RefSeq database growth influences the accuracy of *k*-mer-based species identification

Daniel J. Nasko[1], Sergey Koren[2], Adam M. Phillippy[2], and Todd J. Treangen[1*]

[1] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA.

[2] Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA.

* treangen@umd.edu to whom correspondence should be addressed

ABSTRACT

Accurate species-level taxonomic classification and profiling of complex microbial communities remains a challenge due to homologous regions shared among closely related species and a sparse representation of non-human associated microbes in the database. Although the database undoubtedly has a strong influence on the sensitivity of taxonomic classifiers and profilers, to date, no study has carefully explored this topic on historical RefSeq releases and explored its impact on accuracy. In this study, we examined the influence of the database, over time, on *k*-mer based sequence classification and profiling. We present three major findings: (*i*) database growth over time resulted in more classified reads, but fewer species-level classifications and more species-level misclassifications; (*ii*) Bayesian re-estimation of abundance helped to recover species-level classifications when the exact target strain was present; and (*iii*) Bayesian re-estimation struggled when the database lacked the target strain, resulting in a notable decrease in accuracy. In summary, our findings suggest that the growth of RefSeq over time has strongly influenced the accuracy of *k*-mer based classification and profiling methods, resulting in

different classification results depending on the particular database used. These results suggest a need for new algorithms specially adapted for large genome collections and better measures of classification uncertainty.

Keywords: Taxonomic classification, Reference database, Metagenomics, Microbiome, Comparative analysis

INTRODUCTION

Fundamental questions of a metagenomic survey are: (*i*) what microbes are present in each sample, (*ii*) how abundant is each organism identified in a sample, *(iii)* what role might each microbe play (i.e. what gene functions are present) and (*iv*) how do the previous observations change across samples and time? Specifically, there have been numerous studies highlighting the utility of metagenomic datasets for pathogen detection, disease indicators, and health [1,2]. Addressing each of these fundamental questions is predicated on the ability to assign taxonomy and gene function to unknown sequences.

Several new tools and approaches for taxonomic identification of DNA sequences have emerged [3–5], in addition to community-driven 'bake-offs' and benchmarks [6]. *k*-mer based classification methods such as Kraken or CLARK [3,7] are notable for their exceptional speed and specificity, as both are capable of analyzing hundreds of millions of short reads (ca. 100 base pairs) in a CPU minute. These *k*-mer based algorithms use heuristics to identify unique, informative, *k*-length subsequences (*k*-mers) within a database to help improve both speed and accuracy. A challenge for *k*-mer based classification approaches is that closely related species and strains often contain

47 many identical sequences within their genomes. This challenge is typically addressed by
48 assigning the query sequence with the lowest common ancestor (LCA) of all species that share
49 the sequence. A comprehensive benchmarking survey indicated that Kraken offered the best F_1
50 score (a measure considering both precision and recall) among the *k*-mer based taxonomic
51 classifiers evaluated at the species level [8]. Bracken, a Bayesian method that refines Kraken
52 results, is capable of estimating how much of each species is present among a set of ambiguous
53 species classifications by probabilistically re-distributing reads in a taxonomic tree [9]. We thus
54 selected Kraken and Bracken as representative tools from the genre of *k*-mer based classification
55 methods. The focus on this study was not to examine a specific software tool, but rather to
56 decouple the performance of *k*-mer based methods from the underlying database.
57
58 Available *k*-mer based methods for taxonomic identification and microbiome profiling rely on
59 existing reference databases. While several investigations have examined the influence of
60 contamination in specific database releases, and identified idiosyncrasies specific to a release
61 [10,11], no study has examined the specific influence of perhaps the most popular database from
62 which to build classification databases, the repository of sequenced and assembled microbes
63 (RefSeq), across all releases of the database. Additionally, metagenomic classification and
64 profiling tools are commonly compared to each other using simulated datasets on a fixed
65 database, with leave-one-out analysis, but never compared to each other across recent trajectories
66 in database growth. The aim of this study was to elucidate the influence of RefSeq database
67 growth over time on the performance of *k*-mer based taxonomic identification tools.
68

69 **RESULTS**

RefSeq database growth

Since its release in June 2003 bacterial RefSeq, on average, has doubled in size (giga base pairs, Gbp) every 1.5 years (Fig. 1A), with the number of unique 31-mers in the database growing at a similar rate (Fig. 1B). A more recent release, bacterial RefSeq version 84 (released 9/11/2017), totaled over 700 Gbp of sequence data. The Simpson's index of diversity is a metric with values between zero and one that reports the probability that two individuals randomly selected from a sample will not belong to the same species. Samples with a high Simpson's index of diversity (i.e. closer to one) may be considered more diverse than those with low values (i.e. closer to zero). The diversity for each version of the bacterial RefSeq database increased until April 2013 where the Simpson's index of diversity for each subsequent bacterial RefSeq release has trended downward (Fig. 1C). A slower growth is also seen in the number of new bacterial species in each RefSeq version, indicating many of the same species are being sequenced repeatedly (Fig. 1D).

Taxonomic classification over time with a simulated metagenome

Kraken's own simulated validation set of ten known genomes was searched against nine versions of bacterial RefSeq (1, 10, 20, 30, 40, 50, 60, 70, 80) and the MiniKraken database (4GB version) (Fig. 2). The accuracy of each Kraken run depends on the RefSeq version used in the search (Fig. 2; Table 1). Correct genus-level classifications increased as RefSeq grew, but correct species-level classifications peaked at version 30 and tended to decline thereafter (Fig. 2). The decrease in correct species classifications is due to more closely-related genomes appearing over time in RefSeq, making it difficult for the classifier to distinguish them and forcing a move up to the genus level. Overall, misclassified species-level calls were consistently rare, as reads

were misclassified at the species level an average of 7% of the time (Table 1; Fig. 2). The fraction of reads classified at any taxonomic level, regardless of accuracy, increases as RefSeq grows over time (Fig. 3). However, the fraction of species-level assignments (again, regardless of accuracy) peaks at RefSeq version 30 and begins to decline thereafter, while the fraction of genus-level classifications begins to increase.

Bracken was used to re-estimate the abundances of classifications made by Kraken when searching the simulated reads against eight bacterial RefSeq database versions (1, 10, 20, 30, 40, 50, 60, 70). Bracken first derives probabilities that describe how much sequence from each genome is identical to other genomes in the database. This step requires searching a Kraken database against itself with Kraken, which could not be performed for the MiniKraken DB (as there is no FASTA file for this database) or bacterial RefSeq version 80 (as it would require extensive computation for a database that size). Bracken was able to re-estimate species abundances for 95% of the input data using RefSeq version 70, while Kraken only classified 51% of reads at the species level. Because Bracken may probabilistically distribute a single read's classification across multiple taxonomy nodes, its performance must be measured in terms of the predicted abundances. Bracken typically included the correct species in its re-estimation, but sometimes included incorrect species in the abundance estimation (on average 15% of reads were associated with a genome outside of the ten knowns).

Taxonomic classification of difficult to classify genomes over time

The challenging nature of classifying sequences belonging to the *Bacillus cereus* sensu latu group has been previously documented [12,13]. The *B. anthracis* species within this group is a well-

116 defined monophyletic subclade of the larger *B. cereus* group, and the base of the *B. anthracis*
117 clade is commonly denoted by a single nonsense mutation in the *plcR* gene [14] which is conserved
118 in all known *B. anthracis* genomes and has been shown to confer a regulatory mutation essential
119 for maintaining the pXO1 and pXO2 plasmids that carry the virulence factors characteristic of
120 anthrax [15]. However, not all *B. anthracis* cause disease in humans, such as *B. anthracis* Sterne
121 (missing the pXO2 plasmid) and some *B. cereus* strains do cause anthrax-like disease [16],
122 complicating a precise species definition. Thus, it is not a surprise that accurate species-level
123 classification within this group has proven challenging for *k*-mer based methods, especially those
124 methods not based on phylogenetic evidence. To demonstrate how difficult sequences from this
125 group have been to classify over time, simulated reads were created for two *Bacillus cereus*
126 strains. The first, *B. cereus* VD118, is a strain available in RefSeq version 60 and beyond, and
127 the second, *B. cereus* ISSFR-23F [17], was recently isolated from the International Space Station
128 and is not present in any of the RefSeq releases tested. It is phylogenetically close to *B.*
129 *anthracis*, but lacks the phylogenetic and species characteristics of *B. anthracis*. Again, as
130 bacterial RefSeq grows over time, the number of genus-level classifications made by Kraken
131 increases (Fig. 4). While the number of genus-level calls made by Kraken increases over time
132 the number of unclassified and misclassified species calls decreases (most commonly *B.*
133 *anthracis*, *B. thuringensis*, and *B. weihenstephanensis*).

134

135 Bracken made species-level predictions for all reads no matter which version of bacterial RefSeq
136 was used (Fig. 4). However, the increased rate of species-level predictions came at the cost of
137 accuracy, as Bracken correctly identified *B. cereus* VD118 and *B. cereus* ISSFR-23F an average
138 of 72% and 29% of the time, respectively, across RefSeq versions 1 through 70. The fraction of

reads assigned to each *Bacillus* species varied substantially from each database tested. The range of *Bacillus* species predictions for *B. cereus* VD118 were: *B. cereus* 81% (max=100%, min=18%), *B. anthracis* 48% (max=48%, min=0%), *B. thuringiensis* 23% (max=23%, min=0%), and *B. weihenstephanensis* 76% (max=76%, min=0%). While the range of *Bacillus* species predictions for *B. cereus* ISSFR-23F were: *B. cereus* 45% (max=50%, min=5%), *B. anthracis* 90% (max=95%, min=5%), and *B. thuringiensis* 54% (max=54%, min=0%).

CPU/Memory performance over time

Historical bacterial RefSeq versions were recreated and used to build Kraken databases with default settings. While most databases were constructed with ease and in less than a day, version 70 required 500 GB of RAM and 2 days (single compute node using on 64 cores), while version 80 required ca. 2.5 TB of RAM and ca. 11 days (single compute node using on 64 cores). Given this trend, future releases will likely require over 4 TB of RAM and weeks of computation to build, putting into question the feasibility of building and profiling *k*-mer databases on future RefSeq versions. Recent studies [18] have suggested alternative approaches for database construction that would help to circumvent future computational bottlenecks.

DISCUSSION

The results of our study support three conclusions: (*i*) the RefSeq bacterial database composition and diversity is dynamic, varying from release to release; (*ii*) the database composition strongly influences the performance of *k*-mer based taxonomic identification methods, and (*iii*) Bayesian based methods can help mitigate some of the effect, but struggle with novel genomes that have close relatives in the database.

Database influences on k-mer based taxonomic classification

Using Bracken, the majority of *Bacillus cereus* ISSFR-23F simulated reads were not correctly assigned to *B. cereus* but were more frequently mis-assigned as *Bacillus anthracis* or *Bacillus thuringiensis* (Fig. 4B). This, in part, is not surprising as two of the three species in this group, *B. cereus* and *B. thuringiensis*, have no clear phylogenetically defined boundary, though *B. anthracis* is phylogenetically distinct from *B. cereus* and *B. thuringiensis*. Furthermore, any two genomes within the *Bacillus cereus* sensu lato group are likely to be over 98% identical [9]. Given that *k*-mer based methods are not phylogenetically-grounded, but rather based on sequence composition, they are susceptible to misidentification in clades where the taxonomy is in partial conflict with phylogeny, such as the *Bacillus cereus* sensu lato group. One clear example of misidentification within this group was the false identification Anthrax in public transit systems [19,20].

Another observation worth highlighting is that the fraction of simulated reads classified as one of the three *B. cereus* sensu lato species varied across database versions (Fig. 4), with the exception of *B. cereus* VD118, which was present in RefSeq releases 60 and 70 (Fig. 4A). The variation in species classifications across database versions indicates that even when using the same tools to analyze the same dataset, the conclusions derived from this analysis can vary substantially depending on which version of a database you are searching against, especially for genomes belonging to difficult to classify species (i.e. require phylogenetic-based approaches).

Imperfect data

185 The genomic data deluge has helped to expand public repositories with a broader and deeper
186 view of the tree of life, but has also brought with it contamination and misclassification.
187 Contamination in public databases is well-documented [21] and represents an additional
188 confounding factor for *k*-mer based methods. While several custom tools have been built to deal
189 with imperfect data [22], there is a need for database 'cleaning' tools that can preprocess a database
190 and evaluate it for both contamination (genome assemblies that contain a mixture of species) and
191 misclassified species and strains (genomes that are assigned a taxonomic ID that is inconsistent
192 with its similarity to other genomes in the database). The misclassification issue often is in the
193 eye of the beholder; species have been named based on morphology, ecological niche, toxin
194 presence/absence, isolation location, 16S phylogenetic placement, and average nucleotide
195 identity across the genome. This, coupled with an often ambiguous species concept in microbial
196 genomes due to horizontal gene transfer and mobile elements [23], brings into question the reliance
197 on the current taxonomic structure for assigning names to microbes sequenced in metagenomic
198 samples. A more robust approach would be for the classification databases to derive their own
199 hierarchical structure directly from the data, rather than taxonomy, and then map back the
200 internally derived hierarchy to widely-used taxonomic names.
201

202 ## CONCLUSION

203 Our findings demonstrate that changes in RefSeq over time have influenced the accuracy of *k*-
204 mer based classification and profiling methods. Bayesian re-estimation approaches are helpful
205 for species or strain level prediction but can result in false positives and are computationally
206 prohibitive with larger databases. Despite recent progress in *k*-mer based methods for
207 metagenome profiling and classification these tools should likely be used as step one in a multi-

208 step process, which also includes read mapping, assembly, feature prediction, and annotation.
209 Additionally, priority should be given to the breadth, not depth, of species added to reference
210 databases over time.
211

212 # METHODS

213 **Acquisition of bacterial RefSeq databases versions 1 through 80**

214 FASTA files of previous versions of bacterial RefSeq are not publically available for download.
215 Therefore, sequences from previous versions of bacterial RefSeq were acquired using custom
216 scripts (https://github.com/dnasko/refseq_rollback). Briefly, the process involved downloading
217 the current bacterial RefSeq release (ver. 84 as of the date of the analysis) FASTA files
218 (ftp.ncbi.nlm.nih.gov/refseq/release/bacteria) and concatenating them into one file. Then, the
219 catalog file associated with the desired version is downloaded
220 (ftp.ncbi.nlm.nih.gov/refseq/release/release-catalog/archive), which contains the identifiers for
221 sequences present in that version of bacterial RefSeq. Sequence identifiers in that version's
222 catalog file are pulled from the current RefSeq FASTA file and written to a new file. Using the
223 refseq_rollback.pl script any version of bacterial RefSeq can be created. For this study only
224 versions 1, 10, 20, 30, 40, 50, 60, 70, and 80 were recreated.
225

226 **Taxonomic classification on simulated datasets**

227 Two simulated read datasets were used to test Kraken and Bracken performance with different
228 versions of the bacterial RefSeq database. The first simulated dataset was downloaded from the
229 Kraken website (ccb.jhu.edu/software/kraken) and was previously used in the Kraken manuscript
230 as a validation set [3]. Briefly, this simulated dataset was composed of 10 known bacterial species:

231 *Aeromonas hydrophila* SSU, *Bacillus cereus* VD118, *Bacteroides fragilis* HMW 615,
232 *Mycobacterium abscessus* 6G-0125-R, *Pelosinus fermentans* A11, *Rhodobacter sphaeroides*
233 2.4.1, *Staphylococcus aureus* M0927, *Streptococcus pneumoniae* TIGR4, *Vibrio cholerae*
234 CP1032(5), and *Xanthomonas axonopodis* pv. Manihotis UA323. Each genome had 1,000
235 single-end reads (101 bp in size) for a total of 10,000 reads. We selected this dataset as it has
236 been widely used as a benchmark for other *k*-mer based classification methods [3,7] and represents
237 a breadth of species. This simulated read dataset was classified against each of the recreated
238 bacterial RefSeq databases using Kraken (ver 1.0) with default settings.
239
240 To test the ability to classify reads from genomes not in the bacterial RefSeq database 10,000
241 simulated single-end Illumina reads (101 bp) were created using Grinder [24] with default settings
242 from: (*i*) a *Bacillus cereus* genome, *B. cereus* VD118, not present in RefSeq until verion 60 and
243 beyond; and (*ii*) a novel *B. cereus* genome, *B. cereus* ISSFR-23F [17], never present in any of the
244 RefSeq versions tested. We decided to use these genomes as they are members of the *B. cereus*
245 sensu lato group, containing a collection of species that are known to be challenging for *k*-mer
246 methods to distinguish between [19,20]. These datasets were classified with Kraken (ver. 1.0) and
247 Bracken (ver. 1.0.0) [9] both with default settings (Bracken "read-length" set to 101).
248
249 **Running Bracken on Kraken output**
250 Bracken (ver. 1.0.0) was run on the output of each Kraken search (except for release 80 and
251 KrakenMiniDB). Default parameters were used except for "read-length", which was set to 101.
252

253 **Bacterial RefSeq diversity metric calculations**

254 Diversity metrics were calculated for every version of bacterial RefSeq (1-84) by parsing the
255 catalog files for each version. An operational taxonomic unit (OTU) table was constructed using
256 the NCBI taxonomy identifiers as taxonomic units (see create_otu_table.pl in the refseq_rollback
257 repository). The OTU table was imported to QIIME (ver. MacQIIME 1.9.1-20150604) [25].
258 Diversity metrics (Simpson, Shannon, Richness) were calculated using the "alpha_diversity.py"
259 script and plotted using the R base package.

260

261 ## ABBREVIATIONS

262 OTU: Operational taxonomic unit; LCA: Lowest common ancestor

263 ## DECLARATIONS

264 **Acknowledgements**

265 This work utilized the computational resources of the NIH HPC Biowulf cluster
266 (https://hpc.nih.gov). The authors would like to thank Mihai Pop for his feedback and
267 discussion of this project in its early development.

268

269 **Funding**

270 S.K. and A.M.P. were supported by the Intramural Research Program of the National
271 Human Genome Research Institute, National Institutes of Health. D.J.N. and T.J.T
272 were supported by the FunGCAT program from the Office of the Director of National
273 Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via

the Army Research Office (ARO) under Federal Award No. W911NF-17-2-0089. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government.

Availability of Data and Materials

Scripts used in this analysis are available on GitHub (github.com/dnasko/refseq_rollback). Datasets and genomes used in this analysis are available online and referenced in the text.

Authors' contributions

T.J.T. and D.J.N. designed the experiments. D.J.N. wrote the analysis scripts. D.J.N., S.K, and T.J.T. performed the experiments. D.J.N., S.K., A.M.P. and T.J.T. wrote the paper.

Ethics approval and consent to participate

NA

Consent for publication

NA

Competing interests

NA

Additional files

REFERENCES

1. Nakamura, S. *et al.* Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One* **4,** 1–8 (2009).
2. Greenblum, S., Turnbaugh, P. J. & Borenstein, E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and in fl ammatory bowel disease. *Proc. Natl. Acad. Sci.* **109,** 594–599 (2012).
3. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** R46 (2014).
4. Nguyen, N. P., Mirarab, S., Liu, B., Pop, M. & Warnow, T. TIPP: Taxonomic identification and phylogenetic profiling. *Bioinformatics* **30,** 3548–3555 (2014).
5. Ainsworth, D., Sternberg, M. J. E., Raczy, C. & Butcher, S. A. k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic Acids Res.* **45,** 1649–1656 (2017).
6. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. *Nat. Methods* **14,** 1063–1071 (2017).
7. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16,** 1–13 (2015).
8. McIntyre, A. B. R. *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **18,** 1–19 (2017).
9. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3,** e104 (2017).
10. Schäffer, A. A. *et al.* VecScreen_plus_taxonomy: imposing a tax(onomy) increase on vector contamination screening. *Bioinformatics* 1–5 (2017). doi:10.1093/bioinformatics/btx669
11. Pible, O., Hartmann, E. M., Imbert, G. & Armengaud, J. The importance of recognizing and reporting sequence database contamination for proteomics. *EuPA Open Proteomics* **3,** 246–249 (2014).
12. Helgason, E. *et al.* Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis —— One Species on the Basis of Genetic Evidence One Species on the Basis of Genetic Evidence. *Appl. Environ. Microbiol.* **66,** 2627–2630 (2000).
13. Zwick, M. E. *et al.* Genomic characterization of the *Bacillus cereus* sensu lato species: Backdrop to the evolution of *Bacillus anthracis*. *Genome Res.* **22,** 1512–1524 (2012).
14. Keim, P. *et al.* Anthrax molecular epidemiology and forensics: Using the appropriate marker for different evolutionary scales. *Infect. Genet. Evol.* **4,** 205–213 (2004).
15. Mignot, T. *et al.* The incompatibility between the PlcR- and AtxA-controlled regulons may have selected a nonsense mutation in Bacillus anthracis. *Mol. Microbiol.* **42,** 1189–1198 (2001).
16. Klee, S. R. *et al.* The genome of a Bacillus isolate causing anthrax in chimpanzees combines chromosomal properties of B. cereus with B. anthracis virulence plasmids. *PLoS*

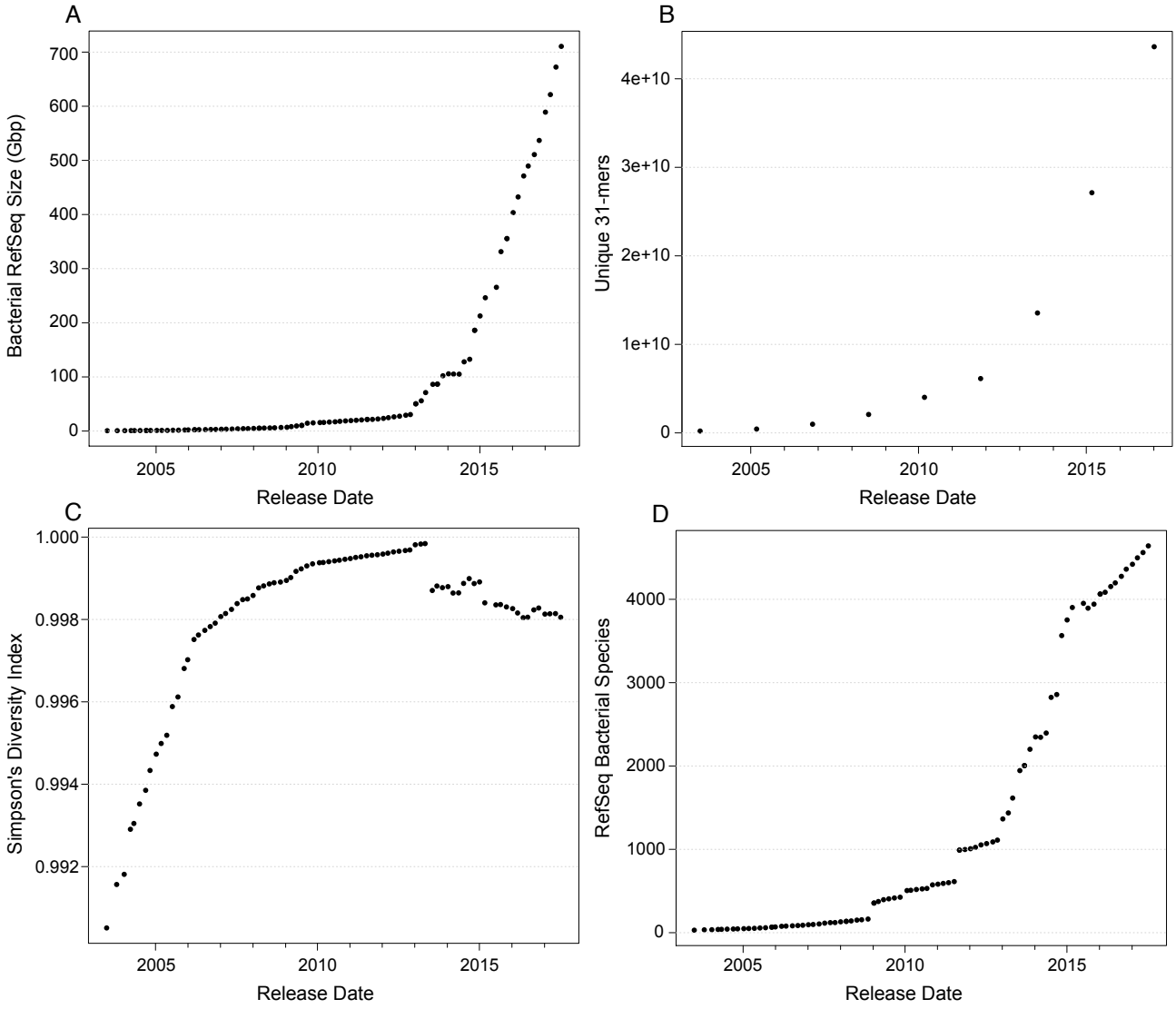
One **5,** (2010).

17. Venkateswaran, K. *et al.* Draft genome sequences from a novel clade of Bacillus cereus Sensu Lato strains, isolated from the International Space Station. *Genome Announc.* **5,** e00680-17 (2017).
18. Zhou, W., Gay, N. & Oh, J. ReprDB and panDB: minimalist databases with maximal microbial representation. *Microbiome* **6,** 15 (2018).
19. Afshinnekoo, E. *et al.* Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics. *Cell Syst.* **1,** 72–87 (2015).
20. The MetaSUB International Consortium. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. *Microbiome* **4,** 24 (2016).
21. Merchant, S., Wood, D. E. & Salzberg, S. L. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* **2,** e675 (2014).
22. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* **6,** (2011).
23. Cohan, F. M. What are Bacterial Species? *Annu. Rev. Microbiol.* **56,** 457–487 (2002).
24. Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P. & Tyson, G. W. Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res* **40,** e94 (2012).
25. Caporaso, J. G. *et al.* QIIME allows analysis of high- throughput community sequencing data. *Nat. Publ. Gr.* **7,** 335–336 (2010).

360 **Table 1. Fractions of unclassified (Unclass.), correctly classified (Correct), and misclassified**
361 **(Misclass.) simulated reads from ten genomes using Kraken against different versions of**
362 **bacterial RefSeq**.

			Genus		Species	
Release	**Date**	**Unclass.**	**Correct**	**Misclass.**	**Correct**	**Misclass.**
1	2003-06-30	0.62	0.38	0.00	0.29	0.08
10	2005-03-06	0.53	0.46	0.01	0.38	0.07
20	2006-11-05	0.49	0.49	0.01	0.40	0.08
30	2008-07-07	0.25	0.74	0.00	0.60	0.07
40	2010-05-07	0.22	0.77	0.00	0.54	0.08
50	2011-11-08	0.21	0.78	0.01	0.52	0.07
60	2013-07-19	0.03	0.96	0.00	0.57	0.09
70	2016-03-03	0.03	0.95	0.01	0.42	0.09
80	2017-01-09	0.03	0.94	0.01	0.28	0.08

363



364

365 **Figure 1: Simpson diversity index of bacterial RefSeq has decreased every release since**

366 **April 2013.** (A) The number of base pairs in bacterial RefSeq continues to grow exponentially,

367 but (B) the number of unique 31-mers and (D) the number of bacterial species added increases

368 slower. (C) The Simpson's diversity index grew every release up to April 2013 where it has

369 declined every release since.

370

371

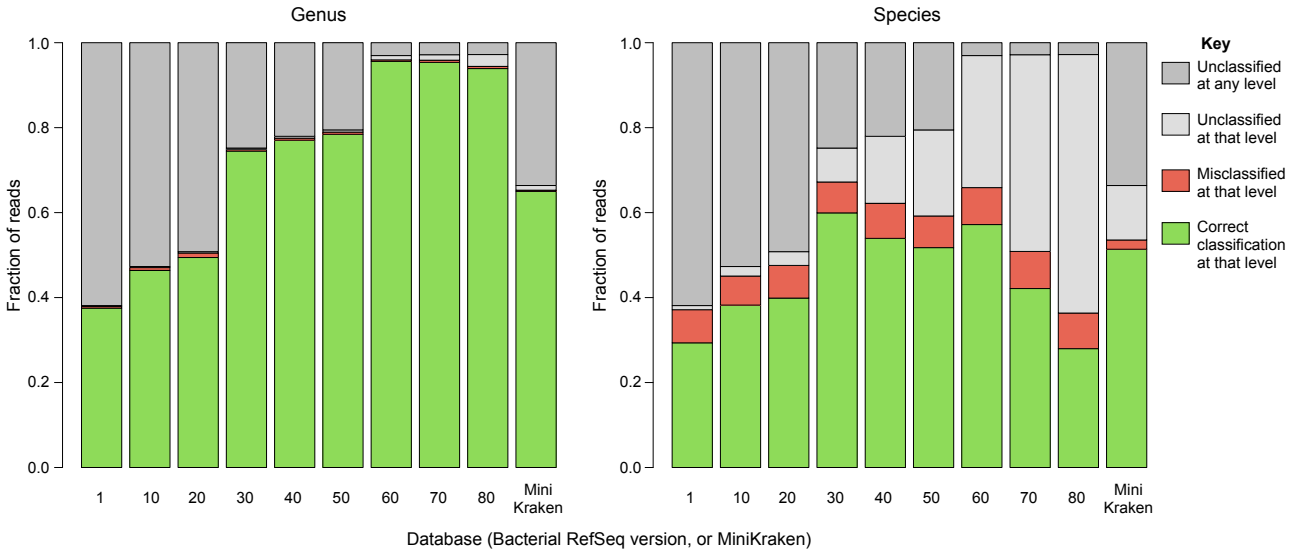


Figure 2: The fraction of correct species classifications (right) decreases in later RefSeq database versions because they are only being classified at the genus level (left). Kraken classification results of simulated reads from known genomes against nine versions the bacterial RefSeq database and the MiniKraken database. Misclassifications at the genus and species levels remain consistently low across database versions.

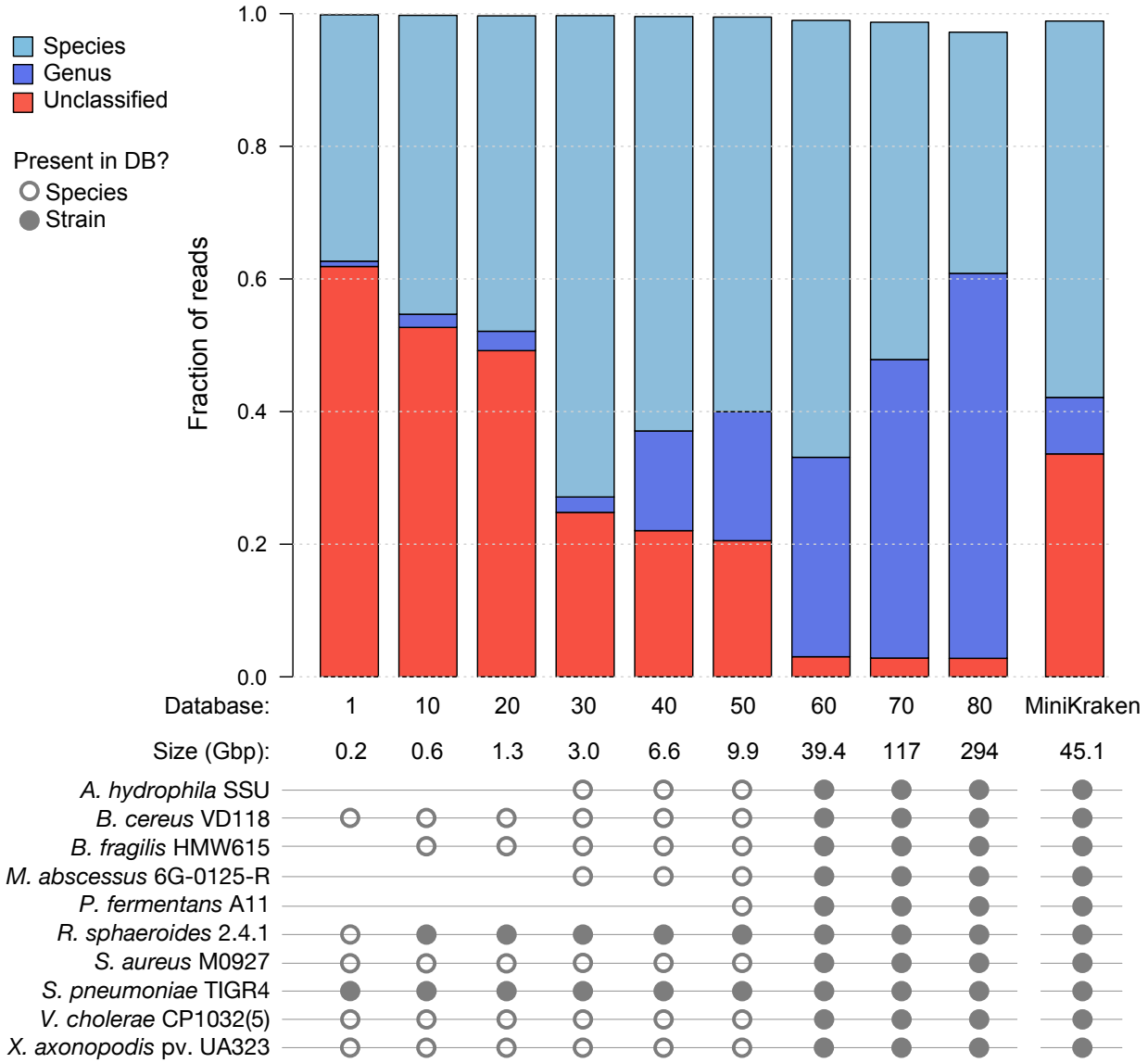


Figure 3: **Species-level classifications decrease, and genus-level classifications increase as bacterial RefSeq grows.** Fraction of simulated reads classified at different taxonomic levels, regardless of accuracy, using Kraken against ten databases. The circles below indicate when each genome's species/strain is in a database. Although the MiniKraken database contains all 10 genomes it yields results comparable to bacterial RefSeq version 40.

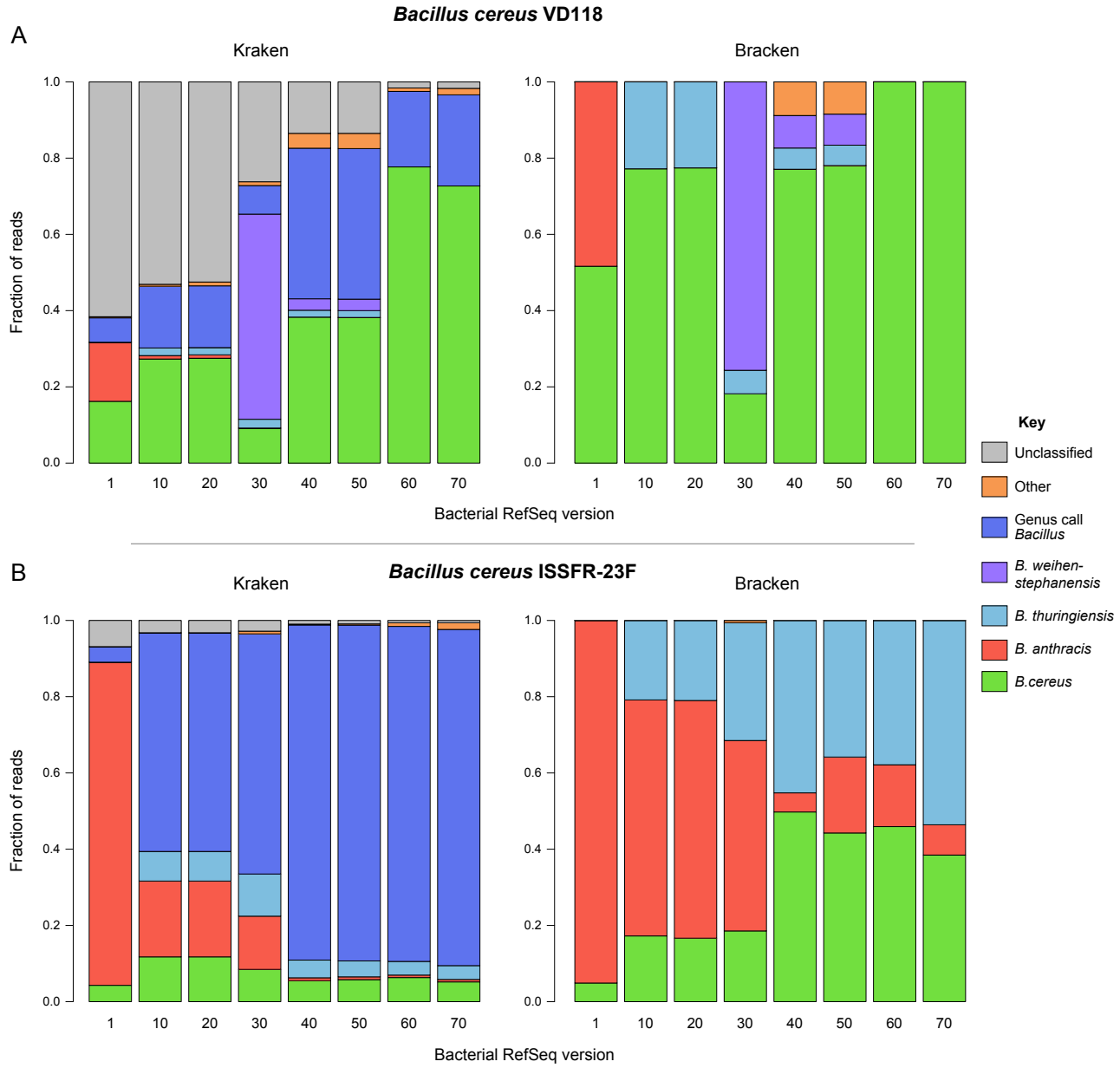


Figure 4: The fraction of simulated reads classified among *Bacillus* species varied considerably depending on which RefSeq version was used, demonstrating the influence of the database on a *k*-mer based taxonomic classification. (A) Classifying simulated *B. cereus* VD118 reads with Kraken (left) and Bracken (right) against different version of RefSeq. Species-level classifications varied, and the fraction of unclassified reads decreased with Kraken,

as the database grew. Once *B. cereus* VD118 appeared in the database (ver. 60) Bracken correctly classified every read. **(B)** Species-level classifications decrease with Kraken as RefSeq grows using simulated reads from an environmental *Bacillus cereus* not in RefSeq. Fraction of simulated *B. cereus* ISSFR-23F reads classified using Kraken ver. 1.0 (left) and Bracken ver. 1.0.0 (right) against different versions of bacterial RefSeq. Bracken classification pushed all reads to a species-level call, though these classifications were often for other *Bacillus* species.