BLAST-based validation of metagenomic sequence assignments

Adam L. Bazinet[1,*], Brian D. Ondov[1,2], Daniel D. Sommer[1], and Shashikala Ratnayake[1]

[1]National Biodefense Analysis and Countermeasures Center, Fort Detrick, MD, 21702, USA
[2]Present address: National Human Genome Research Institute, Bethesda, MD, 20892, USA

[*]adam.bazinet@nbacc.dhs.gov

Abstract

When performing bioforensic casework, it is important to be able to reliably detect the presence of a particular organism in a metagenomic sample, even if the organism is only present in a trace amount. For this task, it is common to use a sequence classification program that determines the taxonomic affiliation of individual sequence reads by comparing them to reference database sequences. As metagenomic data sets often consist of millions or billions of reads that need to be compared to reference databases containing millions of sequences, such sequence classification programs typically use search heuristics and databases with reduced sequence diversity to speed up the analysis, which can lead to incorrect assignments. Thus, in a bioforensic setting where correct assignments are paramount, assignments of interest made by "first-pass" classifiers should be confirmed using the most precise methods and comprehensive databases available. In this study we present a BLAST-based method for validating the assignments made by less precise sequence classification programs, with optimal parameters for filtering of BLAST results determined via simulation of sequence reads from genomes of interest, and we apply the method to the detection of four pathogenic organisms. The software implementing the method is open source and freely available.

Keywords: BLAST; metagenomics; sequence classification; taxonomic assignment; bioforensics; validation

Introduction

In metagenomic analysis, comparing the genomic sequence content of a sample to a reference database is fundamental to understanding which organisms present in the sample were sequenced. There exist many bioinformatics software programs that perform this classification task [6,9]; some programs only estimate overall taxonomic composition and abundance in the sample [18,25], while other programs assign a taxonomic label to each metagenomic sequence [3,13,16,17,24,28]. In a bioforensic setting, one is often concerned with reliably detecting the presence of a particular organism in a metagenomic sample, which may only be present in a trace amount. For this task, one typically uses the latter class of programs just mentioned, which determine the taxonomic affiliation

of each sequence using a reference database [7, 19, 20] and a taxonomy [4]. A canonical metagenomic sequence classification workflow is shown in Figure 1.

When classifying sequences, there is a general trade-off between sensitivity (the proportion of the total number of sequences assigned correctly) and precision (the proportion of assigned sequences assigned correctly), as well as between classification performance (combined sensitivity and precision) and computational resource requirements. Modern metagenomic sequence classification programs often use relatively fast heuristics and databases with limited sequence diversity to increase analysis speed, as metagenomic data sets often consist of millions or billions of sequences that need to be compared to millions of database sequences. Thus, while they are useful in performing a "first-pass" analysis, in a bioforensic setting it is important to validate the assignments of interest made by such programs using the most precise methods available [12]. One could choose to validate only the assignments made to the taxonomic clade of interest (e.g., *Bacillus anthracis*), but depending on the compute capacity one has access to, one might choose to validate all assignments subsumed by a higher ranking taxon (e.g., the *B. cereus* group or the *Bacillus* genus), which would enable the detection of possible false negative assignments as well as false positive assignments made by the first-pass classifier.

In this study, we present a method that uses BLAST [10] (the "gold standard" for sequence comparison), the NCBI non-redundant nucleotide database [11] (`nt`), and the NCBI taxonomy [11] to validate the assignments made by less precise sequence classification programs. For simplicity, we refer to the taxonomic clade of interest in each example as the "target taxon", and we assume all metagenomic sequences are paired-end reads generated by the Illumina HiSeq 2500 sequencer (no assembled sequences). The basic validation procedure involves comparing each read against the `nt` database using BLASTN, and then filtering and interpreting the BLAST results based on data collected from metagenomic simulations aimed at optimising detection of the target taxon. The validation workflow is shown in Figure 2.

We developed two basic ways of filtering BLAST results. The first is an "absolute" filter, which simply removes BLAST hits that do not meet a particular criterion. Various possible criteria include minimum alignment length, minimum alignment percent identity, or maximum E-value. Of these three filters, this study only uses the E-value filter (abbreviated E), as E-value is fundamentally a composite of alignment length and alignment similarity. (Our software supports the use of all three filters, however, either individually or in combination.) If the best BLAST hit matches the target taxon after application of the absolute filters, it is then possible to apply a "relative" filter by computing the difference in E-value or bit score between the best hit and the best hit to a *non-target* taxon (should the latter exist). As very small E-values are typically rounded to zero, our software uses relative bit scores in this context for maximum applicability; we call this quantity the "bit score difference", abbreviated b. If b is greater than or equal to an experimentally determined threshold, then we have validated the assignment of the read to the target taxon. Examples of the application of the bit score difference filter are given in Figures 3 and 4.

When deciding which filters to use and how they should be parameterised, an optimality criterion is needed. In the execution of bioforensic casework, it is important that any assignments made are correct, so ideally one would like to eliminate any incorrect assignments of reads to the target taxon while correctly assigning as many reads as possible. Due to inherent variation in the degree of interrelatedness among genomes, optimal filter parameter values need to be set differently for each target taxon. Here we present examples aimed at detecting a variety of pathogenic target taxa including *Bacillus anthracis*, *Clostridium botulinum*, pathogenic *Escherichia coli*, and *Yersinia pestis*. In order to determine optimal filter settings we must know the true origin

of our test sequences; thus, we simulate reads from each target taxon genome, BLAST them against `nt`, and evaluate classification performance under different combinations of filter parameter settings. The parameter optimisation workflow is shown in Figure 5. In the experiments that follow, a true positive (TP) is a simulated read from the target taxon assigned correctly (either assigned to the target taxon directly or to a more specific taxon beneath the target); a false positive (FP) is a simulated read from a non-target taxon that is incorrectly assigned to the target taxon; and a false negative (FN) is a simulated read from the target taxon assigned incorrectly (i.e., assigned to a taxon that is not part of the target taxon lineage).

The count of true positives and false negatives can be easily determined by parsing the BLAST output associated with the target taxon. Determining the count of false positives is more difficult; naively, it would require simulating reads and evaluating BLAST results for every non-target taxon in the database, but we currently regard this as computationally prohibitive. Instead, we offer two alternatives. The first, which we call "near neighbour", uses the genome in the database that is most globally similar to the target taxon as a proxy for all non-target database taxa. The intuition behind this approach is that a misclassified read (presumably due to sequencing error) is most likely to originate from a database genome that is very similar to the target taxon. Thus, with the near neighbour approach, we choose filters and parameter values that maximize sensitivity (TP/read count) while maintaining perfect precision ($TP/(TP + FP)$). The potential weakness of this approach is that there could be a region of local similarity to the target taxon in a database genome that is not the near neighbour. Thus, we offer a second approach that does not rely on selecting other genomes from the database, which we call the "false negatives" approach. This approach relies on the observation that if the sequencing error process is symmetric — i.e., the probability of an erroneous A to C substitution is the same as that of C to A, insertions are as probable as deletions, and so on — then the process that gives rise to false negatives can be treated as equivalent to the process that gives rise to false positives. While it is known that in practice this assumption of symmetry is violated [26], it may nonetheless suffice to use FN as a proxy for FP in this context. Thus, when using the false negatives approach, it is only necessary to simulate reads from the target taxon, and the calculation of precision uses the formula ($TP/(TP + FN)$). Unfortunately, deciding which of the two heuristics is more effective would require comparison to a provably optimal procedure; thus, we present simulation results from both the near neighbour and false negatives approaches and describe the patterns we observe.

Related work

Platypus Conquistador

An existing software tool, "Platypus Conquistador" [12], also uses BLAST to validate the classification of particular sequences of interest. Platypus requires the user to split their reference sequences into two databases: a database containing only sequences of possible interest, and a database composed of potential "background" sequences. BLAST queries are performed against both databases, and hits may be filtered by various combinations of percent identity or alignment length values, which need to be provided by the user. After filtering, query sequences with hits to the "interest" database are checked to see if they also have hits to the "background" database; if so, the bit scores of the respective best hits are compared and are roughly categorised as "equal", "interest > background", etc. While this could be a helpful diagnostic tool, there is no guidance provided to the user as to what parameter values to use or what difference in bit scores between interest and background should be regarded as significant. Furthermore, this tool no longer appears to be actively developed.

Genomic purity assessment

Whereas in this study we are concerned with the precision with which individual reads are classified so as to be confident in the detection of a target taxon in a metagenomic sample, a recently published study [23] addresses a different, but related problem, namely detecting contaminant organisms in ostensibly axenic (non-metagenomic) samples. Specifically, Olson et al. develop methods to determine the proportion of a contaminant required to be present in an otherwise pure material such that the contaminant can be detected with standard metagenomic sequence classification tools. As in our study, they simulate reads with ART [15] software (in their case from both "material" and potential contaminant genomes) to set up conditions under which sequence classification performance can be assessed. PathoScope [14] is used instead of BLAST for read classification. In general, they find that their method is able to identify contaminants present in a proportion of at least 10^{-3} for most contaminant-material combinations tested.

Outlier detection in BLAST hits

Shah et al. [27] have developed a method that detects outliers among BLAST hits in order to separate the hits most closely related to the query from hits that are phylogenetically more distant using a modified form of Bayesian Integral Log Odds (BILD) scores [2] and a multiple alignment scoring function. In this way, they separate sequences with confident taxonomic assignments from sequences that should be analyzed further. The method was developed for and tested on 16S rRNA data, and thus is currently not applicable to whole genome sequencing (WGS) data sets. As a general-purpose filter, however, it can be used with any organism containing 16S rRNA data, whereas our methods are optimised for detection of specific taxa. It is also interesting to note that in the Shah et al. study, BLAST is used as a first-pass classifier and subsequent analysis is performed with TIPP [21], whereas in the paradigm we present here, a much faster classifier than BLAST would be used for a first-pass (e.g., Kraken [28]), and then our BLAST-based method would be used for validation.

Results and discussion

Taxon selection

To demonstrate the BLAST-based validation procedure, we selected four target taxa, all of which are biological agents that could conceivably be of interest in a bioforensic setting. The first is *Bacillus anthracis*, the bacterium that causes anthrax. The second is *Clostridium botulinum*, a bacterium capable of producing the lethal botulinum neurotoxin. The third is a pathogenic strain of *Escherichia coli*, *E. coli* O157:H7 str. Sakai, a bacterium that has been associated with major outbreaks of foodborne illness. The fourth and final target taxon is *Yersinia pestis*, the bacterium that causes bubonic plague. Thus, three out of the four target taxa represent particular species to be identified (*B. anthracis*, *C. botulinum*, and *Y. pestis*), whereas one target taxon represents a particular strain to be identified (*E. coli* O157:H7 str. Sakai). Species-level evaluations were performed using the representative strains indicated in Table 1. In two of the three evaluations, the genome chosen was a reference genome for the species (*C. botulinum* A str. ATCC 3502 and *Y. pestis* CO92). For the *B. anthracis* evaluation, the genome of the Ames Ancestor strain was used to ensure that the pXO plasmids were included. To evaluate the implications of representative genome choice, an alternate representative genome was selected for each species. Additional information about the target taxa and near neighbours is provided in Table 1.

Experiments

A total of four experiments were performed. In each experiment, 81 different combinations of filter parameter values were tested — i.e., all combinations of maximum E-value $(E) = \{10^0, 10^{-1}, 10^{-2}, 10^{-4}, 10^{-8}, 10^{-16}, 10^{-32}, 10^{-64}, 10^{-128}\}$ and bit score difference $(b) = \{0, 1, 2, 4, 8, 16, 32, 64, 128\}$. Optimal filter parameters were chosen to maximize sensitivity (detection of the target taxon) while maintaining perfect precision (zero incorrect read assignments). In all experiments, we report the most stringent parameter values that yielded exactly the same optimal values of sensitivity and precision. A comparison of sensitivity across experiments on a per-taxon basis is available in Supplementary Figs. S1-S4, online.

250-bp simulated reads

The first experiment simulated 250-bp reads from the target and near neighbour genomes; the results are shown in Table 2.

We observe that when requiring perfect precision, sensitivity was highest for identification of *C. botulinum* (≈99%), followed by much lower sensitivity for *B. anthracis* and *Y. pestis*. These results are understandable, as it is well established that the species that comprise *B. cereus sensu lato* have very similar genomic content [5], and that *Y. pestis* and *Y. pseudotuberculosis* are also very closely related [1]. Sensitivity was lowest for identification of *E. coli* O157:H7 str. Sakai (≈0.4% for near neighbour and ≈0.08% for false negatives). Again, this result is consistent with the expectation that strain-level identification would be substantially more challenging than species-level identification, as the two *E. coli* strains in this case are ≈99.97% identical. Because the reads in this experiment were simulated from genomes that were present in the reference database, almost all read alignments had equally good scores, so the absolute E-value filter had little or no effect until it was set so stringently that it eliminated all TP ($E = 10^{-128}$). In the case of *B. anthracis*, we observe that sensitivity increased from ≈0.8% to ≈8.9% when allowing exactly one FN assignment (precision ≈99.9995%; Table 2). This suggests that if one is willing to relax the perfect precision requirement very slightly, it may be possible to make significant gains in sensitivity. Finally, it is interesting to note that in most cases, $b = 8$ maximized sensitivity while achieving perfect precision. This likely represented a "sweet spot" (at least as compared to $b = 4$ or $b = 16$) for the level of taxonomic specificity represented by the selected target taxa.

250-bp simulated reads, alternate representative genome

Choosing a particular genome to represent a strain, species, or higher-level taxon could in principle have implications for the filter parameter values recommended by the optimisation procedure. While hopefully the taxonomy is structured such that members of a particular clade are more similar to each other than to members of other clades, taxonomies are well known to be imperfect in this regard. To test the implications of representative genome choice, we repeated the species-level evaluations from the 250-bp simulated read experiment, except that we used an alternate representative genome for the target taxon, the database genome that was most distantly related to the original representative genome. The results are shown in Table 3.

In general, the optimal parameter values recommended by this experiment and the resulting values of sensitivity and precision were highly concordant with the results of the first experiment (Tables 2 and 3). The optimal parameter values recommended for classification of *B. anthracis* when using the Cvac02 strain were identical to those recommended when using the Ames Ancestor strain ($E = 10^{-64}$ and $b = 8$), with the exception that it was possible to achieve perfect precision using the false negatives approach when $b = 8$. Likewise, when using *C. botulinum* B1 str. Okra, the false

negatives approach recommended $b = 4$ rather than $b = 8$. These results suggest that the filter parameter values recommended by the false negatives approach are potentially more dependent on representative genome choice than those recommended by the near neighbor approach. The calculation of FP in the false negatives approach is based solely on the classification of reads simulated from the chosen representative genome, whereas in the near neighbour approach, FP can result from the assignment of a near neighbour read to any genome associated with the target taxon (any strain of *B. anthracis*, for example). Thus, it might behoove a user of our method to sample the diversity in their clade of interest by running the optimisation procedure for multiple representatives and using the globally most conservative recommended parameter values for classification (if maximizing precision is the goal). Alternatively, one might devise a method for more exhaustive sampling of the diversity that might exist among target taxon genomes.

150-bp simulated reads

The third experiment was identical to the first, except that a simulated read length of 150 bp was used, thus making the classification task more challenging. The results are shown in Table 4.

Under optimal filter parameter values, we observe that sensitivity in detecting each target taxon decreased relative to the 250-bp experiment — e.g., in the case of *C. botulinum*, sensitivity decreased from ≈99% to ≈96% (Tables 2 and 4). Also, optimal values for the E-value and bit score difference filters varied somewhat relative to the 250-bp experiment, although it was always the case that $E \leq 10^{-64}$ and $b \leq 8$.

Clade-level exclusion

In a final experiment, clade-level exclusion [8] of either species (*B. anthracis*) or strains (*C. botulinum* A str. ATCC 3502 and *Y. pestis* CO92) was performed to assess classification performance in the hypothetical scenario where the taxon for which one has sequence data is not represented in the reference database. In these tests, we simulated 250-bp reads from the *excluded* taxon, omitted the excluded taxon from the reference database when performing BLAST searches, and then obtained optimal filter parameter values for classification of the target taxon, which in this case was the taxon immediately above the excluded taxon in taxonomic rank. Only the false negatives method of computing FP was used; the results are shown in Table 5.

In this experiment, we observe that it was not always possible to achieve perfect precision — maximum precision for identification of the *B. cereus* group when excluding *B. anthracis* was ≈93.9%, and maximum precision for identification of *C. botulinum* when excluding *C. botulinum* A str. ATCC 3502 was ≈99.9%. We note that sensitivity for identification of *C. botulinum* decreased from ≈98.5% in the first experiment (Table 2) to ≈90.5% in the clade-level exclusion experiment (Table 5). By contrast, sensitivity for identification of *Y. pestis* hardly decreased at all (≈10.6% vs. ≈10.5%).

Calculating precision: "near neighbour" versus "false negatives"

Our results show that it was possible in every case (except for the clade-level exclusion experiments) to achieve perfect precision when using either the near neighbour or false negatives approaches. In 7/11 cases, the filter parameter values recommended by the two approaches were identical; in the cases where they differed, the false negatives approach uniformly recommended more stringent filter parameter values than the near neighbour approach, resulting in reduced sensitivity (Tables 2, 3, and 4). As mentioned previously, deciding which of the two approaches to calculating precision is superior would require a comparison to a provably optimal approach, which we currently deem

computationally prohibitive. Each heuristic makes assumptions that may not always hold: the near neighbour approach assumes that a single genome that is closely related to the target taxon is sufficient to serve as a proxy for all other non-target taxa in the database, and the false negatives approach assumes that the sequencing error process is symmetric. When seeking to avoid an erroneous claim that a particular biological agent is present in a sample, one may wish to use the more conservative set of parameter values recommended by the two approaches.

Practical application of the BLAST-based validation procedure

We have shown how BLAST, a very widely used tool for sequence similarity searches, can be used to perform taxonomic assignment with maximal precision by using BLAST result filters (that have been fine-tuned via simulation experiments) in conjunction with an LCA algorithm. Given the massive size of many metagenomic data sets, however, we envision most users employing a "two-step" approach that involves first producing candidate target taxon sequence assignments using a relatively fast classification program, one that is not necessarily optimised for precision, and then confirming the accuracy of these sequence assignments using the BLAST-based validation procedure.

One would be hard-pressed to define a "typical" metagenomic experiment, and the probability that a particular genome that is physically present in a metagenomic sample at some abundance is ultimately represented in the sequencing library and sequenced to a particular degree of coverage is a function of many factors that are outside the scope of this study. The methods we present here are concerned with read-by-read taxonomic assignment (each read interrogated independently of all other reads), and the selection of optimal BLAST result filters for this assignment process — in our case, we define "optimal" to mean correct assignment of the greatest possible number of reads without any incorrect assignments. In a real-world detection scenario, an additional question will often be asked: how many reads should be assigned to a particular target taxon before one deems it "present" in the sample? In principle, if one assumes that the reads in question originate from a genome that is present in the reference database, and that there was no error associated with the simulation process or choice of optimal filter parameter values, then the answer is simply "one read". In practice, however, if only one read out of millions or billions is assigned to a particular taxon, it is only natural that one may hesitate to claim that a potential pathogen or other biological agent is present in a sample on the basis of such scanty evidence. Unfortunately, meaningful additional guidance on this point would require a comprehensive accounting of all possible sources of error associated with the analysis of a metagenomic sample.

Conclusion

We have presented a BLAST-based method to validate specific assignments of interest made by faster classification programs, which involves using simulated reads to determine optimal parameter values to use for filtering BLAST results. We demonstrated the parameter optimisation process for four different pathogenic organisms, and showed that optimal parameter values and resulting values of sensitivity and precision varied significantly depending on the selected taxon, taxonomic rank, read length, and representation of the sequenced taxon in the reference database. Once optimal BLAST filter parameter values for a particular taxon have been determined by a procedure similar to the one presented here, they can be used subsequently to perform validation of sequence assignments to that taxon. Theoretically, however, the addition or removal of a single sequence from the reference database could change the recommended optimal parameter values, so the optimisation process should be re-run every time the

database is updated. Potential users of the software will find scripts for parsing BLAST results, performing parameter sweeps, and assigning final taxon labels to sequences at https://github.com/bioforensics/blast-validate.

Methods

Read simulation

For read simulation we used ART [15] (version 2016-06-05). To ensure thorough sampling, all experiments used simulated reads equivalent in total to 10× coverage of the source genome. For 150-bp reads, we used the built-in HS25 quality profile with an insert size of 200 ± 10 bp (mean ± standard deviation). For 250-bp reads, we used a custom Illumina HiSeq quality profile that we generated from recent runs of our HiSeq 2500, with an insert size of 868 ± 408 bp determined from recent library preparations. We provided this information so that the simulated reads would have characteristics that closely matched what we would expect to obtain from a real HiSeq run in our laboratory, thus ensuring that the simulation results would be maximally useful to us. We recommend that others who emulate our procedures customise the attributes of their simulated reads to correspond to the real data they anticipate analysing.

Selection of near neighbour and alternate representative genomes

For each target taxon, we used the "Genome neighbor report" feature of the NCBI Genome database [11] to select the most closely related complete genome of a different species or strain, as appropriate, to be used as the "near neighbour". For species-level target taxa, we used the Genome neighbor report to select the complete genome of the same species that was most distantly related to the original representative genome, which we call the "alternate representative genome" (Table 1).

Similarity searches

We used BLASTN from the BLAST+ package [10] (version 2.2.25+) together with the NCBI non-redundant nucleotide database [11] (downloaded February 2017) for all classification experiments. Default parameters were used, except when excluding taxa from the reference database, in which case the `-negative_gilist` option was added. The BLAST computation was distributed over many cluster nodes to complete the analyses in a timely manner.

BLAST result parsing

BLAST result parsing and final taxonomic assignment of each read was performed with a custom Perl script that is capable of querying the NCBI taxonomy database [11]. If a target taxon was supplied, assignments to the target taxon lineage that were more specific than the target taxon were simply reassigned to the target taxon. BLAST hits that did not meet the criteria specified by the absolute filters (minimum alignment length, minimum alignment percent identity, or maximum E-value) were removed, as were hits to the "other sequences" clade (NCBI taxon ID 28384), which were presumed to be erroneous. To finally classify each read, the lowest common ancestor (LCA) algorithm [16] was applied to the remaining hits that had a difference in bit score from the best hit less than a specified amount. If multiple parameter values were supplied for one or more filters, the script parsed the BLAST results once for each possible combination of parameter values and wrote the results to separate files, thus enabling the user to efficiently perform parameter sweeps. The final output from the script was one or more

files, each containing the final LCA classification of each read for a particular combination of filter parameter values.

Evaluating results

Counts of true positives, false positives, and false negatives (from which sensitivity and precision were calculated) were obtained using a custom Perl script that parsed the LCA files produced by the BLAST result parser.

Software and data availability

All genome data used in this study is available from the NCBI RefSeq [22] database (assembly accessions provided in Table 1). The software implementing the methods described in this study is available online at the following URL: https://github.com/bioforensics/blast-validate

References

1. M. Achtman, K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel. *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences*, 96(24):14043–14048, 1999.

2. S. F. Altschul, J. C. Wootton, E. Zaslavsky, and Y.-K. Yu. The construction and use of log-odds substitution scores for multiple sequence alignment. *PLOS Computational Biology*, 6(7):1–17, 07 2010.

3. S. K. Ames, D. A. Hysom, S. N. Gardner, G. S. Lloyd, M. B. Gokhale, and J. E. Allen. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, 29(18):2253, 2013.

4. M. Balvočiūtė and D. H. Huson. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics*, 18(2):114, 2017.

5. A. L. Bazinet. Pan-genome and phylogeny of *Bacillus cereus sensu lato*. *BMC Evolutionary Biology*, 17(1):176, Aug 2017.

6. A. L. Bazinet and M. P. Cummings. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13(1):92, May 2012.

7. D. A. Benson, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 42(D1):D32, 2014.

8. A. Brady and S. L. Salzberg. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Meth*, 6(9):673–676, 09 2009.

9. F. P. Breitwieser, J. Lu, and S. L. Salzberg. A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, page bbx120, 2017.

10. C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

11. N. R. Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(Database issue):D7–D19, 01 2016.

12. A. Gonzalez, Y. Vázquez-Baeza, J. B. Pettengill, A. Ottesen, D. McDonald, and R. Knight. Avoiding pandemic fears in the subway and conquering the platypus. *mSystems*, 1(3), 2016.

13. I. Gregor, J. Dröge, M. Schirmer, C. Quince, and A. C. McHardy. *PhyloPythiaS+*: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ*, 4:e1603, Feb. 2016.

14. C. Hong, S. Manimaran, Y. Shen, J. F. Perez-Rogers, A. L. Byrd, E. Castro-Nallar, K. A. Crandall, and W. E. Johnson. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome*, 2(1):33, 2014.

15. W. Huang, L. Li, J. R. Myers, and G. T. Marth. ART: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593, 2011.

16. D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–386, 2007.

17. D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12):1721–1729, 2016.

18. D. Koslicki and D. Falush. MetaPalette: a k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. *mSystems*, 1(3), 2016.

19. T. Kulikova, P. Aldebert, N. Althorpe, W. Baker, K. Bates, P. Browne, A. van den Broek, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, M. Garcia-Pastor, N. Harte, C. Kanz, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, P. Stoehr, G. Stoesser, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler. The EMBL nucleotide sequence database. *Nucleic Acids Research*, 32(suppl_1):D27, 2004.

20. J. Mashima, Y. Kodama, T. Fujisawa, T. Katayama, Y. Okuda, E. Kaminuma, O. Ogasawara, K. Okubo, Y. Nakamura, and T. Takagi. DNA data bank of Japan. *Nucleic Acids Research*, 45(D1):D25, 2017.

21. N.-p. Nguyen, S. Mirarab, B. Liu, M. Pop, and T. Warnow. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics*, 30(24):3548–3555, 2014.

22. N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733, 2015.

23. N. D. Olson, J. M. Zook, J. B. Morrow, and N. J. Lin. Challenging a bioinformatic tool's ability to detect microbial contaminants using *in silico* whole genome sequencing data. *PeerJ*, 5:e3729, Sept. 2017.

24. R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16(1):236, 2015.

25. L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter. Pseudoalignment for metagenomic read assignment. *ArXiv e-prints*, Oct. 2015.

26. M. Schirmer, R. D'Amore, U. Z. Ijaz, N. Hall, and C. Quince. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, 17(1):125, 2016.

27. N. Shah, S. F. Altschul, and M. Pop. Outlier detection in BLAST hits. In R. Schwartz and K. Reinert, editors, *17th International Workshop on Algorithms in Bioinformatics (WABI 2017)*, volume 88 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 23:1–23:11, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

28. D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014.

Acknowledgements

We thank Todd Treangen and Nicholas Bergman for helpful discussions about the project, and Travis Wyman, Timothy Stockwell, and Lisa Rowe for providing feedback on drafts of the manuscript.

Author contributions statement

ALB conceived the experiments and all authors participated in the intellectual development of the project. BDO, DDS, and ALB wrote the software. All authors took part in conducting the experiments and analysing the results. ALB drafted the manuscript and all authors reviewed and revised the text.

Additional information

Competing financial interests

The authors declare no competing financial interests.

Funding

This work was funded under Contract No. HSHQDC-15-C-00064 awarded by the Department of Homeland Security (DHS) 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 author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the DHS or S&T. In no event shall DHS, NBACC, S&T or Battelle National Biodefense Institute have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. DHS does not endorse any products or commercial services mentioned in this publication.

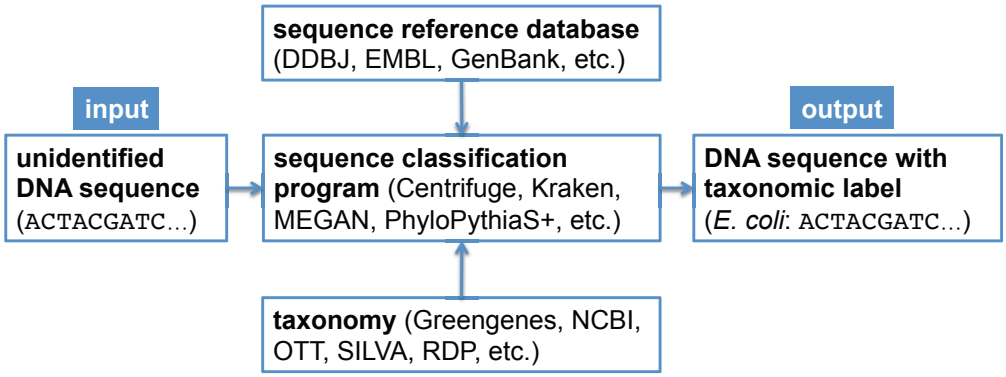


Figure 1. Canonical workflow for the classification of metagenomic sequences.

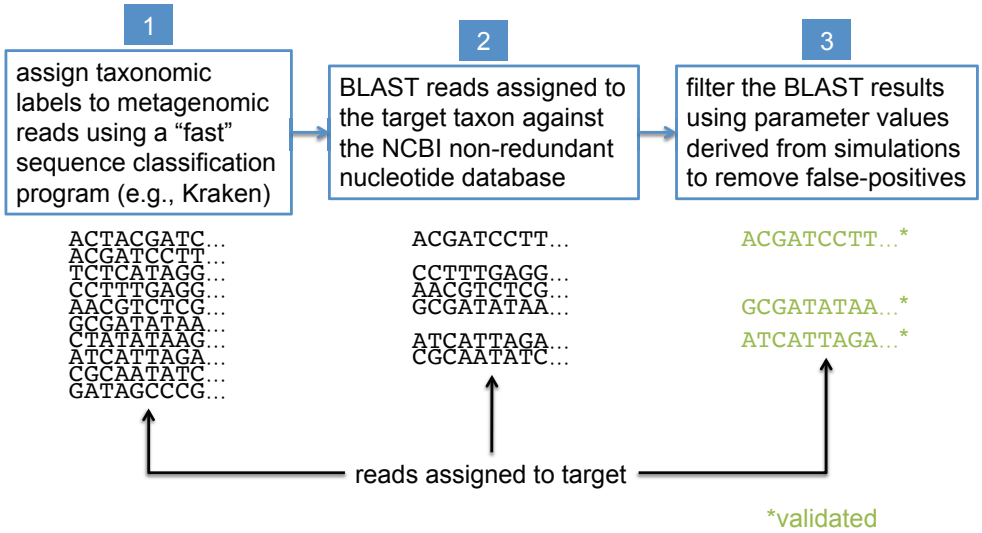


Figure 2. Workflow for BLAST-based validation of taxonomic assignments.

```
# BLASTN 2.2.25+
# Query: NC_003997.3-1994/1
# Database: /data/ncbi/blast/db/nt
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score, taxon
# 372 hits found
NC_003997.3-1994/1	gi|1043168508|gb|CP012728.1|	100.00	250	0	0	1	250	1704042	1703793	2e-123	452	Bacillus anthracis
NC_003997.3-1994/1	gi|1043163212|gb|CP012727.1|	100.00	250	0	0	1	250	1704042	1703793	2e-123	452	Bacillus anthracis
NC_003997.3-1994/1	gi|1043157917|gb|CP012726.1|	100.00	250	0	0	1	250	1704041	1703792	2e-123	452	Bacillus anthracis
NC_003997.3-1994/1	gi|1043152620|gb|CP012725.1|	100.00	250	0	0	1	250	1704043	1703794	2e-123	452	Bacillus anthracis
NC_003997.3-1994/1	gi|1043147323|gb|CP012724.1|	100.00	250	0	0	1	250	1704043	1703794	2e-123	452	Bacillus anthracis
NC_003997.3-1994/1	gi|589079678|gb|CP006742.1|	100.00	250	0	0	1	250	1690737	1690488	2e-123	452	Bacillus anthracis str. SVA11
NC_003997.3-1994/1	gi|570716017|gb|CP001970.1|	100.00	250	0	0	1	250	1703973	1703724	2e-123	452	Bacillus anthracis str. A16
NC_003997.3-1994/1	gi|570710095|gb|CP001974.1|	100.00	250	0	0	1	250	1703925	1703676	2e-123	452	Bacillus anthracis str. A16R
NC_003997.3-1994/1	gi|384383725|gb|CP002091.1|	100.00	250	0	0	1	250	1694749	1694500	2e-123	452	Bacillus anthracis str. H9401
NC_003997.3-1994/1	gi|229264291|gb|CP001598.1|	100.00	250	0	0	1	250	1703999	1703750	2e-123	452	Bacillus anthracis str. A0248
NC_003997.3-1994/1	gi|227002338|gb|CP001215.1|	100.00	250	0	0	1	250	2535117	2535366	2e-123	452	Bacillus anthracis str. CDC 684
NC_003997.3-1994/1	gi|218534755|gb|CP001283.1|	100.00	250	0	0	1	250	1762619	1762370	2e-123	452	Bacillus cereus AH820
...
```

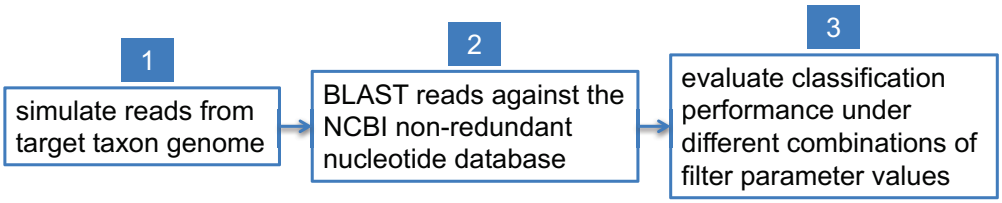
- **bit score of the best hit to the target taxon (*B. anthracis*) = 452**
- **bit score of the best hit to a non-target taxon (*B. cereus*) = 452**
- **452 – 452 = 0; minimum required bit score difference determined by simulation = 8**
- **therefore, this read is <u>not</u> assigned to the target taxon (*B. anthracis*)**

Figure 3. Demonstration of the bit score difference filter (first example).

```
# BLASTN 2.2.25+
# Query: NC_003997.3-1992/1
# Database: /data/ncbi/blast/db/nt
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score, taxon
# 307 hits found
NC_003997.3-1992/1	gi|753450531|gb|CP009541.1|	100.00	250	0	0	1	250	322480	322231	2e-123	452	Bacillus anthracis str. Sterne
NC_003997.3-1992/1	gi|753270362|gb|CP009315.1|	100.00	250	0	0	1	250	4108350	4108599	2e-123	452	Bacillus anthracis str. Turkey32
NC_003997.3-1992/1	gi|673992625|gb|CP007666.1|	100.00	250	0	0	1	250	1176391	1176640	2e-123	452	Bacillus anthracis str. Vollum
NC_003997.3-1992/1	gi|589079678|gb|CP006742.1|	100.00	250	0	0	1	250	2289646	2289895	2e-123	452	Bacillus anthracis str. SVA11
NC_003997.3-1992/1	gi|570716017|gb|CP001970.1|	100.00	250	0	0	1	250	2303524	2303773	2e-123	452	Bacillus anthracis str. A16
NC_003997.3-1992/1	gi|570710095|gb|CP001974.1|	100.00	250	0	0	1	250	2303013	2303262	2e-123	452	Bacillus anthracis str. A16R
NC_003997.3-1992/1	gi|384383725|gb|CP002091.1|	100.00	250	0	0	1	250	2293805	2294054	2e-123	452	Bacillus anthracis str. H9401
NC_003997.3-1992/1	gi|229264291|gb|CP001598.1|	100.00	250	0	0	1	250	2303542	2303791	2e-123	452	Bacillus anthracis str. A0248
NC_003997.3-1992/1	gi|227002338|gb|CP001215.1|	100.00	250	0	0	1	250	1935455	1935206	2e-123	452	Bacillus anthracis str. CDC 684
NC_003997.3-1992/1	gi|49176966|gb|AE017225.1|	100.00	250	0	0	1	250	2303569	2303818	2e-123	452	Bacillus anthracis str. Sterne
NC_003997.3-1992/1	gi|30260185|gb|AE016879.1|	100.00	250	0	0	1	250	2303518	2303767	2e-123	452	Bacillus anthracis str. Ames
NC_003997.3-1992/1	gi|675832311|gb|CP008853.1|	99.60	250	1	0	1	250	2303095	2303344	9e-122	446	Bacillus anthracis
NC_003997.3-1992/1	gi|755995789|gb|CP009605.1|	99.20	250	2	0	1	250	1841890	1841641	1e-120	443	Bacillus cereus
...
```

- **bit score of the best hit to the target taxon (*B. anthracis*) = 452**
- **bit score of the best hit to a non-target taxon (*B. cereus*) = 443**
- **452 – 443 = 9; minimum required bit score difference determined by simulation = 8**
- **therefore, this read <u>is</u> assigned to the target taxon (*B. anthracis*)**

Figure 4. Demonstration of the bit score difference filter (second example).



true positive (*TP*) = a simulated read from the *target* taxon assigned correctly

false negative (*FN*) = a simulated read from the *target* taxon assigned incorrectly

false positive (*FP*) = a simulated read from a *non-target* taxon incorrectly assigned to the *target* taxon

sensitivity = (*TP* / *read count*)

precision = (*TP* / (*TP* + *FP*))[1] or (*TP* / (*TP* + *FN*))[2]

[1] Standard definition of precision, used with the near neighbour approach.

[2] Non-standard definition of precision where *FN* are used as a surrogate for *FP*.

Figure 5. Workflow for determining optimal BLAST filter parameter values.

Taxon	Taxonomic rank	Type	NCBI Taxonomy ID	RefSeq assembly accession
Bacillus cereus group	species group	target	86661	N/A
B. anthracis	species	target	1392	N/A
B. cereus	species	near neighbour	1396	N/A
B. anthracis Ames Ancestor	strain	representative	261594	GCF_000008445
B. anthracis Cvac02	strain	representative (alternate)	N/A	GCF_000747335
B. cereus JEM-2	strain	representative	N/A	GCF_001941925
Clostridium	genus		1485	N/A
C. botulinum	species	target	1491	N/A
C. sporogenes	species	near neighbour	1509	N/A
C. botulinum A str. ATCC 3502	strain	representative	413999	GCF_000063585
C. botulinum B1 str. Okra	strain	representative (alternate)	498213	GCF_000019305
C. sporogenes NCIMB 10696	strain	representative	N/A	GCF_000973705
Escherichia	genus		561	N/A
E. coli	species		562	N/A
E. coli O157:H7 str. Sakai	strain	target	386585	GCF_000008865
E. coli SRCC 1675	strain	near neighbour	N/A	GCF_001612495
Yersinia pseudotuberculosis complex	species group		1649845	N/A
Y. pestis	species	target	632	N/A
Y. pseudotuberculosis	species	near neighbour	633	N/A
Y. pestis CO92	strain	representative	214092	GCF_000009065
Y. pestis Angola	strain	representative (alternate)	349746	GCF_000018805
Y. pseudotuberculosis PB1/+	strain	representative	502801	GCF_000834475

Table 1. Taxonomic data and metadata for target taxa and near neighbour species or strains.

Target taxon	Taxonomic rank	Number of reads	Method of computing FP	Maximum E-value	Bit score difference	Validated reads	Sensitivity	Precision
B. anthracis	species	220,140	near neighbour	10^{-64}	8	19,491	0.088539	1.0
			false negatives	10^{-64}	8	19,491	0.088539	0.999995[1]
			false negatives	10^{-64}	128	1,751	0.007954	1.0
C. botulinum	species	156,120	near neighbour	10^{-64}	8	153,786	0.985050	1.0
			false negatives	10^{-64}	8	153,786	0.985050	1.0
E. coli O157:H7 str. Sakai	strain	223,760	near neighbour	10^{-64}	1	838	0.003745	1.0
			false negatives	10^{-64}	8	184	0.000822	1.0
Y. pestis	species	193,170	near neighbour	10^{-64}	8	20,398	0.105596	1.0
			false negatives	10^{-64}	8	20,398	0.105596	1.0

[1] In the case of *B. anthracis*, we observe that sensitivity increased from ≈0.8% to ≈8.9% when allowing exactly one FN assignment (precision ≈99.9995%).

Table 2. Experimental results using simulated 250-bp reads.

Target taxon	Taxonomic rank	Number of reads	Method of computing FP	Maximum E-value	Bit score difference	Validated reads	Sensitivity	Precision
B. anthracis	species	209,080	near neighbour	10^{-64}	8	20,114	0.096202	1.0
			false negatives	10^{-64}	8	20,114	0.096202	1.0
C. botulinum	species	164,270	near neighbour	10^{-64}	8	162,069	0.986601	1.0
			false negatives	10^{-64}	4	162,594	0.989797	1.0
Y. pestis	species	187,470	near neighbour	10^{-64}	8	22,576	0.120425	1.0
			false negatives	10^{-64}	8	22,576	0.120425	1.0

Table 3. Experimental results using simulated 250-bp reads from alternate representative genomes.

Target taxon	Taxonomic rank	Number of reads	Method of computing FP	Maximum E-value	Bit score difference	Validated reads	Sensitivity	Precision
B. anthracis	species	366,920	near neighbour	10^{-32}	8	16,904	0.046070	1.0
			false negatives	10^{-32}	8	16,904	0.046070	1.0
C. botulinum	species	260,200	near neighbour	10^{-32}	8	250,915	0.964316	1.0
			false negatives	10^{-32}	8	250,915	0.964316	1.0
E. coli O157:H7 str. Sakai	strain	372,960	near neighbour	10^{-64}	4	709	0.001901	1.0
			false negatives	10^{-64}	8	180	0.000483	1.0
Y. pestis	species	321,970	near neighbour	10^{-32}	8	19,965	0.062009	1.0
			false negatives	10^{-32}	8	19,965	0.062009	1.0

Table 4. Experimental results using simulated 150-bp reads.

Target taxon	Taxonomic rank	Excluded taxon	Number of reads	Maximum E-value	Bit score difference	Validated reads	Sensitivity	Precision
B. cereus group	species group	*B. anthracis*	220,140	10^{-64}	64	36,003	0.163546	0.939339
C. botulinum	species	*C. botulinum* A str. ATCC 3502	156,120	10^{-64}	32	141,277	0.904926	0.999385
Y. pestis	species	*Y. pestis* CO92	193,170	10^{-64}	8	20,272	0.104944	1.0

Table 5. Clade-level exclusion experimental results using simulated 250-bp reads. Only the false negatives method of computing FP was used.