Group occasionally associated with food poisoning. International Journal of Systematic and Evolutionary Microbiology 63(1), 31–40 (2013).
2. Liu, B., Liu, G.-H., Hu, G.-P., Cetin, S., Lin, N.-Q., Tang, J.-Y., Tang, W.-Q., Lin, Y.-Z.: *Bacillus bingmayongensis* sp. nov., isolated from the pit soil of Emperor Qin's Terra-cotta warriors in China. Antonie van Leeuwenhoek 105(3), 501–510 (2014).
3. Jung, M.-Y., Paek, W.K., Park, I.-S., Han, J.-R., Sin, Y., Paek, J., Rhee, M.-S., Kim, H., Song, H.S., Chang, Y.-H.: *Bacillus gaemokensis* sp. nov., isolated from foreshore tidal flat sediment from the Yellow Sea. The Journal of Microbiology 49(5), 867–871 (2010).
4. Jung, M.-Y., Kim, J.-S., Paek, W.K., Lim, J., Lee, H., Kim, P.I., Ma, J.Y., Kim, W., Chang, Y.-H.: *Bacillus manliponensis* sp. nov., a new member of the *Bacillus cereus* group isolated from foreshore tidal flat sediment. The Journal of Microbiology 49(6), 1027–1032 (2011).
5. Papazisi, L., Rasko, D.A., Ratnayake, S., Bock, G.R., Remortel, B.G., Appalla, L., Liu, J., Dracheva, T., Braisted, J.C., Shallom, S., Jarrahi, B., Snesrud, E., Ahn, S., Sun, Q., Ristone, J., Ekstad, O.A., Kolsto, A.-B., Fleischmann, R.D., Peterson, S.N.: Investigating the genome diversity of *B. cereus* and evolutionary aspects of *B. anthracis* emergence. Genomics 98(1), 26–39 (2011).
6. Toby, I.T., Widmer, J., Dyer, D.W.: Divergence of protein-coding capacity and regulation in the *Bacillus cereus sensu lato* group. BMC Bioinformatics 15(11), 8 (2014).
7. Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., DeBoy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M.: Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". Proceedings of the National Academy of Sciences of the United States of America 102(39), 13950–13955 (2005).
8. Lapidus, A., Goltsman, E., Auger, S., Galleron, N., Ségurens, B., Dossat, C., Land, M.L., Broussolle, V., Brillard, J., Guinebretière, M.-H., Sanchis, V., Nguen-the, C., Lereclus, D., Richardson, P., Wincker, P., Weissenbach, J., Ehrlich, S.D., Sorokin, A.: Extending the *Bacillus cereus* group genomics to putative food-borne pathogens of different toxicity. Chemico-Biological Interactions 171(2), 236–249 (2008).
9. Zwick, M.E., Joseph, S.J., Didelot, X., Chen, P.E., Bishop-Lilly, K.A., Stewart, A.C., Willner, K., Nolan, N., Lentz, S., Thomason, M.K., Sozhamannan, S., Mateczun, A.J., Du, L., Read, T.D.: Genomic characterization of the *Bacillus cereus sensu lato* species: Backdrop to the evolution of *Bacillus anthracis*. Genome Research 22(8), 1512–1524 (2012).
10. Guinebretière, M.-H., Thompson, F.L., Sorokin, A., Normand, P., Dawyndt, P., Ehling-Schulz, M., Svensson, B., Sanchis, V., Nguyen-The, C., Heyndrickx, M., De Vos, P.: Ecological diversification in the *Bacillus cereus* group. Environmental Microbiology 10(4), 851–865 (2008).
11. Tourasse, N.J., Ekstad, O.A., Kolsto, A.-B.: HyperCAT: an extension of the SuperCAT database for global multi-scheme and multi-datatype phylogenetic analysis of the *Bacillus cereus* group population. Database 2010, 017 (2010).
12. Didelot, X., Barker, M., Falush, D., Priest, F.G.: Evolution of pathogenicity in the *Bacillus cereus* group. Systematic and Applied Microbiology 32(2), 81–90 (2009).
13. Drewniowska, J.M., Swiecicka, I.: Eco-genetic structure of *Bacillus cereus sensu lato* populations from different environments in northeastern Poland. PLOS ONE 8(12), 1–11 (2013).
14. Bohm, M.-E., Huptas, C., Krey, V.M., Scherer, S.: Massive horizontal gene transfer, strictly vertical inheritance and ancient duplications differentially shape the evolution of *Bacillus cereus* enterotoxin operons hbl, cytk and nhe. BMC Evolutionary Biology 15(1), 246 (2015).
15. Schmidt, T.R., Scott, E.J., Dyer, D.W.: Whole-genome phylogenies of the family Bacillaceae and expansion of the sigma factor gene family in the *Bacillus cereus* species-group. BMC Genomics 12(1), 430 (2011).
16. Okinaka, R.T., Keim, P.: The phylogeny of *Bacillus cereus sensu lato*. Microbiology Spectrum 4(1) (2016).
17. Guinebretière, M.-H., Velge, P., Couvert, O., Carlin, F., Debuyser, M.-L., Nguyen-The, C.: Ability of *Bacillus cereus* group strains to cause food poisoning varies according to phylogenetic affiliation (groups I to VII) rather than species affiliation. Journal of Clinical Microbiology 48(9), 3388–3391 (2010).
18. Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M.: Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17(1), 132 (2016).
19. Lefort, V., Desper, R., Gascuel, O.: FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. Molecular Biology and Evolution 32(10), 2798–2800 (2015).
20. Ebersberger, I., Strauss, S., von Haeseler, A.: HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evolutionary Biology 9(1), 157 (2009).
21. Seemann, T.: Prokka: rapid prokaryotic genome annotation. Bioinformatics 30(14), 2068–2069 (2014).
22. Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J.: Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31(22), 3691–3693 (2015).
23. Löytynoja, A., Goldman, N.: An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences of the United States of America 102(30), 10557–10562 (2005).
24. Katoh, K., Frith, M.C.: Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics 28(23), 3144–3146 (2012).
25. Bazinet, A.L., Cummings, M.P., Mitter, K.T., Mitter, C.W.: Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study. PLOS ONE 8(12) (2013).
26. Stamatakis, A.: RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9), 1312–1313 (2014).
27. Bazinet, A.L., Zwickl, D.J., Cummings, M.P.: A gateway for phylogenetic analysis powered by grid computing featuring GARLI 2.0. Systematic Biology 63(5), 812–818 (2014).
28. Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., Moret, B.M.E., Stamatakis, A.: In: Batzoglou, S. (ed.) How Many Bootstrap Replicates Are Necessary?, pp. 184–200. Springer, Berlin, Heidelberg (2009).
29. Bruen, T.C., Philippe, H., Bryant, D.: A simple and robust statistical test for detecting the presence of recombination. Genetics 172(4), 2665–2681 (2006).
30. Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J., Harris, S.R.: Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Research 43(3), 15 (2015).
31. Didelot, X., Wilson, D.J.: ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. PLOS Computational Biology 11(2), 1–18 (2015).
32. Swofford, D.L.: Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland, MA: Sinauer Associates (2002).
33. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical Biosciences 53(1), 131–147 (1981).
34. Rambaut, A.: http://tree.bio.ed.ac.uk/software/figtree/

DISCLAIMER

This work was funded under Agreement No. HSHQDC-15-C-00064 awarded by the Department of Homeland (DHS) Security Science and Technology Directorate (S&T) for the operation and management of the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. In no event shall the DHS, NBACC, or Battelle National Biodefense Institute (BNBI) have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. The Department of Homeland Security does not endorse any products or commercial services mentioned in this publication.