A peer-reviewed version of this preprint was published in PeerJ on 2 November 2018.

View the peer-reviewed version (peerj.com/articles/5882), which is the preferred citable publication unless you specifically need to cite this preprint.

Castro JC, Rodriguez-R LM, Harvey WT, Weigand MR, Hatt JK, Carter MQ, Konstantinidis KT. 2018. imGLAD: accurate detection and quantification of target organisms in metagenomes. PeerJ 6:e5882 https://doi.org/10.7717/peerj.5882

imGLAD: Accurate detection and quantification of target organisms in metagenomes

Juan Castro [1, 2] , **Luis M Rodriguez-R** [1, 3] , **Michael R Weigand** [4] , **Janet K Hatt** [3] , **Michell Q Carter** [5] , **Konstantinos T Konstantinidis** [Corresp. 1, 2, 3]

[1] Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, Georgia, United States

[2] School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, United States

[3] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, United States

[4] Center for Disease Control and Prevention, Atlanta, Georgia, United States

[5] Produce Safety and Microbiology, USDA-ARS Western Regional Research Center, U.S. Department of Agriculture, Albany, California, United States

Corresponding Author: Konstantinos T Konstantinidis
Email address: kostas@ce.gatech.edu

Accurate detection of target microbial species in metagenomic datasets from environmental samples remains limited because the limit of detection of current methods is typically inaccessible and the frequency of false-positives, resulting from inadequate identification of regions of the genome that are either too highly conserved to be diagnostic (*e.g.*, rRNA genes) or prone to frequent horizontal genetic exchange (*e.g.*, mobile elements) remains unknown. To overcome these limitations, we introduce imGLAD, which aims to detect genomic sequences in metagenomic datasets. imGLAD achieves high accuracy because it uses the sequence-discrete population concept for discriminating between metagenomic reads originating from the target organism compared to reads from co-occurring close relatives, masks regions of the genome that are not informative using the MyTaxa engine, and models both the sequencing breadth and depth to determine relative abundance and limit of detection. We validated imGLAD by analysing metagenomic datasets derived from spinach leafs inoculated with the enteric pathogen *Escherichia coli* O157:H7 and showed that its limit of detection is comparable to that of PCR-based approaches (~1 cell/gram).

imGLAD: accurate detection and quantification of target organisms in metagenomes

Juan C. Castro[1,2], Luis M. Rodriguez[1,2,3], Michael R. Weigand[3+], Janet K. Hatt[3], Michell Q. Carter[4] , and Konstantinos T. Konstantinidis[1,2,3+,*]

[1] School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA,
[2] Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, GA 30332, USA,
[3] School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA,
[4] Produce Safety and Microbiology, USDA-ARS Western Regional Research Center, Albany, CA 94710, USA and

[+]Present Address: Michael R. Weigand, Division of Bacterial Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30329, USA

* To whom correspondence should be addressed.

Konstantinos T. Konstantinidis,
School of Civil & Environmental Engineering,
Georgia Institute of Technology.
311 Ferst Drive, ES&T Building, Room 3321,
Atlanta, GA, 30332.
Telephone: 404-639-4292
Email: kostas@ce.gatech.edu

ABSTRACT

Accurate detection of target microbial species in metagenomic datasets from environmental samples remains limited because the limit of detection of current methods is typically inaccessible and the frequency of false-positives, resulting from inadequate identification of regions of the genome that are either too highly conserved to be diagnostic (e.g., rRNA genes) or prone to frequent horizontal genetic exchange (e.g., mobile elements) remains unknown. To overcome these limitations, we introduce imGLAD, which aims to detect genomic sequences in metagenomic datasets. imGLAD achieves high accuracy because it uses the sequence-discrete population concept for discriminating between metagenomic reads originating from the target organism compared to reads from co-occurring close relatives, masks regions of the genome that are not informative using the MyTaxa engine, and models both the sequencing breadth and depth to determine relative abundance and limit of detection. We validated imGLAD by analysing metagenomic datasets derived from spinach leafs inoculated with the enteric pathogen *Escherichia coli* O157:H7 and showed that its limit of detection is comparable to that of PCR-based approaches (~1 cell/gram).

INTRODUCTION

Assessment of the minimum amount of sequencing required for accurate detection of target bacterial species in a background of a complex microbial community remains challenging. This problem has important practical applications in environmental and clinical surveillance studies. Detection limits vary depending on the sequencing effort and technology (e.g., read length), and the complexity of the microbial community sampled, and in most cases these parameters or their effects on the limit of detection remain inaccessible. Experiments with increasing amounts of target DNA added to environmental samples have been performed in the past to empirically establish detection limits [e.g., (1)]. However, a theoretical framework to establish limit of detection based on bioinformatics analysis of metagenomics is still lacking. Furthermore, such empirical approaches are typically computationally expensive, cumbersome, and specific to the system tested.

Several methods to evaluate presence or absence of bacterial species based on best match or Bayesian analysis of read mapping patterns against a reference collection of genome sequences such as Pathoscope or Sigma (2,3) have been recently developed. Additionally, taxonomic profilers such as MetaPhlAn (4,5) or MetaMLST (6) employ species- or strain-specific genetic markers to identify the different members of the community. However, these approaches rely on Single Nucleotide Polymorphism (SNPs) pattern differences against reference genes/genomes, which are difficult to robustly determine, especially in cases of low abundance (i.e., not enough reads available to reliably call SNPs), and are typically computationally intensive. Importantly, no available tool can detect organisms that are not part of a reference genome database, and most tools are not easily adaptable to include new target genomes as references (e.g., the tools require re-computation of the -typically large- training datasets or reference database to include new target organisms). Further, it is not clear how most of these tools perform when relatives of varying relatedness to the target organisms are co-occurring in the sample, as often is the case of environmental samples, and whether or not strain-level resolution can be achieved.

Here, we present imGLAD (*in-silico* metagenomes for Genome Low-Abundance Detection), a new algorithm that incorporates a training step and several computational optimizations in order to address the abovementioned limitations. Application of imGLAD to metagenomes derived from samples of known composition (mock) showed that it can reliably detect target organisms of interest in a background of closely related co-occurring relatives and frequently outperforms other methods.

MATERIAL AND METHODS

Overview of the imGLAD pipeline

imGLAD assumes that reads of a metagenomic dataset originate at random from all regions of the genome. Thus, the fraction of the genome that is recovered in the dataset (sequencing breadth) as well as the number of times each region is sequenced (sequencing depth), both depend on the abundance of the

organism in the community. Highly conserved regions (e.g., rRNA and tRNA genes), as well as regions resulting from recent horizontal gene transfer (e.g., transposase and integrase genes), can recruit reads from other non-target genomes and misleadingly increase the value of sequencing depth (and hence, estimated relative abundance) in some datasets depending on the gene composition of the organisms present. To address this problem, we developed a framework to identify which fraction of a target genome corresponds to reads that belong to the target and what fraction is the result of spurious matches. This framework has two steps: initial training and subsequent prediction (Figure 1). Training set selection can be automatic or user defined. The automatic training generates reads from a randomly selected number of genomes (default is 200 genomes) from RefSeq (7), and builds *in-silico*-generated datasets of about 1 million reads each. Simulated reads from the target genome(s) are then generated in a similar way and added to the former datasets in order to create the positive datasets with decreasing target abundances. Reads from the target genome(s) are omitted for the construction of negative datasets. All other genomes used to create the datasets are sampled in equal proportions (i.e., even richness). The user can also choose the genomes to use to generate the training set (e.g., genomes previously known to co-occur in the same environment). In this case, the construction of the training set will be performed based on these genomes rather than the default genome collection from RefSeq. Simulated Illumina-like reads are generated using ART-MountRainier (8) with default settings. Simulation of reads from additional sequencing platforms is provided as an option. Reads from both positive and negative samples are then recruited against the target genome sequence (reference) using BLAT (9) (or BLAST (10)). By default, reads with identity higher than 95% and at least 90% of the read length aligned are selected to calculate sequencing breadth and sequencing depth, after normalizing for the size of the dataset. This level of identity has been shown to capture well the sequence-discrete populations recovered frequently in metagenomes of natural habitats (11), although different user-defined cut-offs can be used as well. Sequencing depth (SD) is calculated as the number of reads mapping to the genome (N) multiplied by the read length (L) divided by the total length of the genome (G), and sequencing breadth (SB) is calculated as the number of bases covered (B) divided by the total length of the genome as follows.

$$SD = L*N/G \quad (1)$$

$$SB = B/G \quad (2)$$

A logistic function is fitted to the resulting recruitment data that attempts to separate the positive from the negative training datasets in terms of sequencing depth and sequencing breadth. The logistic function is then used to estimate the probability of presence in the prediction step based on the sequencing depth and/or breadth of the target genome in a query (unknown) metagenome (see also below). imGLAD includes an additional, optional step in which MyTaxa (10) can be used to identify and mask the reference genes that do not have robust phylogenetic signals at the species level due either to insufficient diversity (high sequence conservation) or because of frequently undergoing horizontal gene transfer. This step, in general, improves results for several datasets and genomes (see for instance Fig 4).

Model training

The logistic model estimates the probability of presence based on two predictor variables, i.e., sequencing breadth, or a combination of sequencing breadth and sequencing depth. Regression coefficients are calculated for these variables as well as for an intercept term and thus, the final model estimates three parameters, i.e., sequencing depth, breadth and intercept. These parameters are estimated via a training set that consists of at least 200 simulated metagenome-like datasets. These datasets include 100 datasets with the target genome (positive samples), and 100 samples without this genome (negative samples).

Construction of training datasets

For each dataset, BLAT (9) is employed to align the reads to the target genome sequence. Alternatively, BLAST can be used to improve sensitivity at the expense of computational time (10). The resulting alignments are used to calculate the sequencing depth and sequencing breadth of the target genome in each dataset using the nucleotide cut-off mentioned above. Sequencing breadth is calculated as the fraction of bases of the total genome sequence that recruit at least one read (equation **2** above). If the genome consists of more than one contig (e.g., draft genomes), the length is assumed to be the sum of the lengths of all contigs. Sequencing depth is calculated in an analogous way by counting the times that each base of the reference genome sequence is covered by a read, on average (equation **1** above).

Sequencing breadth and depth are subsequently used to calculate the regression parameters of a logistic model using log likelihood maximization via gradient, which modifies the parameters values until the error is minimized. This approach calculates the optimal set of parameters by computing the error in the training set (i.e., what sequencing breadth and/or depth values are observed for positive vs. negative samples) and modifying the parameter accordingly to reduce the error until convergence is reached. Final parameters of the model are estimated by default only based on sequencing breadth, as this variable was found to be the most discriminating parameter for positive vs. negative samples (see also below). However, an estimation including sequencing depth is also provided as an option in order to produce, in addition to the probability of presence/absence, an accurate estimation of the abundance of the target genome.

Probability estimation

Once the logistic model has been built, sequencing breadth can be used to reliably predict the probability of presence of the target genome in any number of query metagenomes after the reads of the query have been recruited against the target genome as described above for training datasets. The probability of presence is estimated according to:

$$p = 1 - \frac{1}{1 - e^{-z}} \quad (3)$$

Where **z** is a linear function of the form $\beta^T t$, β represents the regression parameters and ***t*** is either a vector composed of the sequencing depth (***d***) and sequencing breadth (***b***) or by default, a one-dimensional variable corresponding to ***b***. Based on the model parameters, it is possible to establish a detection limit for the target genome in each metagenomic dataset analysed. This limit is defined as the minimum fraction (sequencing breadth) that needs to be sampled in order to estimate a probability of presence at 0.95. The result is displayed as a black solid line in a 2D plot of sequencing breadth and sequencing depth (e.g., Figure 2). The sequencing depth value observed based on the read recruitment, when corresponding to a probability value equal or higher to 0.95, is then used to estimate the relative abundance of the organism in the sample. The sequencing depth corresponding to 0.95 probability then provides the limit of detection in terms of relative abundance.

Filtering of conserved regions

To avoid spurious results from reads mapping on conserved or highly mobile regions of the (target) genome, the user can create a filter for these regions using MyTaxa. This filter is created by predicting genes in the target genome and determining their classification weight using MyTaxa. If the MyTaxa classification score is at the bottom 5% or the gene is not scored (e.g., some hypothetical proteins), the gene is removed from the genome and further analysis. The filtered version of the genome is subsequently used for the model training and probability estimation steps.

Parameters of software used

MetaPhlAn V2 (5) was run with the default settings using Bowtie version 2.2.8 (13) for read mapping. MetaMLST (6) was used with default settings. PathoScope 2.0 (3) was run with default settings, using the same set of reference genomes that were used to build the training datasets for imGLAD.

Four tests were performed to assess specificity and sensitivity. In all cases, sensitivity was calculated as the proportion of properly classified positive datasets among the total number of positive datasets. Specificity was defined instead as the fraction of correctly identified negative datasets among all negative datasets examined. For the first test, metagenomic datasets were created with similar parameters to the training dataset of *E. coli* (i.e., 100 datasets from RefSeq genomes). These datasets were spiked with seven different concentrations of the *E. coli* genome, ranging from 1% to 7%. In the second test, Human Microbiome Project (HMP) metagenomes were spiked with reads from the *E. coli* ranging from 1% to 7% relative abundance. 571 HMP datasets were used for each *E. coli* concentration. In the third test, the datasets constructed in test 1 were spiked with reads from close relatives of *E. col*, i.e., *Klebsiella* (81% ANI), *Salmonella* (82% ANI), and *Escherichia fergusoni* (92% ANI), at random concentrations for each genome in addition to the *E. coli* reads. Finally, a test using close relatives, e.g., >95% ANI, was performed in the HMP datasets in a similar way as described above for test #3.

Leaf inoculation experiments to test imGLAD

Fifty grams of field grown spinach leaves were inoculated (spiked in) with cells of *Escherichia coli* O157:H7 strain RM6067, a strain linked to the 2006 spinach-associated outbreak in the U.S.A (14). Three serial dilutions were performed resulting in three inoculation concentrations: 80, 8 x 10^3 and 8 x 10^5 cells per pellet, plus a control sample with no inoculated cells. Cells for inoculation were obtained from single colonies that were grown overnight, and cell concentrations were determined by enumeration of colony forming units (CFUs) on LB agar plates. Leaves were subsequently washed, the leaf wash was filtered to remove plant debris, and leaf-associated microorganisms were pelleted by centrifugation at 10,000g for 10 min at 4°C. DNA extraction was performed using MoBio UltraClean Microbial DNA isolation kit according to manufacturer's instruction (MoBio).

DNA sequencing libraries were prepared using the Illumina Nextera XT DNA library prep kit according to manufacturer's recommendations, except that the protocol was terminated after isolation of cleaned amplified double stranded libraries. Library concentrations were determined by fluorescent quantification using a Qubit HS DNA kit and Qubit 2.0 fluorometer (ThermoFisher Scientific, formerly Life Technologies) according to manufacturer's recommendations and libraries were run on a High Sensitivity DNA chip using the Bioanalyzer 2100 instrument (Agilent) to determine average library insert sizes. An equimolar mixture of the libraries (final loading concentration of 11 pM) was sequenced using a MiSeq reagent v3 kit for 600 cycles (2 x 300 bp paired end run) on an in-house Illumina MiSeq instrument (Georgia Institute of Technology), running the MiSeq control software v2.4.0.4 (MCS). Adapter trimming and demultiplexing of sequenced samples was carried out by the MCS. Additionally, we used metagenomic datasets inoculated with *Bacillus anthracis* DNA, which were made available previously (1).

Availiability and dependencies of imGLAD

imGLAD is available through http://enve-omics.ce.gatech.edu/imGLAD/. Source code is available under GNU General Public License v3.0 https://github.com/jccastrog/imGLAD. imGLAD execution requires BLAT or BLAST to be installed, ART and the Python modules "scipy", "numpy", "screed", "statsmodels", and "Bio".

RESULTS

Training set for *E. coli* and *B. anthracis*

We evaluated imGLAD's performance on training datasets with *E. coli* strain O157:H7 EC4115, a strain almost genetically identical to RM6067 used in the spinach inoculation experiments, i.e., >99.9% average nucleotide identity (or ANI), and *B. anthracis* strain Ames as target genomes. The datasets included closely related species of the same genus with ANI around 95% or lower (Figure 1; see also below for evaluation of the effect of relatedness of co-occurring genomes), which corresponds to the frequently used standard for species demarcation (15) and encompass the sequence-discrete populations recovered frequently in metagenomes of natural habitats (11). Although the predicted detection limit (from the training step) varied slightly for each species, it was always possible to have confident detection (probability of presence >99%) when sequencing breadth was about 0.03 (or 3% of the total genome) or more based on

the training datasets used (Figure 2 & Table 1). The model for *E. coli* was able to accurately separate positive from negative samples (probability of presence >95%) to a minimal value of sequencing breadth of 0.01 (Figure 2A).

The logistic models from the training datasets were then applied to metagenomic datasets originating from environmental samples, spiked-in with the target genome (see Materials and Methods for details). Samples from spinach leaf surface spiked with different concentrations of *E. coli* strain O157:H7 cells were sequenced at about 1-2 Gbp/sample on an Illumina Mi-Seq platform. Four different concentrations were tested, three with different inoculum concentrations, 80, 8 x 10^3 and 8 x 10^5 target *E. coli* cells per leaf microbiome that was recovered from 100g of spinach leaf material, and a negative control (no cells were spiked in; although *E. coli* cells might have been present in the background leaf microbial community in low concentrations). The model was able to detect the target *E. coli* genome in all samples, even as low as 80 cells (Table 1). For the negative control, imGLAD provided values of sequencing breadth (0.007) and sequencing depth (0.042) that were consistent with the values of negative samples in the training set, i.e., the target genome was not present at the limit of detection of the approach (Figure 2A; p-value for presence: 0.847, Table 1). The matching reads in this case probably originated from natural *E. coli* populations present on the spinach leaves at low abundance or close relatives and/or spurious matches (Suppl Fig 1).

The datasets made available by Be and colleagues consisted of a soil microbial community DNA sample spiked with known quantities (genome equivalents) of DNA of *B. anthracis* strain Ames (Table 2), and sequenced using the Illumina technology (1). A training set for *B. anthracis* was built in a similar way to the *E. coli* set; however, genomes that belong to *Bacillus cereus* were excluded from the training dataset in this case as they show ANI values higher than 95% to *B. anthracis*. Based on the training datasets, a slightly higher limit of detection than the one for *E. coli* was obtained (probability of presence >95%), with a minimum value of sequencing breadth of 0.039 (Figure 2B). Among the 6 samples tested, a significant probability of presence (p > 99%) was obtained in samples with 100 (3.8% of the genome recovered), 1,000 (56.2%), 10^4 (98.3%), and 10^5 (99.9%) *B. anthracis* genomes. The samples with lower genome copy number (1 and 10 genomes) were not identified as positive. Manual inspection of the number and position of matching reads to the *B. anthracis* reference genome in the latter two datasets revealed about 2000 reads for the 10 genome copy dataset and about 4000 reads for the 1 genome copy dataset, i.e., more reads were obtained with the lower abundance dataset, indicating spurious matches (each dataset was on average 5.6 Gbp in size). Further, the reads were concentrated in a few regions of the genome (not randomly distributed), which was indistinguishable from negative datasets (Table 2, Figure 2B, and Suppl Fig 2). Thus, it appears that the *B. anthracis* genomes might not have been sequenced adequately in the low copy number datasets. This interpretation is also consistent with estimates that 100 Gbp or more are required to cover the complete genome diversity within typical soil microbial communities as described previously (16) and the conclusions of the original study by Be and colleagues.

Comparison to other tools

We compared the performance of imGLAD with other available platforms that can be used to identify the taxa present in the sample. It should be pointed out, however, that these tools do not target a specific organism/genome of interest but instead assess the total microbial community composition and thus, their objective is slightly different than imGLAD's. Nonetheless, we were able to obtain meaningful results by comparing imGLAD with popular tools for these purposes such as MetaPhlAn, MetaMLST, and Pathoscope, which illustrated the advantages of imGLAD. In the *E. coli* and *B. anthracis* metagenomes described above, imGLAD provided higher sensitivity than other tools, especially at low levels of sequencing breadth. For instance, with 2% of the genome covered by sequencing reads in training datasets, imGLAD can accurately classify as positive 95% of the datasets, whereas Pathoscope and MetaPhlAn can classify only about 47% and 16% of the datasets, respectively. Only when sampling 7% of the genome or more, did these tools yield similar results to imGLAD (Figure 3A). It should be noted that 7% is more than twice the genome breadth (i.e., 3% of the genome) that imGLAD required to reach 100% classification sensitivity. Comparisons against available tools that require higher target abundance to make confident calls such as ConStrains (17) were not attempted as this would have been an unfair comparison.

Additionally, we used a set of 571 metagenomic datasets of the HMP project (http://www.hmpdacc.org/), in which different concentrations of *E. coli* (target organism) reads were spiked to further test the specificity of imGLAD against a naturally occurring background community (as opposed to *in-silico* generated datasets) (Figure 3B). The datasets were selected because they did not have any detectable amounts of *E. coli* by any of the three tools to confound results. MetaPhlAn, which is optimized for human-associated microbial communities, had better performance when tested against these HMP datasets relative to the *E. coli* or *B. anthracis* datasets mentioned above. However, MetaPhlAn still required at least 5% of the genome to be recovered in order to provide high confidence (positive) detection whereas imGLAD achieved similar confidence with only 3% of the genome. Hence, imGLAD's performance is superior, especially in cases of low abundance of the target genome(s).

Improved detection was also observed in *in-silico* synthesized datasets that included close relatives (ANI greater than 95% up to 98% compared to the target), although a larger fraction of the genome was typically required in these cases (~7%) in order to achieve high specificity and sensitivity by imGLAD. PathoScope and MetaPhlAn required an even higher fraction (at least 10%) of the target genome for comparable specificity and sensitivity (Figure 3C; Suppl Fig 3 shows similar results but the background metagenome was HMP instead of the *in-silico* synthesized datasets). In all cases high specificity (>97%) was achieved by imGLAD, which resulted in a low false positive rate (i.e., <3%). In comparison, the other three tools never reached higher than 90% specificity on the same four tests (Suppl Fig 4 & 5).

Filtering of conserved regions

In addition to creating a model using the whole genome, regions of the genome that provide a less reliable phylogenetic signal (e.g., regions that are highly conserved or contain mobile elements; see Material and Methods for details) can be identified by MyTaxa and removed/masked so that the prediction and/or the

training steps can be repeated with the filtered genome for more accurate results. Filtering in general, improved the detection limit because reads mapping on masked regions were not counted (Figure 4). For instance, filtering lowered the minimum sequencing depth required for robust detection from 0.123 (no filtering applied) to 0.061 (p-value < 0.05) in the training datasets for *E. coli*. The reduction in sequencing breadth however was not as dramatic as sequencing depth (e.g., 0.014 to 0.009 for the same datasets). The larger effect of filtering on sequencing depth than breadth was presumably attributable to the fact that filtering typically removed only a small part of the target genome (i.e., <5% by default settings) that recruited a disproportionally high number of reads encoding highly conserved or frequently transferred genes. This interpretation is also consistent with the sigmoidal relationship between sequencing depth and breadth, which tends to flatten at high values of sequencing depth and becomes linear at lower values (18). Hence, filtering with MyTaxa is recommended, in general.

Effect of relatedness of co-occurring genomes and strain-level resolution

When building the training set, the user is able to add any non-target genomes that could be relevant for optimizing detection of the target genome such as genomes that are known to be present and relatively abundant in the sample or closely related species that should not contribute positive signal (i.e., reads mapping on shared regions of the genome). In general, imGLAD's sequencing breadth and/or depth for positive detection (i.e., the detection limit) was expected to be higher with higher relatedness of the non-target genomes in the training set to the target genome. For instance, we tested three different training sets that included relatives at different levels of ANI to the target genome, i.e., 95-98% (within species resolution), 90-95% (resolution between closely related species), and 80-90% (resolution between moderately related species). Consistent with our expectations, higher sequencing depth and breadth were required for robust detection when relatives showing 95-98% ANI to the target co-occurred in the training dataset compared to relatives showing 90% or 80% ANI. This is due to the fact that more conserved and/or identical regions are present in the genome of the former relative to the latter. In fact, when co-occurring relatives were members of different species than the target species (i.e., show <95% ANI), imGLAD's limit of detection was very similar to that of the training datasets without close relatives, i.e., 3% of the genome needed to be recovered for confident detection in most cases. When genomes of the same species were present (i.e., show 95-98% ANI), at least 10-12% of the genome was required, depending on the exact genomes considered and their relatedness (Figure 5). Furthermore, as the relatedness increased, the contribution of the sequencing depth metric became less important. In fact, in some training datasets where close relatives with high identity to the target genome (95% ANI or higher) were present and in relative high abundance, the estimated parameters showed high variation during the training step and resulted, for instance, in a positive slope between sequencing breadth and depth, which was not reliable for estimating relative abundance and detection limit (Figure 5, black dotted and dashed line). In such cases, imGLAD can be customized to perform a cross-validation analysis in order to derive more robust parameters. Under this configuration, imGLAD generates a 10-fold larger training dataset, which is subsequently divided into ten subsets. Each subset is used independently to fit the model parameters.

The mean value for each parameter is taken as the consensus value and used to establish the relative abundance and detection limit (Figure 5, grey dotted and dashed line).

In summary, gene-content differences among the target genome and the co-occurring, non-target close relatives become increasingly more important for robust detection in cases where the non-target genome(s) show increasing genetic relatedness to the target. Strain-level resolution was achievable by imGLAD in such cases, and the resolving power decreased with smaller gene-content differences between target and non-target genomes. Hence, training with close relatives, and possibly cross-validation analysis, are important for more stringent results, especially in cases that the close relatives are known or highly anticipated to co-occur in the same samples with the target organism.

DISCUSSION

We presented imGLAD, a novel algorithm that utilizes a logistic model-based learning approach for accurate detection of target bacterial species in complex metagenomes, and for establishing detection limits in a target species- and microbial community-specific manner. By building and analysing training datasets with decreasing abundances of spiked-in reads originating from the target genome, imGLAD allows for highly reliable calls, while reducing the number of false positives (Figure 2). Further, and contrary to available tools, imGLAD allows for reliable estimation of the detection limit of the metagenomic sequencing effort applied based on the training datasets of decreasing target genome abundance and a linear combination of both genome sequencing depth and genome sequencing breadth, or only sequencing breadth. The degree of sequence conservation of the genes of the target genome and their extent of horizontal gene transfer are also taken into account in estimating the limit of detection, which represents a substantial advantage over existing tools in minimizing false-positive calls. The results using both simulated datasets (e.g., Figure 2) as well as experimental metagenomes (Tables 1 and 2) highlighted these advantages of imGLAD. However imGLAD is not designed to detect all species present in a sample. Thus, it differs from taxonomic profiling software, and is computationally more expensive, due to the training step, if the goal is to detect more than a couple of targets. Rather, the goal of imGLAD is to provide highly accurate detection of specific, user-provided target species (e.g., pathogens), including newly sequenced genomes. Further, imGLAD's logistic model, while computationally demanding to create (e.g., building *in-silico* training datasets), need to be built only once and can subsequently be used multiple times, such as with different metagenomes. This way, imGLAD could be used to detect several target organisms in an environmental sample (by building a model for each target).

A distinguishing strength of imGLAD is the detection of low abundance target genomes. Current tools for metagenomic profiling use specific markers or SNP patterns to identify and classify the species present in the sample [e.g., (4)]. However, at low levels of abundance, these markers may not be found, and SNPs cannot be called, and in some cases, the SNPs are called incorrectly such as in the case of MetaMLST (6), which requires high abundances (above 2X) to make confident calls and thus, performed poorly in the tests we conducted compared to other tools or imGLAD (e.g., Fig 3). Our approach is not focused on a particular region of the genome, but instead takes into account the whole genomic context. This provides

higher recall while preserving precision (Figure 3 & Suppl Fig 4). Further, methods based on read assignment depend on the comprehensiveness of their reference database and do not provide high precision when challenged with samples containing closely-related species (3). Accordingly, the tools evaluated here provided high false positive rates in such cases (Suppl Fig 4 & 5), which can be concerning, for instance, in pathogen surveillance studies and environmental samples, where closely related strains of the same species may co-occur. imGLAD can provide reliable prediction even in such cases, although at the expense of a lower detection limit, assuming the close relatives are known and available and, hence, can be used as part of the training step as exemplified in the *E. coli* case above. However, if the query metagenome(s) include relatively abundant, non-target genomes more related to the target genome than any of the genomes used to construct the training datasets, then the predictions of imGLAD (or other tools) might not be highly accurate. In such cases, the user needs to recover the genome sequences of the relatives from the metagenome(s) using genome binning techniques, if the representative sequences are not available otherwise, in order to include them in the training dataset. The results presented here (e.g., Figure 3A & Figure 5) provide a quantitative picture of this issue and its consequences on the accuracy of imGLAD as well as other tools.

Electrochemical immunoassays have shown promise in detecting pathogens such as *B. anthracis* or their toxins, and can sometimes offer strain-level resolution. The limit of detection of these techniques can, in some cases, be ~1pg/ml (19), which is below the limit of detection of imGLAD (56 pg/ml-560 pg/ml corresponding to 10-100 cells, respectively) based on the *E. coli* spike in experiment on spinach and current best practices for metagenomic sequencing and the samples analysed here. Thus, immunoassays and culture-based approaches are still more sensitive than metagenomics, at least for highly complex metagenomes such as those of soils (but probably not as much for food and agricultural samples, the human gut or habitats of similar complexity), and could be used in combination with tools like imGLAD for more reliable and comprehensive results. A key advantage of imGLAD is that it has high specificity, which sometimes cannot be achieved by immunoassays or culture-based approaches. It should be noted that imGLAD might be able to offer resolution within species as well, e.g., by including in the training dataset genomes that are members of the same species but show sequence divergence from the target genome higher than that of the sequencing errors (e.g., 99% ANI or less, assuming a sequencing error of <1%) and/or have substantial gene content differences (which can be captured by the sequencing breadth parameter). Sub-species resolution can also be obtained by analysing the reads identified by imGLAD as representing the target genome in the query metagenome for their SNP pattern against a collection of genomes related to the target genome, using -for instance- the PathoScope approach (3).

A key advantage of imGLAD is that the training step is easily customizable; thus, the algorithm can be optimized to evaluate samples of different microbial community complexity and co-occurring relatives of varied genetic similarity to the target organism as well as different target genomes. For instance, more complex communities can be simulated in the training step by including a higher number of different genomes in the training datasets (200 genomes by default) and/or with different species abundance distributions, e.g., power law as opposed to equal abundances (default setting). We have also found that

training datasets with 200 genomes work well for most natural communities of medium-to-high complexity while increasing the number of genomes only marginally increased the specificity or sensitivity of imGLAD (Suppl Fig 6 & 7), in general, especially given the extra computational time required. Specifically, our assessment showed that if the richness of the targeted microbial community (i.e., number of 95% ANI defined species or clusters) is within one order of magnitude of the number of genomes used in the training (i.e., 1 through 2000 species, for 200 genomes in the training datasets), the estimated imGLAD models are robust. Hence, the default number of genomes (n=200) should work for most microbial communities, and smaller number of genomes could be used for less complex communities (e.g., n=100). Moreover, analysing the target metagenome with profiling tools such as MethaPlan and MyTaxa (5,12) in advance can provide the end user with pertinent information on the taxonomic composition of the target community. This information can guide the selection of the genomes used for the training datasets so that close relatives, when present in the metagenome, can be included for more robust results (e.g., Figure. 5). Further, imGLAD allows one to include new target genomes, including draft assemblies, in the training datasets, with little effort, which may be important for practical applications. Thus, the training step of imGLAD can be optimized with specific microbial communities or habitats in mind such as the human gut and provide comparable, if not better results than tools that are already optimized for these communities. In contrast, most tools available require time and CPU-intensive updates of their reference databases to include new targets. Similarly, imGLAD can be easily optimized for different sequencing technologies as long as the training datasets are produced with reads simulating these technologies. This flexibility of imGLAD is an important advantage because the tenet "one approach fits all" does not apply well in the case of microbial detection in environmental samples, which are typically characterized by different degrees of microbial community complexity and co-occurring (non-target) relatives and are often sequenced based on different strategies nowadays.

CONCLUSIONS

The decreasing costs of sequencing as well as technological improvements in sequencing throughput and read length make it possible to use metagenomics to track specific bacterial populations in time series data or monitor the presence of pathogens in clinical or environmental samples. As studies with a focus on metagenomic datasets continue to increase, the need for fast, reliable and flexible bioinformatics analysis tools to detect and characterize target populations will also continue to grow, particularly in cases where isolation is not possible or is expensive. imGLAD represents an effective way to accomplish this objective and to robustly evaluate the limitations of the underlying sequencing technology or effort. imGLAD's default settings should work for most target microbial communities and genomes, and the results presented here represent a guide for further optimization depending on the specific goals of the study and the samples analysed. Therefore, we anticipate that imGLAD will find applications across the fields of clinical and environmental microbiology.

AVAILABILITY

imGLAD is open source software available in the GitHub repository (https://github.com/jccastrog/imGLAD).

FUNDING

This work was supported by the U.S.D.A. [award 2030-42000-046-10] and the US National Science Foundation [award 1356288].

CONFLICT OF INTEREST

Conflict of interest statement. None declared.

FIGURE LEGENDS

Figure 1. Schematic representation of imGLAD's pipeline. imGLAD has two main components. The first part (training) consists of a learning procedure, in which a set of *in-silico* generated datasets are fitted through a logistic model that aims to separate positive from negative datasets. For this, a database of 200 genomes is used to generate the simulated Illumina reads of these datasets. Reads simulated from the target genome are then incorporated into half of the simulated datasets. The resulting datasets are marked as positive for training while the other half is marked as negative. Sequencing depth and breadth of the target (reference) genome are calculated for each dataset. A logistic function is then fitted to the data to separate positive from negative examples. The regression parameters are stored for further use. The second part (estimation) consists of estimating the sequencing breadth and/or depth values of the target genome provided by the (recruited) reads of the experimental metagenomes, and comparison of the derived sequencing depth and breadth values to those of the logistic function from the training step.

Figure 2. Identification of target genomes in metagenomic datasets with imGLAD. Positive datasets (crosses) are separated form negative datasets (dots) through a logistic function (solid line) based on *in-silico* training datasets. (**A**) Datasets with reads of *E. coli* are separated form negative datasets. (**B**) Datasets with reads of *B. anthracis* are separated form negative datasets. Red asterisks denote the position of the experimental metagenomes (remaining dots represent *in-silico* generated datasets). Note the differences in scale on the x-axes between positive and negative datasets.

Figure 3. Performance of imGLAD in comparison to Pathoscope, MetaPhlAn and MetaMLST. (**A**) *in-silico* synthesized datasets from 200 RefSeq genomes were spiked with *E. coli* EC11 reads at different abundances (reflected by sequencing breadth, x-axes) to test the sensitivity of imGLAD (y-axes), i.e., the proportion of properly classified positive datasets among the total number of positive datasets. (**B**) Similar comparisons based on 571 datasets from the HMP project, which did not contain any *E. coli* signal and were spiked with different concentrations of *E. coli* EC11 reads. (**C**) imGLAD was evaluated in the same datasets as in panel A but this time the datasets included, in addition to the RefSeq genomes, 10 *E. coli* genomes with ANI ranging between 95-98% to the target *E. coli* EC11 genome spiked in at the same concentration (i.e., 0.3X). Note that as sequencing breadth increases the sensitivity of the prediction is higher for all tools tested, with the exception of MetaMLST that requires at least 2X sequencing depth for robust detection (see text for details). However, imGLAD can effectively classify samples at 100% sensitivity (as positive samples in this case) with a sequencing breadth as low as 0.03 (i.e., 3% of the target genome recovered) or less, whereas the other tools show lower sensitivity at these levels in all cases evaluated.

Figure 4. Effect of filtering of less informative genes by MyTaxa on minimum sequencing breadth and depth. Genome regions of the *E. coli* target genome were classified by MyTaxa, and regions with low scores (bottom 5%) or no scores because the corresponding genes were not indexed by MyTaxa were excluded from further analysis (filtered genome). Note that detection limit for the filtered genome (dashed line) is lower than the unfiltered genome (solid line).

Figure 5. Detection limits when co-occurring relatives are present. Negative and positive examples were constructed using default imGLAD settings except that closely related (non-target) genomes to the target genome at three levels of ANI, i.e., 95-98% (dotted and dashed line), 90-95% (dashed line), and 80-90% (solid line) were included in the datasets. 20 genomes of relatives were used at each ANI level, e.g., 20 *E. coli* genomes showing between 95 and 98% ANI to the target *E. coli* O157-H7 strain were used in the first set. Note that as relatedness increases, the fraction of the genome required for making a confident decision also increases since negative datasets (that include the relatives) have higher values of sequencing depth, which causes the decision line to have even a positive slope for the 95-98% ANI set and thus, the estimated detection limit to not be reliable. In the latter case, the detection limit should be calculated based on a cross-validation approach (grey lines; see text for details), which can provide more reliable estimates of model parameters and thus, relative abundance and detection limit. For relatives showing ANI lower than 95% identity to the target, cross-validation was not necessary (e.g., grey and black lines coincided).

REFERENCES

1. Be, N.A. *et al*. (2013). Detection of Bacillus Anthracis DNA in Complex Soil and Air Samples Using Next-Generation Sequencing. *PLoS ONE* **8**(9): 1-16.

2. Ahn, T.H. *et al*. (2015). Sigma: Strain-Level Inference of Genomes from Metagenomic Analysis for Biosurveillance. *Bioinformatics* **31**(2):170-77.

3. Hong, C. *et al*. (2014). PathoScope 2.0: A Complete Computational Frame-work for Strain Identification in Environmental or Clinical Sequencing Samples. *Microbiome* **2**(1): 33.

4. Segata, N. *et al*. (2012). Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes. *Nature methods* **9**(8): 811-14.

5. Truong, D.T. *et al*. (2015). MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling. *Nature Methods* **12**(10): 902-3.

6. Zolfo, Moreno *et al*. (2017). MetaMLST: Multi-Locus Strain-Level Bacterial Typing from Metagenomic Samples. *Nucleic Acids Research* **45**(2)

7. Altschul, S.F. *et al*. (1990). Basic Local Alignment Search Tool. *Journal of molecular biology* **215**(3): 403-10.

8. Caro-Quintero, A., and Konstantinidis K.T. (2012). Bacterial Species May Exist, Metagenomics Reveal. *Environmental Microbiology* **14**(2): 347-55.

9. Carter, M.Q. *et al*. (2011). Distinct Acid Resistance and Survival Fitness Displayed by Curli Variants of Enterohemorrhagic *Escherichia coli* O157:H7. *Applied and Environmental Microbiology* **77**(11): 3685-95.

10. Goris, J. *et al*. (2007). DNA-DNA Hybridization Values and Their Relationship to Whole-Genome Sequence Similarities. *International Journal of Systematic and Evolutionary Microbiology* **57**(1): 81-91.

11. Huang, W. *et al*. (2012). ART: A next-Generation Sequencing Read Simulator. *Bioinformatics* **28**(4): 593-94.

12. Kent, W. J. (2002). BLAT - The BLAST -Like Alignment Tool. *Genome research* **12**: 656-64.

13. Langmead, B., and Salzberg S.L. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nature Methods* **9**(4): 357-59.

14. Luo, C. et al. (2014). MyTaxa: An Advanced Taxonomic Classifier for Genomic and Metagenomic Sequences. *Nucleic Acids Research* **42**(8): 1-12.

15. Luo, C *et al*. 2015. ConStrains Identifies Microbial Strains in Metagenomic Datasets. *Nature Biotechnology* **33**(10): 1045–52.

16. Pruitt, K. D. *et al*. (2004). NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Research* **33**(suppl_1): D501–4.

17. Rodriguez-R, L. M, and Konstantinidis K.T. (2014). Estimating Coverage in Metagenomic Data Sets and Why It Matters. *The ISME journal* **8**(11): 2349-51.

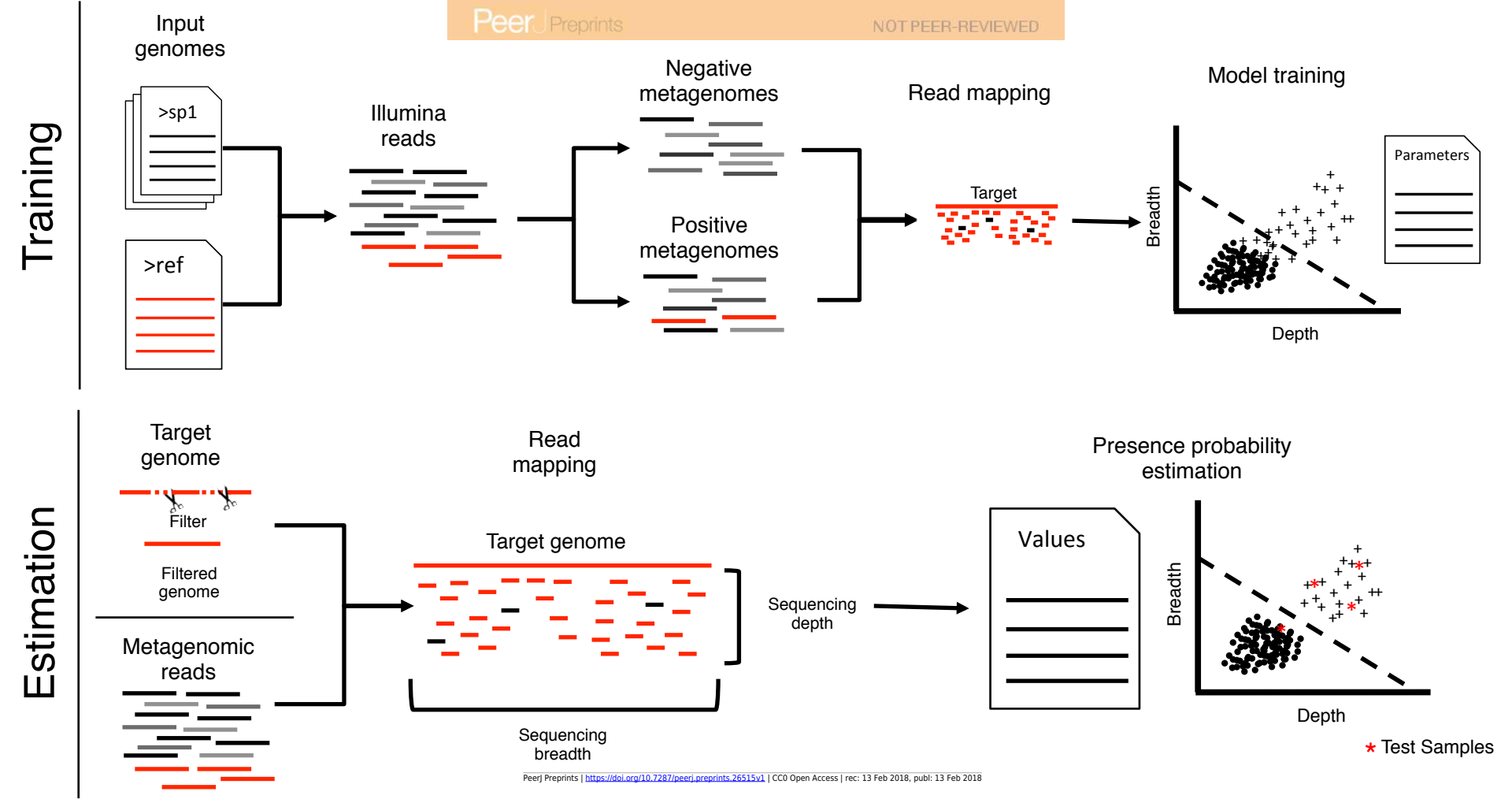
18. Sharma, M. K. *et al*. (2016). Ultrasensitive Electrochemical Immunoassay for Surface Array Protein, a Bacillus Anthracis Biomarker Using Au-Pd Nanocrystals Loaded on Boron-Nitride Nanosheets as Catalytic Labels. *Biosensors and Bioelectronics 80*: 442-49.

19. Wendl, Michael C. *et al*. (2013). Coverage Theories for Metagenomic DNA Sequencing Based on a Generalization of Stevens' Theorem. *Journal of Mathematical Biology* **67**(5): 1141-61.

Figure 1(on next page)

Schematic representation of imGLAD's pipeline.

imGLAD has two main components. The first part (training) consists of a learning procedure, in which a set of *in-silico* generated datasets are fitted through a logistic model that aims to separate positive from negative datasets. For this, a database of 200 genomes is used to generate the simulated Illumina reads of these datasets. Reads simulated from the target genome are then incorporated into half of the simulated datasets. The resulting datasets are marked as positive for training while the other half is marked as negative. Sequencing depth and breadth of the target (reference) genome are calculated for each dataset. A logistic function is then fitted to the data to separate positive from negative examples. The regression parameters are stored for further use. The second part (estimation) consists of estimating the sequencing breadth and/or depth values of the target genome provided by the (recruited) reads of the experimental metagenomes, and comparison of the derived sequencing depth and breadth values to those of the logistic function from the training step.



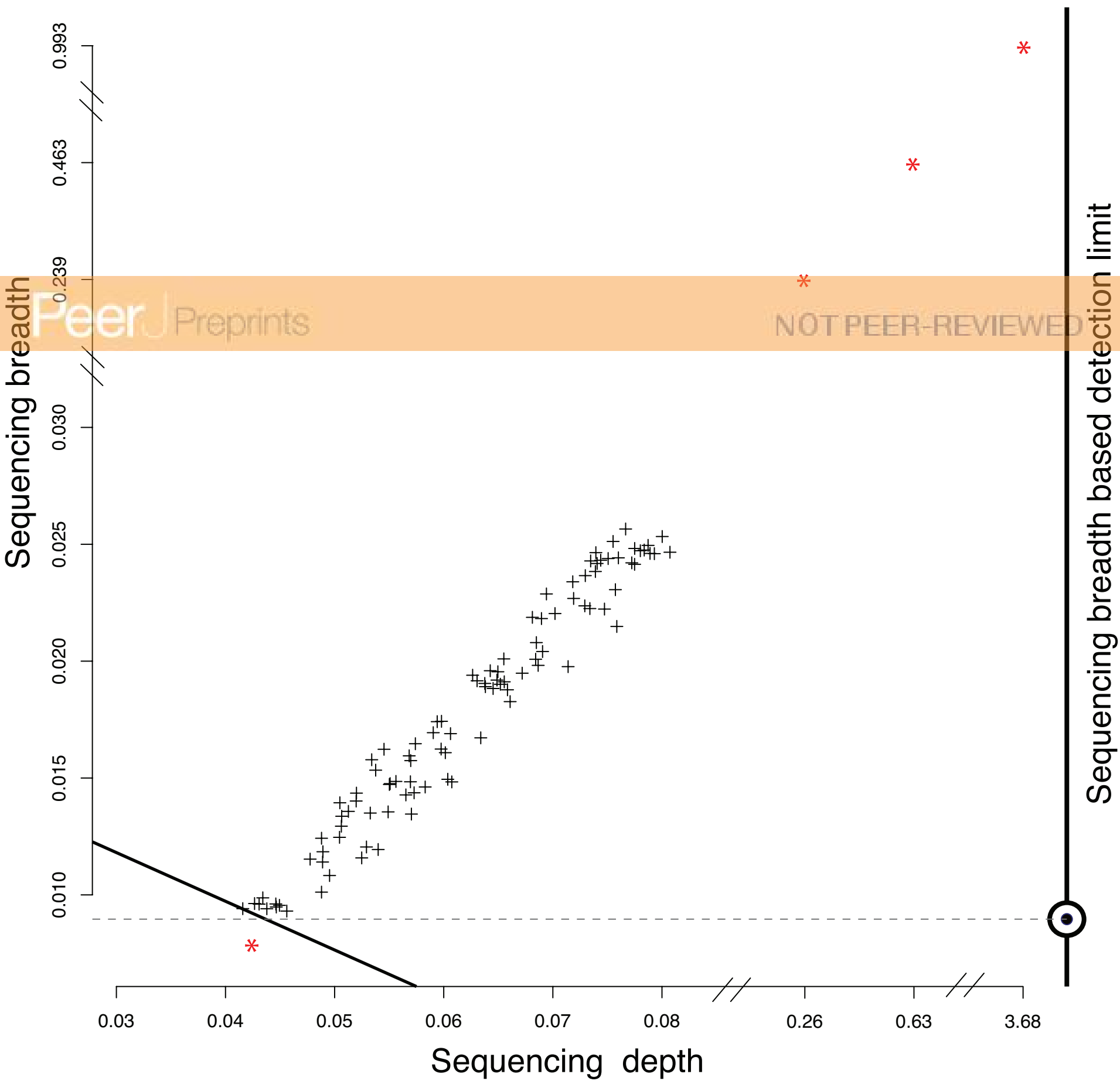
Peerj Preprints | https://doi.org/10.7287/peerj.preprints.26515v1 | CC0 Open Access | rec: 13 Feb 2018, publ: 13 Feb 2018

Figure 2(on next page)

Identification of target genomes in metagenomic datasets with imGLAD.

Positive datasets (crosses) are separated form negative datasets (dots) through a logistic function (solid line) based on *in-silico* training datasets. (**A**) Datasets with reads of *E. coli* are separated form negative datasets. (**B**) Datasets with reads of *B. anthracis* are separated form negative datasets. Red asterisks denote the position of the experimental metagenomes (remaining dots represent *in-silico* generated datasets). Note the differences in scale on the x-axes between positive and negative datasets.

A



B

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26515v1 | CC0 Open Access | rec: 13 Feb 2018, publ: 13 Feb 2018

Figure 3(on next page)

Performance of imGLAD in comparison to Pathoscope, MetaPhlAn and MetaMLST.

(**A**) *in-silico* synthesized datasets from 200 RefSeq genomes were spiked with *E. coli* EC11 reads at different abundances (reflected by sequencing breadth, x-axes) to test the sensitivity of imGLAD (y-axes), i.e., the proportion of properly classified positive datasets among the total number of positive datasets. (**B**) Similar comparisons based on 571 datasets from the HMP project, which did not contain any *E. coli* signal and were spiked with different concentrations of *E. coli* EC11 reads. (**C**) imGLAD was evaluated in the same datasets as in panel A but this time the datasets included, in addition to the RefSeq genomes, 10 *E. coli* genomes with ANI ranging between 95-98% to the target *E. coli* EC11 genome spiked in at the same concentration (i.e., 0.3X). Note that as sequencing breadth increases the sensitivity of the prediction is higher for all tools tested, with the exception of MetaMLST that requires at least 2X sequencing depth for robust detection (see text for details). However, imGLAD can effectively classify samples at 100% sensitivity (as positive samples in this case) with a sequencing breadth as low as 0.03 (i.e., 3% of the target genome recovered) or less, whereas the other tools show lower sensitivity at these levels in all cases evaluated.

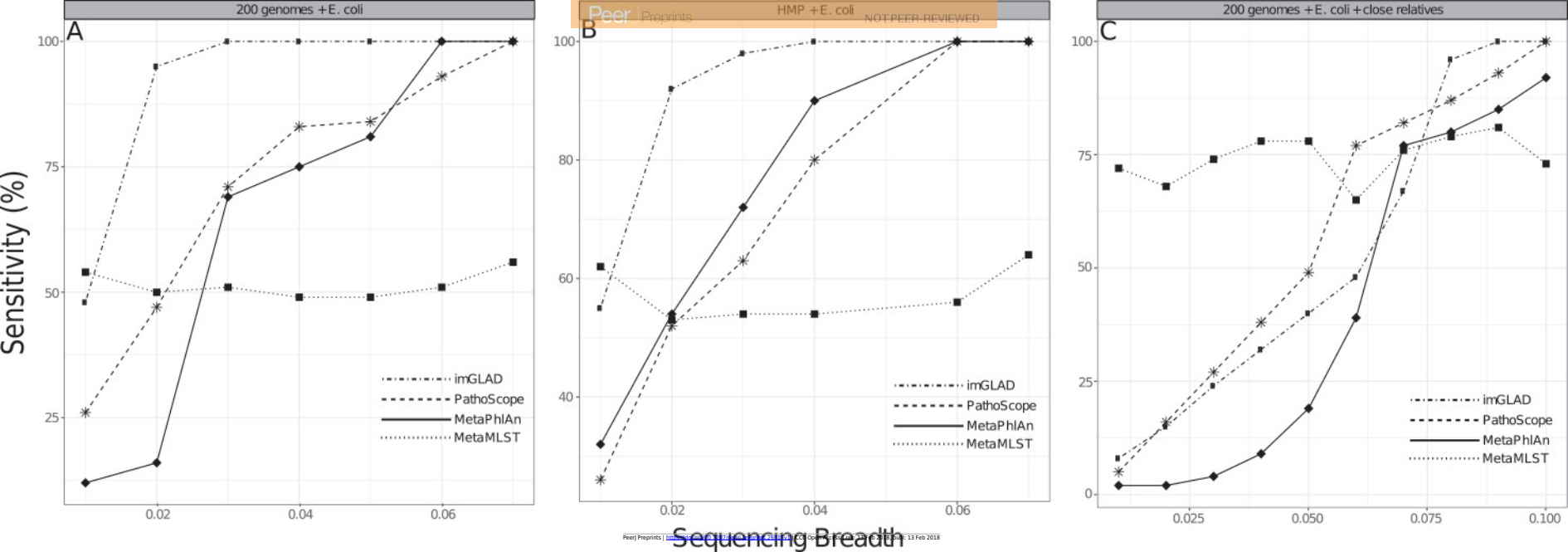


Figure 4(on next page)

Effect of filtering of less informative genes by MyTaxa on minimum sequencing breadth and depth.

Genome regions of the *E. coli* target genome were classified by MyTaxa, and regions with low scores (bottom 5%) or no scores because the corresponding genes were not indexed by MyTaxa were excluded from further analysis (filtered genome). Note that detection limit for the filtered genome (dashed line) is lower than the unfiltered genome (solid line).

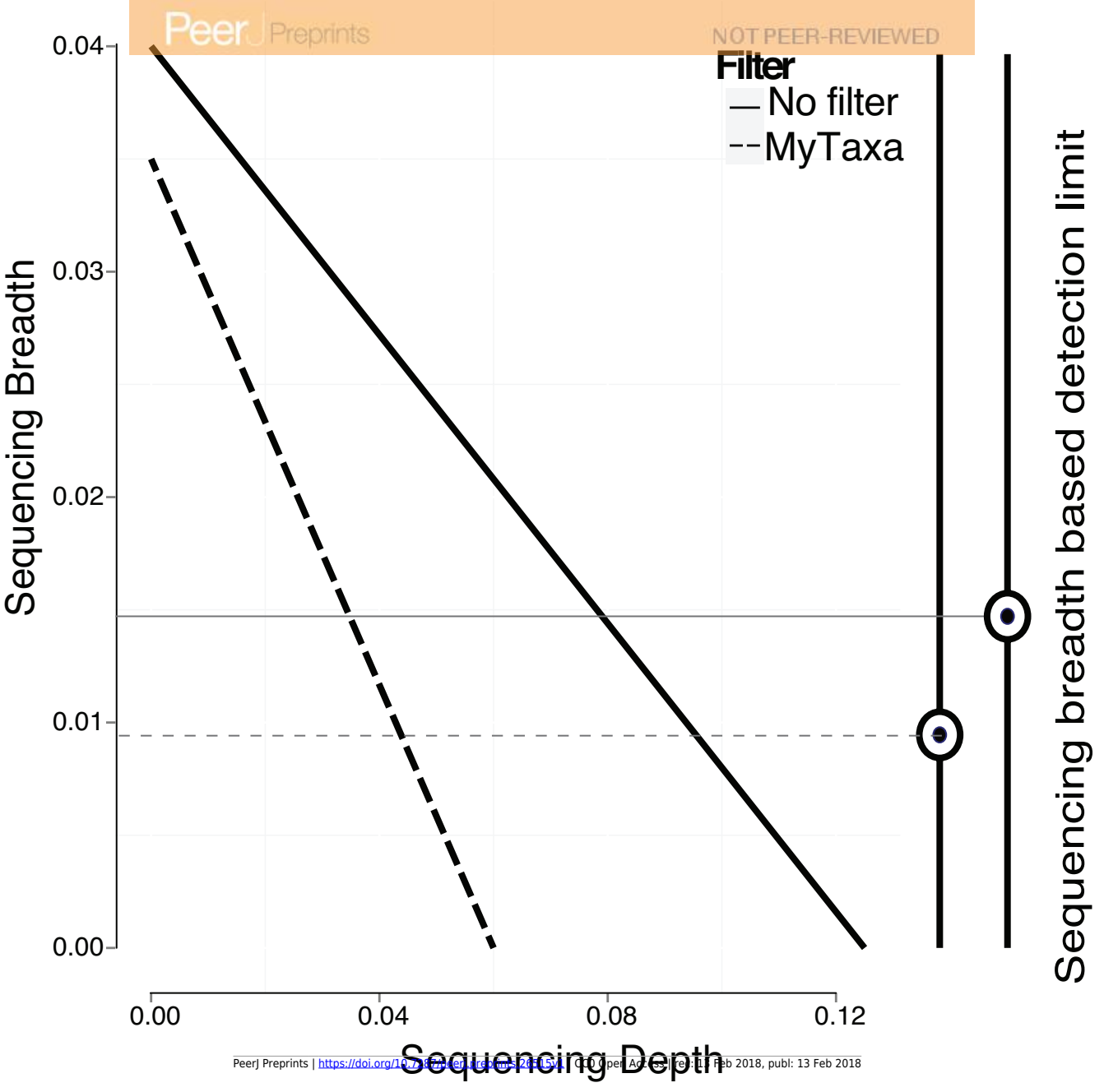
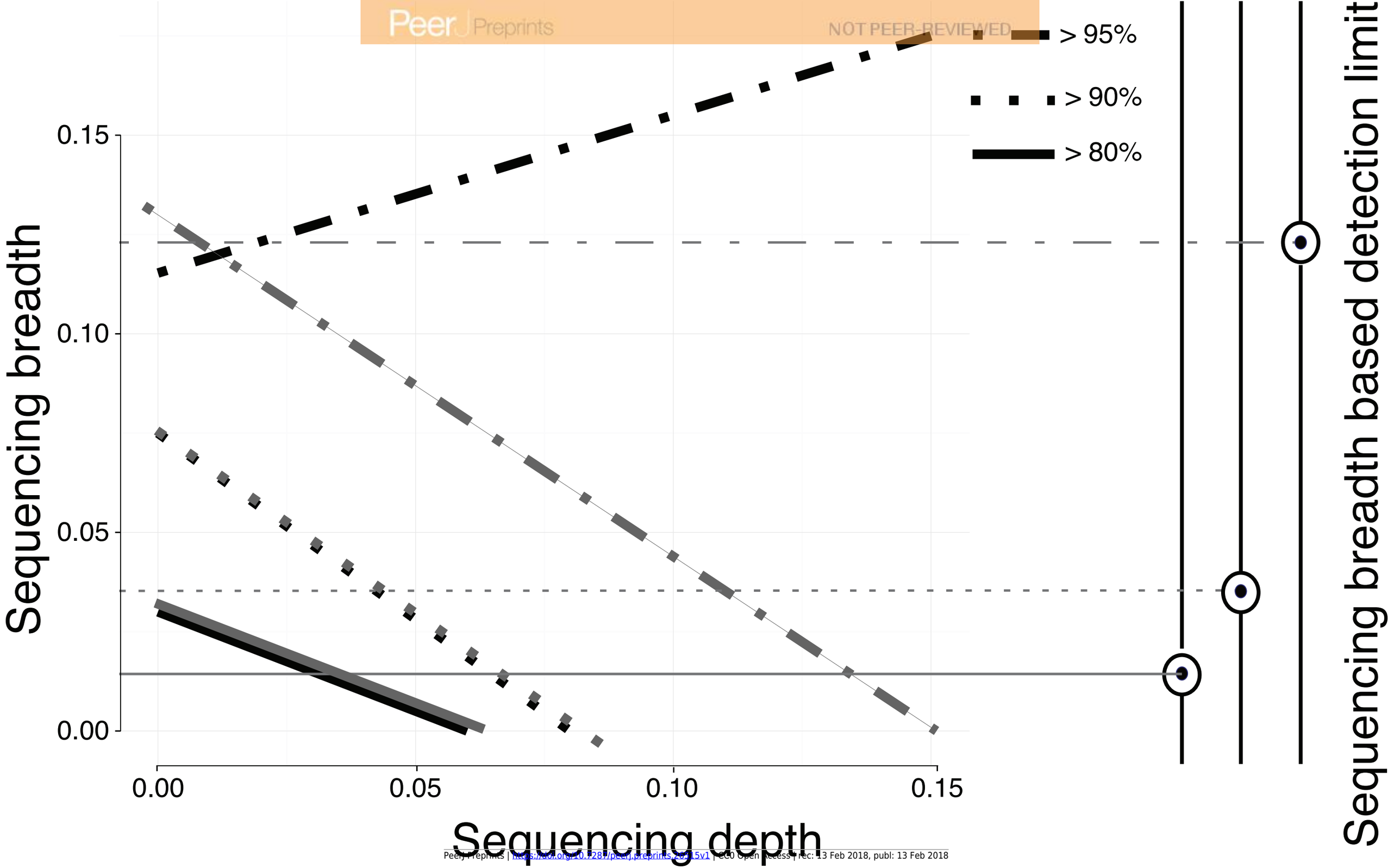


Figure 5(on next page)

Detection limits when co-occurring relatives are present.

Negative and positive examples were constructed using default imGLAD settings except that closely related (non-target) genomes to the target genome at three levels of ANI, i.e., 95-98% (dotted and dashed line), 90-95% (dashed line), and 80-90% (solid line) were included in the datasets. 20 genomes of relatives were used at each ANI level, e.g., 20 *E. coli* genomes showing between 95 and 98% ANI to the target *E. coli* O157-H7 strain were used in the first set. Note that as relatedness increases, the fraction of the genome required for making a confident decision also increases since negative datasets (that include the relatives) have higher values of sequencing depth, which causes the decision line to have even a positive slope for the 95-98% ANI set and thus, the estimated detection limit to not be reliable. In the latter case, the detection limit should be calculated based on a cross-validation approach (grey lines; see text for details), which can provide more reliable estimates of model parameters and thus, relative abundance and detection limit. For relatives showing ANI lower than 95% identity to the target, cross-validation was not necessary (e.g., grey and black lines coincided).



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2815v1 | CC0 Open Access | rec: 13 Feb 2018, publ: 13 Feb 2018

Table 1 (on next page)

Samples inoculated with different cell concentrations of *E. coli* (1st column) were classified by imGLAD as present/positive or absent/negative.

The calculated breadth of the *E. coli* reference genome recovered (2nd column) and the sequencing depth (3rd column) as well as the derived probability of presence (4th column) are shown. All samples were found positive for presence of *E. coli* (p-value = 0.004) except the control sample without inoculated *E. coli* cells.

Table 1. Samples inoculated with different cell concentrations of *E. coli* (1st column) were classified by imGLAD as present/positive or absent/negative.

Sample	Sequencing breadth	Sequencing depth	*E. coli* presence (p-value)
Control	0.007	0.042	0.847
80 Cells	0.239	0.262	0.004
8×10^3 Cells	0.463	0.639	7.1×10^{-4}
8×10^5 Cells	0.993	3.683	1×10^{-5}

The calculated breadth of the *E. coli* reference genome recovered (2nd column) and the sequencing depth (3rd column) as well as the derived probability of presence (4th column) are shown. All samples were found positive for presence of *E. coli* (p-value = 0.004) except the control sample without inoculated *E. coli* cells.

Table 2 (on next page)

Soil samples inoculated with different copies of *B. anthracis* strain Ames genomic DNA (1st column) were classified by imGLAD as present/positive or absent/negative.

The calculated breadth of the *B. anthracis* reference genome recovered (2nd column) and the sequencing depth (3rd column) as well as the derived probability of presence (4th column) are shown. Samples with a number of genome higher or equal to 100 genomes were classified as positive samples. Samples with 1 and 10 genomic copies were indistinguishable from the negative samples of the training set.

Table 2. Soil samples inoculated with different copies of *B. anthracis* strain Ames genomic DNA (1st column) were classified by imGLAD as present/positive or absent/negative.

Sample	Sequencing breadth	Sequencing depth	*B. anthracis* presence (p-value)
1 Genome	1.56×10^{-3}	2.0×10^{-3}	0.999
10 Genome	0.001	0.003	0.998
100 Genome	0.039	0.128	0.002
10^3 Genome	0.562	0.732	1.4×10^{-3}
10^4 Genome	0.983	1.563	0
10^5 Genome	0.999	4.34	0

The calculated breadth of the *B. anthracis* reference genome recovered (2nd column) and the sequencing depth (3rd column) as well as the derived probability of presence (4th column) are shown. Samples with a number of genome higher or equal to 100 genomes were classified as positive samples. Samples with 1 and 10 genomic copies were indistinguishable from the negative samples of the training set.