# Bracken: estimating species abundance in metagenomics data

Jennifer Lu[1,2], Florian P. Breitwieser[2], Peter Thielen[3] and Steven L. Salzberg[1,2,4]

[1] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, United States
[2] Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, United States
[3] Applied Physics Laboratory, Johns Hopkins University, Laurel, MD, United States
[4] Departments of Computer Science and Biostatistics, Johns Hopkins University, Baltimore, MD, United States

## ABSTRACT

Metagenomic experiments attempt to characterize microbial communities using high-throughput DNA sequencing. Identification of the microorganisms in a sample provides information about the genetic profile, population structure, and role of microorganisms within an environment. Until recently, most metagenomics studies focused on high-level characterization at the level of phyla, or alternatively sequenced the 16S ribosomal RNA gene that is present in bacterial species. As the cost of sequencing has fallen, though, metagenomics experiments have increasingly used unbiased shotgun sequencing to capture all the organisms in a sample. This approach requires a method for estimating abundance directly from the raw read data. Here we describe a fast, accurate new method that computes the abundance at the species level using the reads collected in a metagenomics experiment. Bracken (Bayesian Reestimation of Abundance after Classification with KrakEN) uses the taxonomic assignments made by Kraken, a very fast read-level classifier, along with information about the genomes themselves to estimate abundance at the species level, the genus level, or above. We demonstrate that Bracken can produce accurate species- and genus-level abundance estimates even when a sample contains multiple near-identical species.

**Subjects** Bioinformatics, Computational Biology
**Keywords** Metagenomics, Species abundance, Microbiome, Bayesian estimation

Submitted 7 July 2016
Accepted 28 November 2016
Published 2 January 2017

Corresponding author
Jennifer Lu, jlu26@jhmi.edu

Academic editor
James Procter

Additional Information and Declarations can be found on page 13

DOI **10.7717/peerj-cs.104**

## INTRODUCTION

Metagenomics is a rapidly growing field of study, driven in part by our ability to generate enormous amounts of DNA sequence rapidly and inexpensively. Since the human genome was first published in 2001 (*The International Human Genome Sequencing Consortium, 2001*; *Venter et al., 2001*), sequencing technology has become approximately one million times faster and cheaper, making it possible for individual labs to generate as much sequence data as the entire Human Genome Project in just a few days. In the context of metagenomics experiments, this makes it possible to sample a complex mixture of microbes by "shotgun" sequencing, which involves simply isolating DNA, preparing the DNA for sequencing, and sequencing the mixture as deeply as possible. Shotgun sequencing is relatively unbiased compared to targeted sequencing methods (*Venter et al., 2004*), including widely-used 16S ribosomal RNA sequencing, and it has the additional advantage that it captures any species

**How to cite this article** Lu et al. (2017), Bracken: estimating species abundance in metagenomics data. **PeerJ Comput. Sci. 3:e104; DOI 10.7717/peerj-cs.104**

with a DNA-based genome, including eukaryotes that lack a 16S rRNA gene. Because it is unbiased, shotgun sequencing can also be used to estimate the abundance of each taxon (species, genus, phylum, etc.) in the original sample, by counting the number of reads belonging to each taxon.

Along with the technological advances, the number of finished and draft genomes has also grown exponentially over the past decade. At present there are thousands of complete bacterial genomes, 20,000 draft bacterial genomes, and 80,000 full or partial virus genomes in the public GenBank archive (*Benson et al., 2015*). This rich resource of sequenced genomes now makes it possible to sequence uncultured, unprocessed microbial DNA from almost any environment, ranging from soil to the deep ocean to the human body, and use computational sequence comparisons to identify many of the formerly hidden species in these environments (*Riesenfeld, Schloss & Handelsman, 2004*). Several accurate methods have appeared that can align a sequence "read" to a database of microbial genomes rapidly and accurately (see below), but this step alone is not sufficient to estimate how much of a species is present. Complications arise when closely related species are present in the same sample–a situation that arises quite frequently–because many reads align equally well to more than one species. This requires a separate abundance estimation algorithm to resolve. In this paper, we describe a new method, Bracken, that goes beyond simply classifying individual reads and computes the abundance of species, genera, or other taxonomic categories from the DNA sequences collected in a metagenomics experiment.

When it was first published in 2014, the Kraken metagenomics classifier represented a major enhancement in the speed with which large metagenomics sequence data could be processed (*Wood & Salzberg, 2014*), running over 900 times faster than MegaBlast (*Morgulis et al., 2008*), the closest competitor at the time. Kraken's success and accuracy rely on its use of a very large, efficient index of short sequences of length $k$, which it builds into a specialized database. If $k$ is chosen appropriately, then most sequences of length $k$ in the database will be unique to a single species, and many will also be unique to a particular strain or genome. Larger values of $k$ will yield a database in which even more of each genome is uniquely covered by $k$-mers; obviously, though, $k$ should not be longer than the length of a sequencing read, and metagenomics projects currently generate reads as short as 75–100 base pairs (bp). Longer $k$-mers are also more likely to contain errors, meaning that more reads will be left unclassified if $k$ is too long. Smaller $k$-mers, in contrast, will yield higher sensitivity because the minimum match length is shorter.

When used to identify the taxonomic label of metagenomics sequences, the Kraken system for classification of metagenomics sequences is extremely fast and accurate (*Wood & Salzberg, 2014*). When classifying raw sequence reads, though, many reads correspond to identical regions between two or more genomes. (The number of such ambiguous reads decreases as reads get longer.) Kraken solves this problem by labeling the sequence with the lowest common ancestor (LCA) of all species that share that sequence, as discussed further below.

## Ambiguity among microbial species and strains

As the database of bacterial genomes has grown, an increasing number of genomes share large portions of their sequence with other genomes. In many cases, these genomes are nearly identical; indeed, sequencing has revealed to scientists that many formerly distinct species and genera are far closer than were known prior to sequencing. Many species have been renamed as a result, in a process that is continual and ongoing, but many other species have retained their old names, often for historical or other reasons.

For example, the species *Mycobacterium bovis* is over 99.95% identical to *Mycobacterium tuberculosis* (*Garnier et al., 2003*), and many cases of human tuberculosis are caused by *M. bovis* (which also infects cows) rather than *M. tuberculosis* (*Grange, 2001*). Their high sequence identity indicates that they should be considered as two strains of a single species, but they retain different species names. As a compromise, taxonomists have created the category *Mycobacterium tuberculosis complex* (*Brosch et al., 2002*) to represent a collection of taxa that now includes more than 100 strains of five different species. This category sits above the species level but below the genus level in the current microbial taxonomy, but it can best be described as a species.

Other examples are numerous and still growing. The three species *Bacillus anthracis* (the causative agent of anthrax), *Bacillus cereus*, and *Bacillus thuringiensis* are well over 99% identical and should all be designated as a single species (*Helgason et al., 2000*), although their names have not been changed despite their near-identity revealed by sequencing. As a compromise, taxonomists created the category *Bacillus cereus group*, between the level of species and genus, to include these three species and at least five others (*Liu et al., 2015*), all of which are extremely similar to one another. In some cases, two organisms that should be called the same species may even have different genus names. For example, *Escherichia coli* and *Shigella flexneri* are classified in different genera, but we know from sequence analysis that they represent the same species (*Lan & Reeves, 2002*).

Failure to recognize the mutability of the bacterial taxonomy can lead to erroneous conclusions about the performance of metagenomic classifiers. For example, one recent study (*Peabody et al., 2015*) created a mock community of 11 species, one of which was *Anabaena variabilis* ATCC 29413, not realizing that this genome had been renamed and was synonymous with species in the genus *Nostoc* (*Thiel et al., 2014*). When *Anabaena* was removed from the database, Kraken correctly identified the reads as *Nostoc*, but Peabody et al. erroneously considered all these reads to be misclassified.

## Classification versus abundance estimation

Kraken attempts to assign a taxonomy label to every read in a metagenomics sample using a custom-built database that may contain any species the user chooses. Among the current set of finished bacterial and archaeal genomes, hundreds of species can be found for which large fractions of their sequence are identical to other genomes belonging to distinct strains, species, or even genera. The reads arising from common regions in these species result in a tie when analyzed with Kraken's classification algorithm, so Kraken correctly reports only the lowest common ancestor (LCA) (*Wood & Salzberg, 2014*). It follows that for well-populated clades with low genome diversity, Kraken only reports species-level assignments for reads from unique regions, and a true indication of total abundance can only be made by taking both species and genus (or higher) level assignments into account. This implies that for some species, the majority of reads might be classified at a higher level of the taxonomy. Kraken thus leaves many reads "stranded" above the species level, meaning that the number of reads classified directly to a species may be far lower than the actual number present.

Therefore, any assumption that Kraken's raw read assignments can be directly translated into species- or strain-level abundance estimates (e.g., *Schaeffer et al., 2015*) is flawed, as ignoring reads at higher levels of the taxonomy will grossly underestimate some species, and creates the erroneous impression that Kraken's assignments themselves were incorrect.

Nonetheless, metagenomics analysis often involves estimating the abundance of the species in a particular sample. Although we cannot unambiguously assign each read to a species, we would like to estimate how much of each species is present, specifically by estimating the number or percentage of reads in the sample. Several software tools have been developed to estimate species abundances in metagenomics samples [MetaPhlAn, ConStrains, GAAS, GASiC, TAEC, GRAMMy] (*Angly et al., 2009*; *Lindner & Renard, 2012*; *Luo et al., 2015*; *Segata et al., 2012*; *Sohn et al., 2014*; *Xia et al., 2011*). These tools, however, employ different strategies for read-level classification which are not always as accurate and efficient as Kraken's *k*-mer approach (*Lindgreen, Adair & Gardner, 2016*). Rather than re-engineer Kraken to address the ambiguous read classification issue and to provide abundance estimates directly, we decided to implement the new species-level abundance estimation method described here as a separate program. This preserves both backwards compatibility for existing Kraken users, and offers the ability to generate more accurate species abundance estimates for datasets already processed by Kraken. Note that if Kraken fails to identify a species (e.g., if the species was missing from the Kraken database), Bracken too will not identify that species.

## MATERIALS AND METHODS

Our new method, Bracken (Bayesian Reestimation of Abundance after Classification with KrakEN), estimates species abundances in metagenomics samples by probabilistically re-distributing reads in the taxonomic tree. Reads assigned to nodes above the species level are distributed down to the species nodes, while reads assigned at the strain level are re-distributed upward to their parent species. For example, in Fig. 1 we would distribute reads assigned to the *Mycobacteriaceae* family and the *Mycobacterium* genus down to *M. marinum* and *M. avium*, and reads assigned to each *M. marinum* strain would be reassigned to the *M. marinum* species. As we show below, Bracken can easily reestimate abundances at other taxonomic levels (e.g., genus or phylum) using the same algorithm.

In order to re-assign reads classified at higher-level nodes in the taxonomy, we need to compute a probabilistic estimate of the number of reads that should be distributed to the species below that node. To illustrate using the nodes in Fig. 1, we need to allocate all reads assigned to *Mycobacterium* (G1) to *M. marinum* (S1) and *M. avium* (S2) below

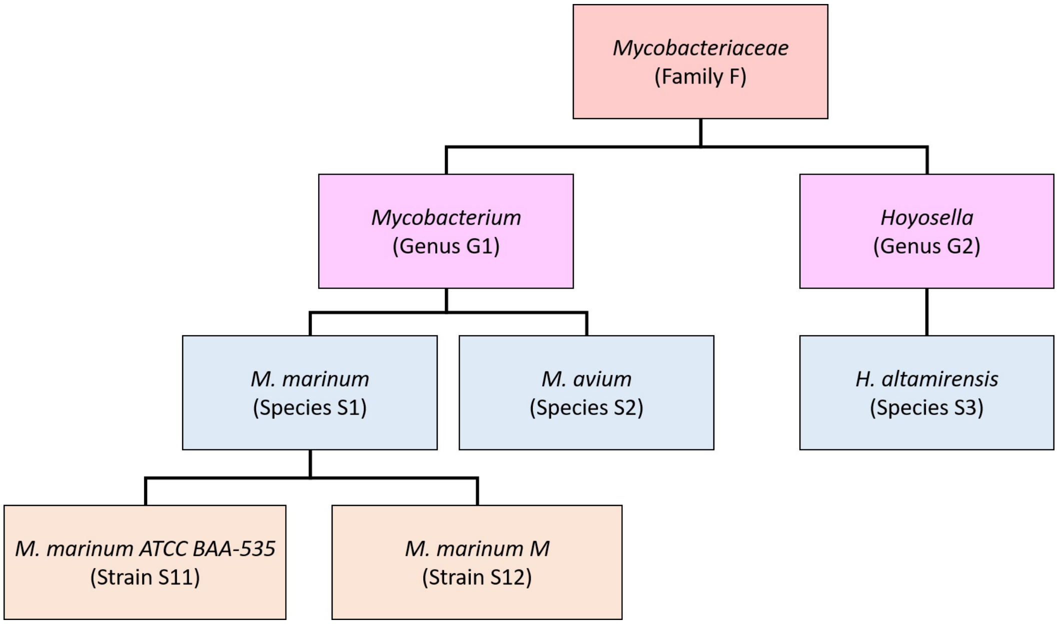

**Figure 1** **Schematic showing a partial taxonomic tree for the *Mycobacteriaceae* family.**

it, and reads assigned to the *Mycobacteriaceae* would have to be allocated to *M. marinum* (S1), *M. avium* (S2), and*Hoyosella altamirensis* (S3).

Reallocating reads from a genus-level node in the taxonomy to each genome below it can be accomplished using Bayes' theorem, if the appropriate probabilities can be computed. Let $P(S_i)$ be the probability that a read in the sample belongs to genome $S_i$, $P(G_j)$ be the probability that a read is classified by Kraken at the genus level $G_j$, and $P(G_j|S_i)$ be the probability that a read from genome $S_i$ is classified by Kraken as the parent genus $G_j$. Then the probability that a read classified at genus $G_j$ belongs to the genome $S_i$ can be expressed as Eq. (1):

$$P(S_i|G_j) = \frac{P(G_j|S_i)P(S_i)}{P(G_j)}. \tag{1}$$

Note that because we began by assuming that a read was classified at node $G_j$, $P(G_j) = 1$.

Next we consider how to compute $P(G_j|S_i)$, the probability that a read from genome $S_i$ will be classified by Kraken at the parent genus $G_j$. We estimate this probability for reads of length $r$ by classifying the sequences (genomes) that we used to build the database using that same database, as follows. For each $k$-mer in the sequences, Kraken assigns it a taxonomy ID by a fast lookup in its database. To assign a taxonomy ID for a read of length $r$, Kraken examines all $k$-mer classifications in that read. For example, for $k = 31$ and $r = 75$, the read will contain 45 $k$-mers. Our procedure examines, for each genome in the database, a sliding window of length $r$ across the entire genome.

To find the taxonomy ID Kraken would assign to each window, we simply find the deepest taxonomy node in the set of $k$-mers in that window. Since each $k$-mer in a database sequence is assigned to a taxonomy ID somewhere along the path from the genome's taxonomy ID to the root, the highest-weighted root-to-leaf path (and thus the Kraken classification) corresponds to the deepest node.

For each genome $S_i$ of length $L_i$ we thus generate $(L_i - r + 1)$ mappings to taxonomical IDs. For node $G_j$, we then count the number of reads from $\mathrm{S}_i$ that are assigned to it, $N_{G_j(Si)}$. $P(G_j|S_i)$ is then the proportion of reads from $S_i$ that were assigned to the genus node $G_j$; i.e., $P(G_j|S_i) = N_{Gj(Si)}/(L_i - r + 1)$. We also calculate the proportion of reads from $S_i$ that were assigned to every node from genome $S_i$ to the root node of the taxonomy tree.

The final term that we must calculate from Eq. (1) is $P(S_i)$, the probability that a read in the sample belongs to genome $S_i$, which is computed in relation to other genomes from the same genus. For example, if the sample contains three genomes in the same genus, and if 30% of all reads from those three genomes belong to $S_i$, then $P(S_i) = 0.3$. We estimate this probability using the reads that are uniquely assigned by Kraken to genome $S_i$, as follows.

If we let $U_{Si}$ be the proportion of genome $S_i$ that is unique, then

$$U_{S_i} = \frac{N_{S_i}}{L_i - r + 1} \tag{2}$$

where $N_{Si}$ is the number of $k$-mers of length $r$ that are uniquely assigned to genome $S_i$ by Kraken, and $L_i$ is the genome length. For example, if $L_i = 1$ Mbp and only 250,000 $k$-mers are unique to genome $S_i$, then $U_{Si} = 0.25$.

Then, using the number of reads $K_{Si}$ from a sample that Kraken actually assigns to $S_i$, we can estimate the number of reads that likely derive from $S_i$ as:

$$\hat{K}_{S_i} = \frac{K_{Si}}{U_{Si}}. \tag{3}$$

For example, if Kraken classifies 1,000 reads as genome $S_i$ and 25% of the reads from $S_i$ are unique, then we would estimate that 4,000 reads (1,000/0.25) from $S_i$ are contained in the sample.

If genus $G_j$ contains $n$ genomes, we estimate the number of reads $\hat{K}_S$ for each of the $n$ genomes and then calculate $P(S_i)$ by:

$$P(S_i) = \frac{\hat{K}_{S_i}}{\sum_{a=1}^{n} \hat{K}_{S_a}}. \tag{4}$$

Using this result in Eq. (1) above allows us to compute $P(S_i|G_j)$ for each genome $S_i$. Each probability $P(S_i|G_j)$ is then used to estimate the proportion of the reads assigned to genus $G_j$ that belong to each of the genomes below it.

These calculations are repeated for each taxonomic level above the genus level (family, class, etc.), with read distribution at each level going to all genomes classified within that taxonomic subtree.

To compute species abundance, any genome-level (strain-level) reads are simply added together at the species level. In cases where only one genome from a given species is detected by Kraken in the dataset, we simply add the reads distributed downward from the genus level (and above) to the reads already assigned by Kraken to the species level. In cases where multiple genomes exist for a given species, the reads distributed to each genome are combined and added to the Kraken-assigned species level reads. The added reads give the final species-level abundance estimates.

This method can also estimate abundance for other taxonomic levels. In such cases, only higher nodes within the taxonomy tree undergo read distribution. After distributing reads downward, we estimate abundance for a node at the level specified by combining the distributed reads across all genomes within that node's subtree.

### Software and data availability

Bracken is written in Perl and Python and is freely available for download at http://ccb.jhu.edu/software/bracken/. The reads from the skin microbiome experiment are freely available from NCBI under BioProject PRJNA316735.

## RESULTS AND DISCUSSION

We applied the statistical re-assignment method described here to create species-level abundance estimates for several metagenomics data sets. The overall procedure works as follows. First, we compute a set of probabilities from the Kraken database by computing, for every sequence of length R in every genome, where it will be assigned in the taxonomy (see 'Methods'). For our experiments, we set $R = 75$ as our datasets contain 75-bp reads. Bracken can use these probabilities for any metagenomics data set, including data with different read lengths, although the estimates might be slightly improved by re-computing with a read length that matches the experimental data.

Second, we run Kraken on the dataset to produce read-level taxonomic classifications. We then apply our abundance estimator, Bracken, which uses the numbers of reads assigned by Kraken at every level of the taxonomy to estimate the abundances at a single level (e.g., species). Note that to exclude false positives, Bracken ignores species with counts below a user-adjustable threshold. In our experiments, we selected a threshold of 10 reads.

### Experiments on a 100-genome metagenomics data set

For our first experiments, we used a data set containing simulated Illumina reads from 100 genomes. This data, which we call here the i100 dataset, was used previously in a comparison of metagenomic assembly algorithms (*Mende et al., 2012*). The data contains 53.3 million paired reads (26.7M pairs) from 100 genomes representing 85 species. The reads have error profiles based on quality values found in real Illumina reads (*Mende et al., 2012*). The i100 dataset includes several very challenging genomes for this task, including multiple strains and species in the genera *Bacillus* and *Mycobacteria*, some of which are nearly identical to one another. The i100 data are freely available at http://www.bork.embl.de/~mende/simulated_data.

The difficulty of estimating species abundance increases as the database itself contains more species. For example, it would clearly be easier to estimate abundances in the i100 dataset if we used a Kraken database containing only the 100 genomes in that dataset. To make the problem more realistic, we built two different databases and estimated abundance using both. The first ("small") database contains 693 genomes including the i100 genomes;

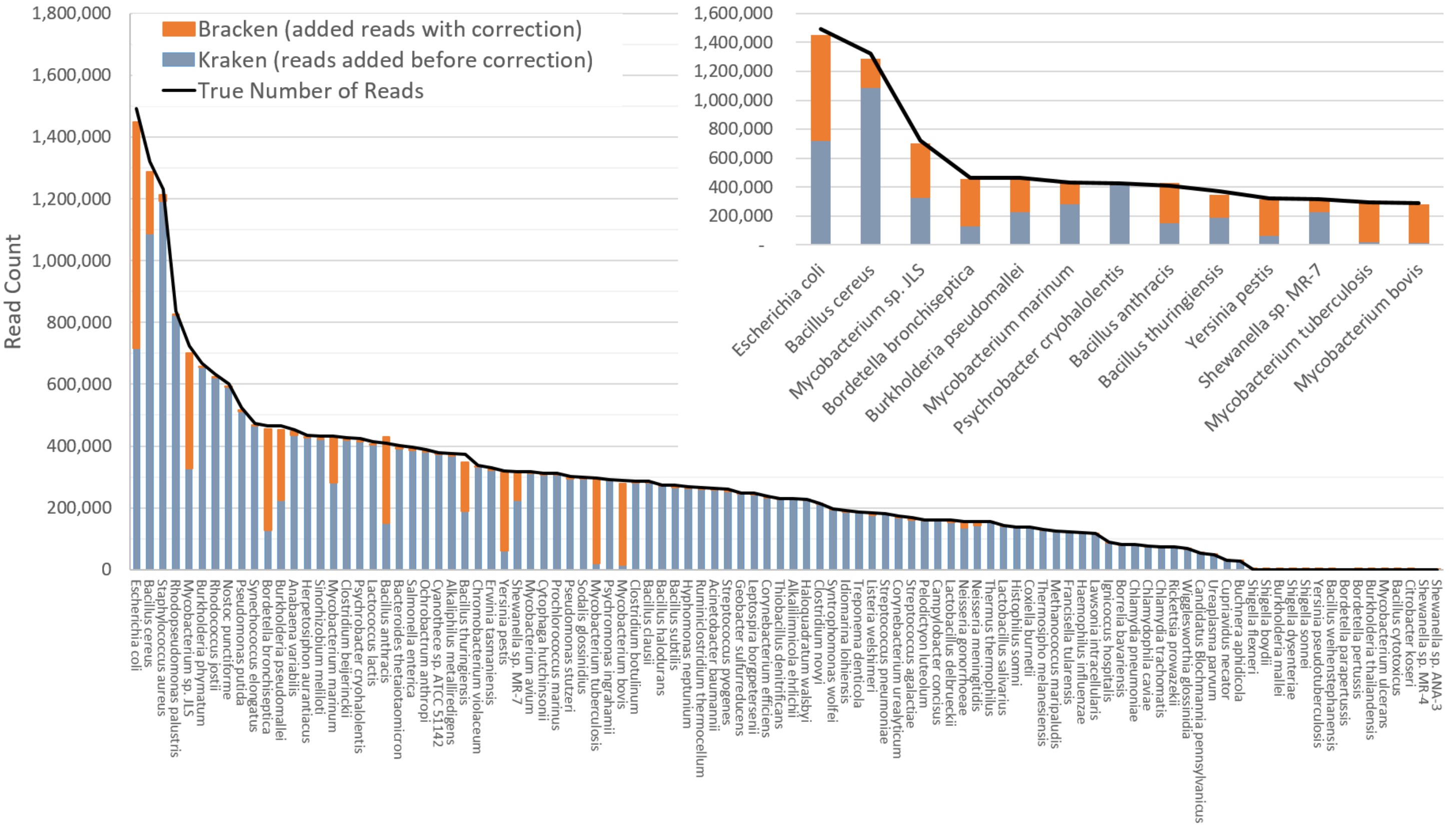

**Figure 2 Estimates of species abundance in the i100 metagenomics dataset computed by Kraken (blue) and Bracken (blue + orange).** For this result, the Kraken database contained 693 genomes that included the i100 genomes. The smaller graph displays results for the subset of species for which Bracken made the largest adjustments. The black line shows the true number of reads from each species. Precise numbers for the Kraken classification, true read counts, and Bracken estimates are contained in Table S2A.

this is the full database from the simulation study by *Mende et al. (2012)*. The results when using the small database for classification are shown in Fig. 2. For several species, the initial Kraken numbers (reads assigned to a particular species) are far too low, because many of the reads (for some genomes, a large majority) were assigned labels at the genus level or above. After reestimation with Bracken, these reads were redistributed to the species level, with the result that almost all the abundance estimates were 98–99% correct, as shown in the figure.

The second (“large”) database contains all genomes used in the synthetic and spike-in experiments, as well as a broad background of bacterial genomes. In particular, it includes all complete bacterial and archaeal genomes from RefSeq as of 25 July 2014 (archived at ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq), which total 2596 distinct taxa, plus those i100 genomes that were not present in the RefSeq data. (We excluded draft genomes because they often contain vector sequences or other contaminants.) We also added the nine genomes used in our skin bacteria spike-in experiment (described below) resulting in a total of 2635 distinct taxa. The complete list of sequences in the large database can be found in Table S1. The resulting Kraken database has a size of 74 GB.

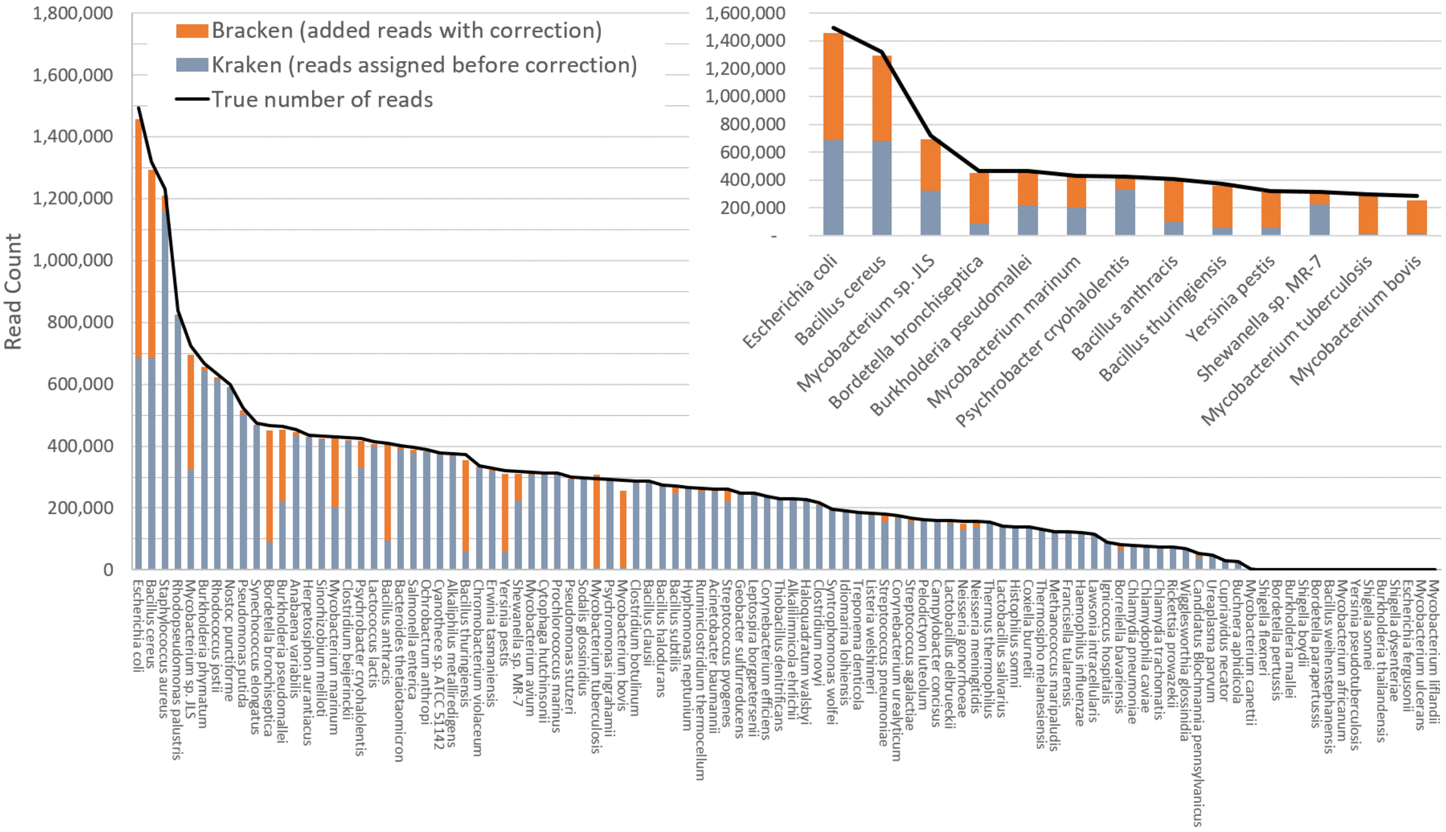

**Figure 3 Estimates of species abundance computed by Kraken (blue) and Bracken (blue + orange) for the i100 metagenomics data.** For this result, the Kraken database contained 2,635 distinct bacterial and archaeal taxa. The black line shows the true number of reads from each species. The smaller graph displays results for the subset of species for which Bracken made the largest adjustments. Precise numbers for the Kraken classification, true read counts, and Bracken estimates are contained in Table S2B.

Figure 3 shows results when using the large database to estimate abundance for the i100 genomes. This test is much more difficult because of the large number of similar and near-identical genomes in the database. Many more reads are ambiguous, mapping identically to two or more species, which means that Kraken assigns them to the LCA of those species. Nonetheless, Bracken brings the estimated abundance of all species within 4% of the true abundance, and most fall within 1%. Note that when the re-estimation procedure distributes reads from higher nodes in the taxonomy down to multiple species within a single genus, it may over-estimate one species and underestimate its sister species if the re-allocation is imperfect.

Tables S2A–S2B contains the detailed numbers for all species in Figs. 2 and 3, along with an error rate for each species in the i100 data, expressed as the difference between the true and estimated proportions. We calculated the average error as:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\left|R_{\text{true}}^{(i)}-R_{\text{est}}^{(i)}\right|}{R_{\text{true}}^{(i)}} \tag{5}$$

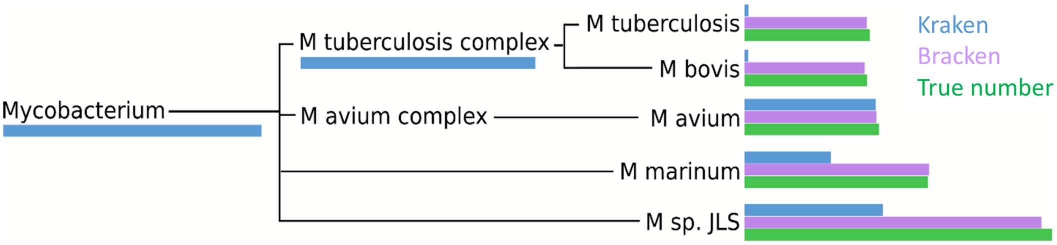

**Figure 4 Number of reads within the *Mycobacterium* genus as assigned by Kraken (blue), estimated by Bracken (purple) and compared to the true read counts (green).** Initially, Kraken assigned only 325,073 reads to *Mycobacterium sp. JLS* although 722,880 reads originated from this species. Bracken reassigned 370,601 reads from the *Mycobacterium* genus to *M. sp. JLS*. Bracken's re-estimated abundance for *M. sp. JLS* is much closer to the true read count. Table S3 contains precise numbers for all species shown here.

where $n$ is the number of species in the i100 data, $R^{(i)}_{\text{true}}$ is the true number of reads for species $i$, and $R^{(i)}_{\text{est}}$ is the Bracken estimate of the number of reads for species $i$. When using the small database, the average relative error of Bracken is 1.75% across all 85 species in the i100 data. For the larger database, the average relative error is 1.89%. We also calculated false positive rates for the i100 data as the percentage of total reads incorrectly classified after Bracken abundance estimation. For the small database, the false positive rate is 0.13% and for the large database, the false positive rate is 0.24%.

Within the i100 genomes, the five species belonging to the *Mycobacterium* genus (*M. tuberculosis, M. bovis, M. avium, M. marinum*, and *M. sp. JLS*) pose a particular challenge for abundance estimation due to the similarities among their individual genomes. For example, Kraken classified only 9,733 *M. tuberculosis* reads at the species level, and classified the remaining 285,414 reads as either *Mycobacterium* (a genus) or *M. tuberculosis complex* (a taxonomic class intermediate between genus and species), as shown in Fig. 4 and Table S3. For these *Mycobacteria* genomes, Bracken reallocated the reads from higher-level nodes to yield species abundance estimates within 4% of the true abundance. Figure 4 and Table S3 show the number of reads assigned to each species by Kraken, the true number of reads, and the number of reads assigned to each species by Bracken after abundance reestimation.

The five species of the *Mycobacterium* genus also provide an example of potential overestimation by Bracken. Bracken apportions all ambiguous reads classified by Kraken at the genus level (and above) to the existing species identified by Kraken. Because Bracken uses a probabilistic method in distributing the reads, one species may receive too many reads while another may receive too few. For example, Kraken assigned 543,916 reads to *M. tuberculosis complex*. Bracken re-allocated 296,543 of these reads to *M. tuberculosis* and the remaining 247,453 reads to *M. bovis*. When added to Kraken's original assignments, Bracken estimated that 306,792 reads belonged to *M. tuberculosis* (11,645 reads more than the true number) that 256,927 reads belonged to *M. bovis* (31,473 reads less than the true number). It is likely that some of the additional reads Bracken allocated to *M. tuberculosis* originated from *M. bovis* instead. However, despite the over- and under-estimation, Bracken's estimates fell within 4% of the true number of reads for both species.

If *M. bovis* were excluded from the database, the 8,965 reads unique to *M. bovis*, as identified by Kraken, would be unclassified, while all 543,916 reads assigned to the

*M. tuberculosis complex* would assigned to *M. tuberculosis* by Kraken. These reads would no longer be ambiguous because no other *Mycobacterium* species from the *M. tuberculosis complex* would be present in the database. In general, reads belonging to species excluded from the database will either be assigned to species with very high similarity to the missing species or will remain unclassified.

## Experiments on a real metagenomics sample created from known species

For a more realistic evaluation of the performance of Bracken, we generated new sequence data using a set of bacteria that are commonly found on healthy human skin. This mock community was assembled by combining purified DNA from nine isolates that were identified and sequenced during the initial phase of the Human Microbiome Project (*Human Microbiome Project, 2012*): *Acinetobacter radioresistens* strain SK82, *Corynebacterium amycolatum* strain SK46, *Micrococcus luteus* strain SK58, *Rhodococcus erythropolis* strain SK121, *Staphylococcus capitis* strain SK14, *Staphylococcus epidermidis* strain SK135, *Staphylococcus hominis* strain SK119, *Staphylococcus warneri* strain SK66, and *Propionibacterium acnes* strain SK137. To generate the skin microbiome community, purified DNA was obtained from the Biodefense and Emerging Infections Research Resources Repository (BEI Resources). Each of the nine bacterial isolates was grown under conditions recommended by BEI Resources, collected by centrifugation during log growth phase at a 600nm optical density ($OD^{600}$) of 0.8–1.2, and genomic DNA was isolated using MasterPure DNA isolation reagents (Epicentre). Purified genomic DNA was quantified using the high sensitivity picogreen assay (Invitrogen), pooled in equal amounts by mass, and prepared for sequencing using Nextera XT library preparation reagents (Illumina). The sample was then sequenced on a HiSeq sequencer, generating a total of 78,439,985 million read pairs (157 million reads), all of them 100 bp in length. These were then classified as pairs by Kraken, which concatenates the two reads from each pair and assigns them to a single taxonomic category.

We used Bracken to estimate both species and genus-level abundance in the skin microbiome community. In the Bracken results, the nine true species comprise over 99% of the species-level abundance estimates. The mixture was created with approximately equal amounts of each of the nine genomes, so the expectation was that each species would account for ∼11% of the total. However, as shown in Fig. 5, the estimates varied from 7.3% to 14.8%. Details for the exact number of reads assigned by Kraken and the abundance estimates by Bracken are shown in Table S4.

Deviations from the expected abundance could arise from a variety of factors. The process of quantifying DNA and mixing in equal amounts can be influenced by pipetting consistency. Second, library amplification by PCR, an integral step in the Nextera library preparation process, can exaggerate small differences in quantities and lead to significant biases in abundance (*Bowers et al., 2015*). We examined a sample of the classified reads by hand, and could find no evidence that Kraken mis-classified reads from *M. luteus* (the smallest portion of the community, estimated at 7.3%) to any of the other species or

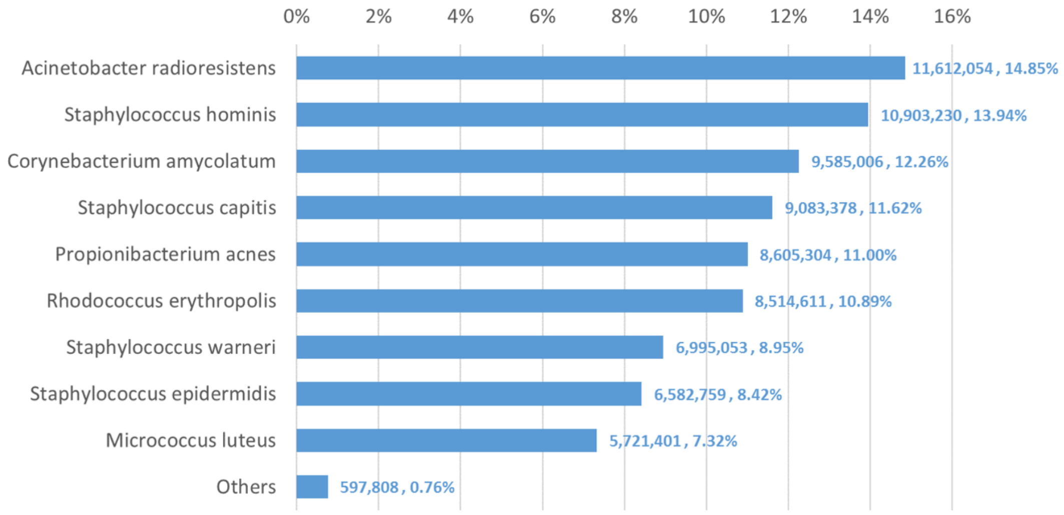

**Figure 5 Estimates of species abundance made by Bracken for the metagenomics community containing isolates of nine bacterial species commonly found on human skin.** Precise numbers can be found in Table S4.

genera. The abundances found in this data, therefore, may correspond fairly closely with the true abundances.

The genus-level abundance estimates computed by Bracken also correspond closely to the expected abundances for the six genera included in the sample. Four of the nine species belong to the genus Staphylococcus, which was thus expected to comprise 44% (4 × 11%) of the sample. The Bracken estimate was 43.3%. Each of the other genus classifications has only one species present, and their abundance estimates are the same for both genus and species.

The comparison between the Kraken classification of reads and Bracken's reassignment revealed that the nine species are sufficiently distinct to allow Kraken to classify a large majority of reads at the species level, with very few reads being classified at higher levels of the taxonomy. Specifically, Kraken classified 76.4 million reads to the nine species included in the sample. Only 1.3 million reads out of the 78.2 million total (1.6%) were classified by Kraken at the genus level or above. (The remaining reads were unclassified.) In this case Bracken does not provide a substantial benefit, because reassignment of the 1.3 million reads could yield at most a 1.6% change in the estimated composition of the sample.

## Abundance estimation timing and resource requirements

In the i100 data experiment with the small database, we used 188 gigabytes (GB) of RAM with 10 threads to build the Kraken database and generate the $k$-mer distribution file required by Bracken. In total, these steps completed in about 2 h and yielded files that can be used across multiple datasets. The resulting Kraken database and distribution files use 53 GB of space. Kraken classification of the i100 dataset took 18 min, using 10 threads and 107 GB of RAM. This step is limited by the size of the database, which is loaded into RAM during classification. Bracken alone runs in under a second, using 35 MB of RAM. The Kraken classification file for the i100 data is 1.6 GB, while Bracken abundance estimation files require ∼65 KB of space. Table S5 lists detailed timing, RAM, and space requirements for each file and step of the Bracken abundance estimation algorithm.

## CONCLUSION

Estimating the abundance of species, genera, phyla, or other taxonomic groups is a central step in the analysis of many metagenomics datasets. Metagenomics classifiers like Kraken provide a very fast and accurate way to label individual reads, and at higher taxonomic levels such as phyla, these assignments can be directly translated to abundance estimates. However, many reads cannot be unambiguously assigned to a single strain or species, for at least two reasons. First, many bacterial species are nearly identical, meaning that a read can match identically to two or more distinct species. Second, the bacterial taxonomy itself is undergoing constant revisions and updates, as genome sequencing reveals the need to re-assign species to new names. These revisions sometimes create new taxa that share near-identical sequence with a distinct species. In these situations, Kraken correctly assigns the read to a higher-level taxonomic category such as genus or family. This creates a problem in that Kraken's classifications cannot be used directly for species abundance estimation.

Bracken addresses this problem by probabilistically re-assigning reads from intermediate taxonomic nodes to the species level or above. As we have shown here, these re-assignments produce species-level abundance estimates that are very accurate, typically 98% correct or higher. For genus-level abundance, accuracy is even higher because fewer reads have ambiguous assignments at that level. For abundance estimation at higher levels, ranging from family up to phylum, Kraken's original read assignments can be used directly to create abundance estimates.

## ACKNOWLEDGEMENTS

We would like to thank Derrick Wood for valuable suggestions on the implementation of the algorithm, and Kasper Hansen for helpful comments and feedback on a draft version of this manuscript.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

This work was supported in part by the US National Institutes of Health R01-HG006677 and R01-GM083873 and by the US Army Research Office W911NF-1410490. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:
US National Institutes of Health: R01-HG006677, R01-GM083873.
US Army Research Office: W911NF-1410490.

## Competing Interests

Steven L. Salzberg is currently serving as an Academic Editor for PeerJ.

## Author Contributions

- Jennifer Lu and Florian P. Breitwieser conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, performed the computation work, reviewed drafts of the paper.
- Peter Thielen performed the experiments, analyzed the data, wrote the paper, reviewed drafts of the paper.
- Steven L. Salzberg conceived and designed the experiments, analyzed the data, wrote the paper, reviewed drafts of the paper.

## Data Availability

The following information was supplied regarding data availability:

Bracken is written in Perl and Python and is freely available for download at http://ccb.jhu.edu/software/bracken/. The reads from the skin microbiome experiment are freely available from NCBI under BioProject PRJNA316735.

## Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.104#supplemental-information.

# REFERENCES

**Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux R, Rodriguez B, Stevens RK, Wegley L, Zhang L, Zhu B, Rohwer F. 2009.** The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLos Computational Biology* **5**:e1000593 DOI 10.1371/journal.pcbi.1000593.

**Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2015.** GenBank. *Nucleic Acids Research* **43**:D30–D35 DOI 10.1093/nar/gku1216.

**Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, Ngan CY, Cheng JF, Tringe SG, Woyke T. 2015.** Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* **16**:856 DOI 10.1186/s12864-015-2063-6.

**Brosch R, Gordon SV, Marmiesse M, Brodin P, Buchrieser C, Eiglmeier K, Garnier T, Gutierrez C, Hewinson G, Kremer K, Parsons LM, Pym AS, Samper S, Van Soolingen D, Cole ST. 2002.** A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences of the United States of America* **99**:3684–3689 DOI 10.1073/pnas.052548299.

**Garnier T, Eiglmeier K, Camus JC, Medina N, Mansoor H, Pryor M, Duthoy S, Grondin S, Lacroix C, Monsempe C, Simon S, Harris B, Atkin R, Doggett J, Mayes R, Keating L, Wheeler PR, Parkhill J, Barrell BG, Cole ST, Gordon SV, Hewinson RG. 2003.** The complete genome sequence of *Mycobacterium bovis*. *Proceedings of the National Academy of Sciences of the United States of America* **100**:7877–7882 DOI 10.1073/pnas.1130426100.

**Grange JM. 2001.** *Mycobacterium bovis* infection in human beings. *Tuberculosis* **81**:71–77 DOI 10.1054/tube.2000.0263.

**Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M, Hegna I, Kolsto AB. 2000.** *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*–one species on the basis of genetic evidence. *Applied and Environmental Microbiology* **66**:2627–2630 DOI 10.1128/AEM.66.6.2627-2630.2000.

**Human Microbiome Project C. 2012.** Structure, function and diversity of the healthy human microbiome. *Nature* **486**:207–214 DOI 10.1038/nature11234.

**Lan R, Reeves PR. 2002.** *Escherichia coli* in disguise: molecular origins of Shigella. *Microbes and Infection* **4**:1125–1132 DOI 10.1016/S1286-4579(02)01637-4.

**Lindgreen S, Adair KL, Gardner PP. 2016.** An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports* **6**:Article 19233 DOI 10.1038/srep19233.

**Lindner MS, Renard BY. 2012.** Metagenomic abundance estimation and diagnostic testing on species level. *Nucleic Acids Research* **41**:e10 DOI 10.1093/nar/gks803.

**Liu Y, Lai Q, Göker M, Meier-Kolthoff JP, Wang M, Sun Y, Wang L, Shao Z. 2015.** Genomic insights into the taxonomic status of the Bacillus cereus group. *Scientific Reports* **5**:Article 14082 DOI 10.1038/srep14082.

**Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. 2015.** ConStrains identifies microbial strains in metagenomic datasets. *Nature Biotechnology* **33**:1045–1052 DOI 10.1038/nbt.3319.

**Mende DR, Waller AS, Sunagawa S, Jarvelin AI, Chan MM, Arumugam M, Raes J, Bork P. 2012.** Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* **7**:e31386 DOI 10.1371/journal.pone.0031386.

**Morgulis A, Coulouris G, Raytselis Y, Madden TL, Agarwala R, Schaffer AA. 2008.** Database indexing for production MegaBLAST searches. *Bioinformatics* **24**:1757–1764 DOI 10.1093/bioinformatics/btn322.

**Peabody MA, Van Rossum T, Lo R, Brinkman FS. 2015.** Evaluation of shotgun metagenomics sequence classification methods using in silico and *in vitro* simulated communities. *BMC Bioinformatics* **16**:363 DOI 10.1186/s12859-015-0788-5.

**Riesenfeld CS, Schloss PD, Handelsman J. 2004.** Metagenomics: genomic analysis of microbial communities. *Annual Review of Genetics* **38**:525–552 DOI 10.1146/annurev.genet.38.072902.091216.

**Schaeffer L, Pimentel H, Bray N, Melsted P, Pachter L. 2015.** Pseudoalignment for metagenomic read assignment. ArXiv preprint. arXiv:1510.07371v07372.

**Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. 2012.** Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods* **9**:811–814 DOI 10.1038/nmeth.2066.

**Sohn M, An L, Pookhao N, Li Q. 2014.** Accurate genome relative abundance estimation for closely related species in a metagnomic sample. *BMC Bioinformatics* **15**:Article 242 DOI 10.1186/1471-2105-15-242.

**The International Human Genome Sequencing Consortium. 2001.** Initial sequencing and analysis of the human genome. *Nature* **409**:860–921 DOI 10.1038/35057062.

**Thiel T, Pratte BS, Zhong J, Goodwin L, Copeland A, Lucas S, Han C, Pitluck S, Land ML, Kyrpides NC, Woyke T. 2014.** Complete genome sequence of *Anabaena variabilis* ATCC 29413. *Standards in Genomic Sciences* **9**:562–573 DOI 10.4056/sigs.3899418.

**Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, et al. 2001.** The sequence of the human genome. *Science* **291**:1304–1351 DOI 10.1126/science.1058040.

**Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO. 2004.** Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**:66–74 DOI 10.1126/science.1093857.

**Wood DE, Salzberg SL. 2014.** Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **15**:Article R46 DOI 10.1186/gb-2014-15-3-r46.

**Xia LC, Cram JA, Chen T, Fuhrman JA, Fengzhu S. 2011.** Accurate genome relative abundance estimation based on shotgun metgenomic reads. *PLoS ONE* **6**:e27992 DOI 10.1371/journal.pone.0027992.