Removal of sequencing adapter contamination improves microbial genome databases

Andrew Moeller (andrew.h.moeller@gmail.com)
Princeton University https://orcid.org/0000-0002-8377-4647
Brian Dillard
Cornell University https://orcid.org/0000-0003-1845-2980
Samantha Goldman
Cornell University
Madalena Real
Cornell University
Daniel Sprockett
Cornell University

Brief Communication

Keywords:

Posted Date: January 23rd, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3888769/v1

Additional Declarations: There is **NO** Competing Interest.

Abstract

Advances in assembling microbial genomes have led to growth of reference genome databases, which have been transformative for applied and basic microbiome research. Here we show that published microbial genome databases from humans, mice, cows, pigs, fish, honeybees, and marine environments contain significant levels of sequencing adapter contamination that systematically reduces assembly quality. By removing the adapter-contaminated ends of contiguous sequences and reassembling, we improve the accuracy and contiguousness of genome assemblies in these databases.

Main Text

Recent methodological advances have enabled the generation of unprecedented numbers of microbial genome sequences from eukaryotic hosts and free-living habitats[1,2,3]. Large-scale databases of thousands of microbial genomes from isolates and metagenomes (i.e., metagenome-assembled genomes—MAGs) are now available from humans[4], mice[5], cows[6], pigs[7], chicken[8], fish[3], and honeybees[9] as well as marine environments[3]. These resources represent milestones for microbiome research, enabling previously intractable functional and evolutionary studies[10,11].

The large size of recently published microbial genome databases requires automated approaches for inspection and quality control of individual assemblies[12]. For example, automated tools for detecting chimeras and strain heterogeneity have been developed[13,14]. Sequencing adapter contamination is a known issue with assembly of reads from commonly used technologies (e.g., Illumina) in which sequences from adapters used during the sequencing process are erroneously incorporated into assemblies[15]. To mitigate this issue, studies that generated the microbial genomes in recently published microbial genome databases made efforts to remove adapter sequences from reads using tools such as cutadapt[16] prior to assembly[10], such that adapter contamination is not expected to be prevalent in these databases. Here, we demonstrate a significant extent of sequencing adapter contamination in recently published microbial genome databases—and develop an approach for its elimination—with the goal of improving the accuracy, contiguousness, and utility of these resources for future studies.

To evaluate whether an assembly contains significant evidence of sequencing-adapter contamination, it is first necessary to calculate the baseline rate at which adapter sequences are expected to be observed by chance in a genome assembly of a given length. Illumina sequencing of TruSeq libraries employs the 12-base universal adapter sequence 'AGATCGGAAGAG'. The probability of observing a sequence of 12 bases in length by chance is $\frac{1}{4^{12}}$, assuming equal probability of each nucleotide at each site and independence among sites. Thus, 'λ', the number of adapter sequences expected to be observed by chance in an assembly of 'X' length containing 'y' contiguous sequences (contigs) (each of which contains 11 bases that cannot correspond to the first base of the 12-base adapter sequence), is:

$$\text{Equation1} : \lambda = \frac{(X - 11y)}{4^{12}}$$

and the probability of observing ‘O’ greater than or equal to a specific number of ‘k’ sequences in a sequence of ‘X’ sites can be calculated using the Poisson cumulative distribution function (Online Methods):

$$\text{Equation2}: \Pr(O \geq k) = 1 - e^{-\lambda} \sum_{j=0}^{\lfloor k-1 \rfloor} \frac{\lambda^j}{j!}$$

Equation 2 provides a *p*-value corresponding to the probability of observing by chance a number of adapter sequences greater than or equal to the number of adapter sequences observed in a real assembly.

Using this approach, we quantified adapter enrichment in every species reference genome assembly in published microbial genome databases from environments represented in MGnify[3], including ‘human gut’, ‘human oral’, ‘human vaginal’, ‘mouse gut’, ‘pig gut’, ‘cow rumen’, ‘honeybee gut’, ‘non-model fish gut’, ‘zebrafish fecal’, ‘chicken gut’, and ‘marine’. The number of adapter sequences observed per assembly (including both forward and reverse complement orientations of the adapter sequence) ranged from 0 to 805, with a *Paenibacillus lactis* assembly from the human gut (accession MGYG000003402) displaying the most adapter sequences. A histogram of assemblies containing 10 or more adapter sequences is shown in Fig. 1a. By chance, only ~ 157, ~15.7, and ~ 1.57e-12 assemblies were expected to be observed at the thresholds of *p*-value < 0.01, 0.001, and 1e-16, respectively. In contrast, of the 15657 species reference genome assemblies in all MGnify databases, 1110, 888, and 433 assemblies contained significant enrichment of adapters at these *p*-value thresholds, respectively. (Fig. 1). Thus, although each study that generated the original microbial genomes reported efforts to remove adapter sequences from raw reads using published tools such as ‘cutadapt’[16], these results show significant adapter contamination in microbial genome assemblies in published databases.

We next tested whether each individual database contained more assemblies showing significant evidence of adapter contamination than expected by chance. We found that 8 out of the 11 databases contained > 1.5-fold more assemblies displaying an enrichment of adapter sequences at a threshold of *p*-value < 0.01 than expected by chance (Fig. 1c–j). Enrichment of adapters was evident in assemblies derived from isolates as well as MAGs (Table S1). A table of summary statistics reporting the number of adapter sequences observed per assembly, the expected number per assembly, the *p*-value, and other information about the location of adapter sequences within assemblies is presented in Table S1. Cumulatively, these results demonstrate significant adapter contamination in published microbial genome databases.

Interestingly, adapter sequences were concentrated at the ends of contigs, and the reverse complements of adapter sequences were concentrated at the beginnings of contigs (Table S1) (Fig. 2a). For example, in the *Paenibacillus lactis* assembly containing 319 adapter sequences in the forward orientation, the average distance of this sequence to the end of the contig in which it was found was only ~ 11 bases, with a maximum distance of 75 bases and a minimum distance of 1 base, despite the average length of

contigs in this assembly being ~ 2900 bases (Fig. 2b, Table S1). Conversely, the reverse complements of the adapter sequence were clustered near the beginnings of contigs in this assembly (Fig. 2c, Table S1). Instances of the adapter sequence were also adjacent to portions of known forward- or reverse-specific adapter sequences ('CACACGTCTGAACTCCAGTCA' and 'CGTCGTGTAGGGAAAGAGTGT', respectively) or their reverse complements (Fig. 2b, c). Concentration of contamination at the beginning or ends of contigs was also observed in the other adapter-contaminated assemblies (Table S1).

That adapter contamination was clustered at the beginnings or ends of contigs is consistent with the possibility that adapter contamination broke contigs during the assembly process. We reasoned that the contiguousness of assemblies might be improved by trimming the ends of contaminated contigs and attempting to stitch together the trimmed contigs. We trimmed the last (or first) 450 bases of every contig containing an adapter sequence within 300 bases (corresponding to ~ 3 reads generated by 150bp Illumina sequencing) of the end (or beginning) of the contig—thereby removing the adapter sequences and their flanking regions—and reassembled the trimmed contigs of every species reference genome assembly (Online Methods). Reassembly increased contiguousness in 649 of the 1110 assemblies with significant adapter contamination at the *p*-value < 0.01 threshold. On average, ~ 2 contigs per assembly, corresponding to ~ 0.8% of contigs, were able to be merged with other contigs by reassembly after removing adapter contamination by trimming the original contigs. This approach improved contiguousness of assemblies in each database (Fig. 2d). Moreover, we observed a positive relationship between the number of adapter sequences present in an assembly and the number of contigs that were able to be merged with other contigs by reassembly after removing adapter contamination (Fig. 2e) (generalized linear model with Poisson-distributed errors for count data, *p*-value = 1.99e-5). This result further indicates that adapter contamination has negative effects on assembly contiguousness and that these negative effects can be mitigated by removal of adapter contamination and reassembly.

Corrected assemblies generated by this study (trimmed assembles and reassemblies) are available at https://zenodo.org/records/10547057. Scripts for detecting and assessing the extent of adapter contamination in assemblies, removing the ends of adapter-contaminated contigs, and reassembling trimmed contigs are available at github.com/CUMoellerLab/MalAdapter. The increased contiguousness of assemblies may improve their utility for studies focused on structural features of microbial genomes (e.g., synteny, genomic rearrangement, etc.). Moreover, removing adapter sequences increases assembly accuracy, which may improve the utility of assemblies for any future study.

Online Methods

Data sources

Genome assemblies for this study were downloaded from the MGnify[3] ftp site at https://ftp.ebi.ac.uk/pub/databases/. Most recent versions of each database were used as follows: 'chicken-gut' = v1.0.1, 'cow-rumen' = v1.0.1, 'honeybee-gut'=v1.0.1, 'human-gut'=v2.0.2, 'human-

oral'=v1.0.1, 'human-vaginal'=v1.0, 'marine'=v1.0, 'mouse-gut'=v1.0, 'non-model-fish-gut'=v2.0, 'pig-gut'=v1.0, 'zebrafish-fecal'=v1.0.

Derivation of expectations and probabilities

The number '11' in Eq. 1 reflects that the last 11 bases in each contig cannot be the start of a 12-base adapter sequence. The term 'k minus 1' in Eq. 2 allows the calculation of the probability of observing greater than or equal to 'k' adapter sequences.

Identification and removal of adapter sequences

Illumina universal adapter sequences ('AGATCGGAAGAG') and their reverse complements ('CTCTTCCGATCT') were identified and counted in assemblies using custom bash scripts available at github.com/CUMoellerLab/MalAdapter. When adapter sequences were detected within the first or last 300 bases of a contig, the first or last 450 of the contig was removed, respectively. The choice of length to trim can be adjusted but was set as 450 here to increase the chances of removing all erroneous regions of contigs flanking adapter sequences.

Reassembly of contigs and calculation of N50

To reassemble trimmed contigs after the removal of adapter contamination at the ends of contigs, we employed CAP3[17] using the following settings: -z 1 -y 6 -f 2 -p 99. These choices enabled the stitching together of the ends of contigs with perfectly overlapping and identical regions. N50 was calculated using custom bash scripts available at github.com/CUMoellerLab/MalAdapter.

Declarations

Acknowledgements

We thank Howard Ochman for useful discussion and for suggesting the name 'MalAdapter'.

References

1. Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., … Smith, T. P. (2022). Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nature Biotechnology, *40*(5), 711–719.
2. Sanders, J. G., Yan, W., Mjungu, D., Lonsdorf, E. V., Hart, J. A., Sanz, C. M., … Moeller, A. H. (2022). A low-cost genomics workflow enables isolate screening and strain-level analyses within microbiomes. Genome Biology, *23*(1), 212.
3. Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., … Finn, R. D. (2020). MGnify: the microbiome analysis resource in 2020. Nucleic Acids Research, *48*(D1), D570-D578.
4. Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., … Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. Nature

Biotechnology, *39*(1), 105–114.

5. Beresford-Jones, B. S., Forster, S. C., Stares, M. D., Notley, G., Viciani, E., Browne, H. P., … Pedicord, V. A. (2022). The Mouse Gastrointestinal Bacteria Catalogue enables translation between the mouse and human gut microbiotas via functional mapping. Cell Host & Microbe, *30*(1), 124–138.
6. Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., … Watson, M. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nature Communications, *9*(1), 870.
7. Chen, C., Zhou, Y., Fu, H., Xiong, X., Fang, S., Jiang, H., … Huang, L. (2021). Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. Nature Communications, *12*(1), 1106.
8. Glendinning, L., Stewart, R. D., Pallen, M. J., Watson, K. A., & Watson, M. (2020). Assembly of hundreds of novel bacterial genomes from the chicken caecum. Genome Biology, *21*(1), 1–16.
9. Li, Y., Leonard, S. P., Powell, J. E., & Moran, N. A. (2022). Species divergence in gut-restricted bacteria of social bees. *Proceedings of the National Academy of Sciences*, *119*(18), e2115013119.
10. Pasolli, E., De Filippis, F., Mauriello, I. E., Cumbo, F., Walsh, A. M., Leech, J., … Ercolini, D. (2020). Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome. Nature Communications, *11*(1), 2610.
11. Sanders, J. G., Sprockett, D. D., Li, Y., Mjungu, D., Lonsdorf, E. V., Ndjango, J. B. N., … Moeller, A. H. (2023). Widespread extinctions of co-diversified primate gut bacterial symbionts from humans. Nature Microbiology, 1–12.
12. Shaiber, A., & Eren, A. M. (2019). Composite metagenome-assembled genomes reduce the quality of public genome repositories. mBio, *10*(3), 10–1128.
13. Orakov, A., Fullam, A., Coelho, L. P., Khedkar, S., Szklarczyk, D., Mende, D. R., … Bork, P. (2021). GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biology, *22*, 1–19.
14. Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., … Segata, N. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell, *176*(3), 649–662.
15. Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D. L., Sims, Y., … Wood, J. (2021). Significantly improving the quality of genome assemblies through curation. Gigascience, *10*(1), giaa153.
16. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, *17*(1), 10–12.
17. Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Research, *9*(9), 868–877.

Figures

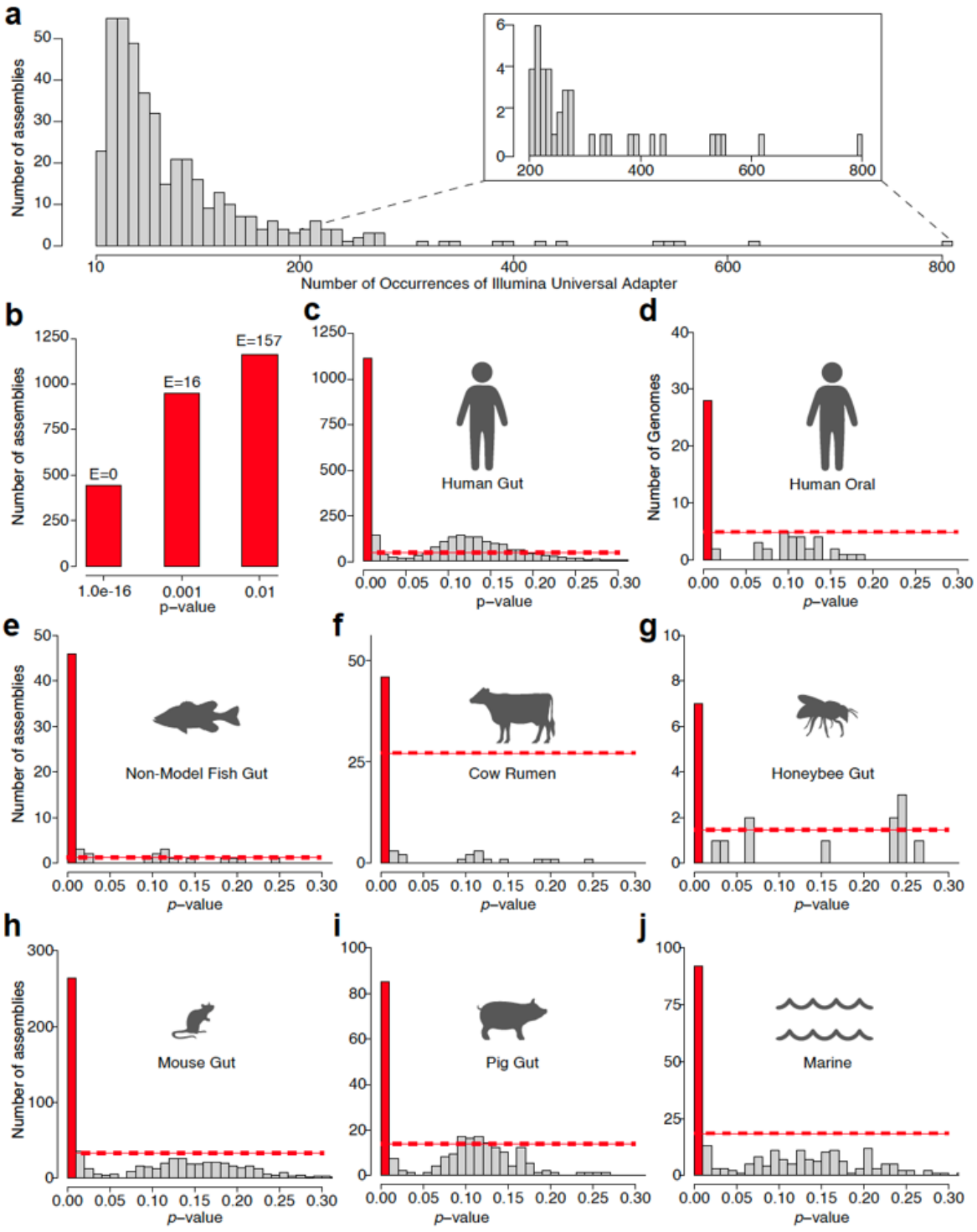

Figure 1

Significant enrichment of Illumina adapter sequences in published microbial genome databases. **(a)**Histogram shows the number of assemblies in all databases containing 10 or more exact matches to the Illumina universal adapter sequence or its reverse complement. Of the 15657 species reference genome assemblies, the number of assemblies expected to contain 10 or more exact matches by chance was ~1.57e-12, i.e., ~0. **(b)** Bar plot shows the number of assemblies displaying significant evidence of

adapter enrichment at three *p*-value thresholds. Expected number of assemblies (E) is shown for each threshold. **(c–j)** Histograms show the number of assemblies in individual databases for specific ranges of *p*-values. In **(c–j)**, Red bars indicate the number of assemblies for which *p*-values were <0.01. Dashed red lines indicate the number of assemblies expected to display *p*-values of <0.01 by chance (i.e., ~1% of assemblies in each database).

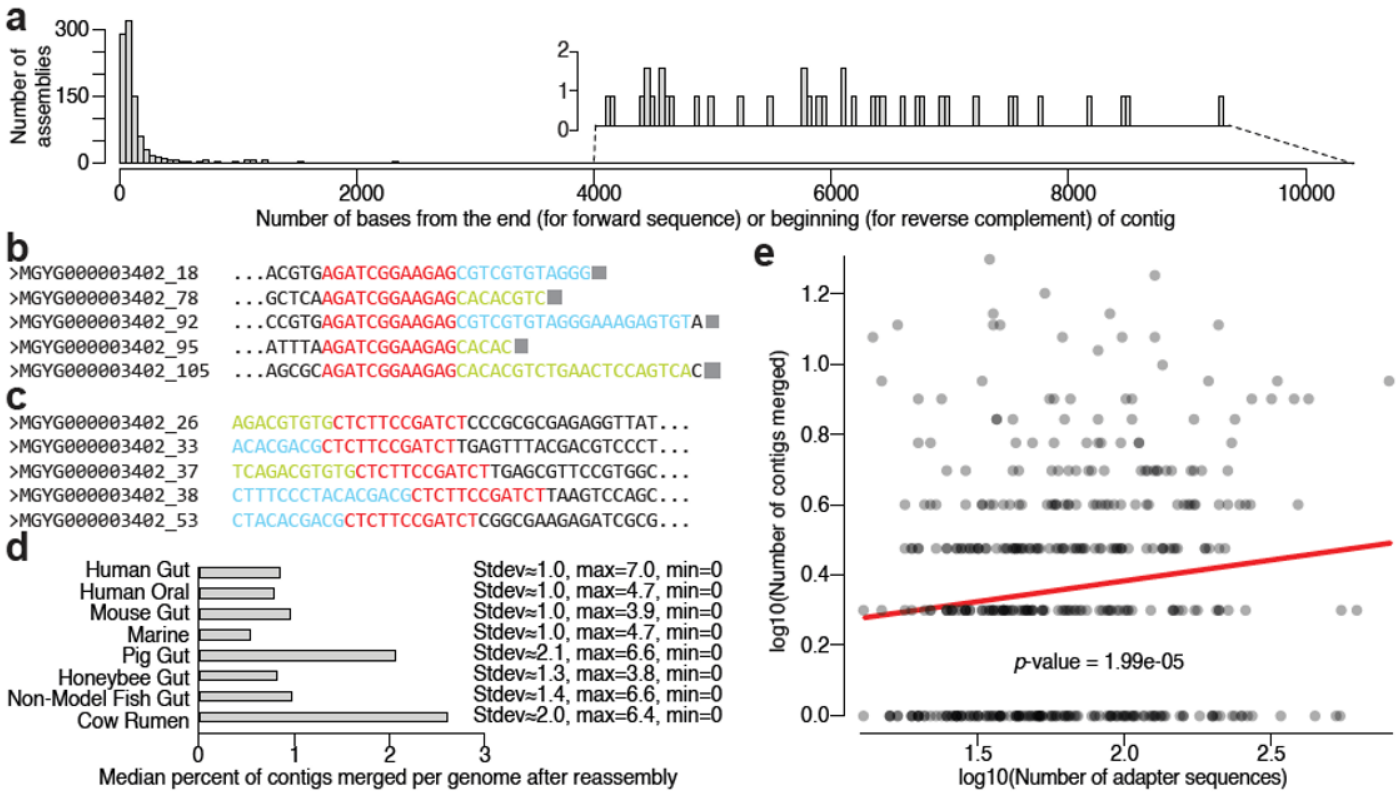

Figure 2

Adapter contamination is concentrated at the beginnings and ends of contigs, and its removal improves assembly contiguousness. (a) Histogram shows the concentration of Illumina universal adapter sequences near the extremities of contigs in the genomes showing significant evidence of adapter contamination (*p*-value < 0.01). Mean distances in bases from beginnings or ends of contigs were calculated for adapter sequences and reverse complements of adapter sequences, respectively. **(b)** DNA sequences show five examples of contamination by Illumina adapters (red sequences) at the ends of contigs (grey squares) in *Paenibacillus lactis* assembly MGYG000003402 from the human gut. **(c)** DNA sequences show five examples of contamination by the reverse complement of Illumina adapters (red sequences) at the beginnings of contigs in assembly MGYG000003402. In **(b)** and **(c)** blue and yellow sequences correspond to forward- and reverse-specific adapter sequences, respectively, adjacent to the universal adapter sequence. **(d)** Barplot shows for each database the per-assembly (of the 1110 contaminated assemblies at) average number of contigs merged with other contigs after the removal of adapter contamination and reassembly. Bars indicate standard deviations. **(e)** Scatterplot shows the

positive relationship between the number of adapter sequences present in assemblies showing the most significant evidence of contamination (*p*-value < 1e-16) (x-axis) and the number of contigs that were able to be merged by reassembly after adapter contamination removal (y-axis). Red line shows best-fit regression of log transformed values (transformation was made to reduce heteroscedasticity). The *p*-value was calculated from a generalized linear model with Poisson-distributed errors for count data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS112224.xlsx