applied microbiology

Communication

Evaluation of Genomic Contamination Detection Tools and Influence of Horizontal Gene Transfer on Their Efficiency through Contamination Simulations at Various Taxonomic Ranks

Luc Cornet [1,2,3,4,*], Valérian Lupo [2], Stéphane Declerck [2] and Denis Baurain [4]

[1] BCCM/IHEM, Mycology and Aerobiology, Sciensano, 1050 Brussels, Belgium
[2] BCCM/MUCL and Laboratory of Mycology, Earth and Life Institute, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium; valerian.lupo@uclouvain.be (V.L.); stephan.declerck@uclouvain.be (S.D.)
[3] BCCM/ULC, InBioS–Molecular Diversity and Ecology of Cyanobacteria, University of Liège, 4000 Liège, Belgium
[4] InBioS–PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, 4000 Liège, Belgium; denis.baurain@uliege.be
* Correspondence: luc.cornet@uliege.be

Abstract: Genomic contamination remains a pervasive challenge in (meta)genomics, prompting the development of numerous detection tools. Despite the attention that this issue has attracted, a comprehensive comparison of the available tools is absent from the literature. Furthermore, the potential effect of horizontal gene transfer on the detection of genomic contamination has been little studied. In this study, we evaluated the efficiency of detection of six widely used contamination detection tools. To this end, we developed a simulation framework using orthologous group inference as a robust basis for the simulation of contamination. Additionally, we implemented a variable mutation rate to simulate horizontal transfer. Our simulations covered six distinct taxonomic ranks, ranging from phylum to species. The evaluation of contamination levels revealed the suboptimal precision of the tools, attributed to significant cases of both over-detection and under-detection, particularly at the genus and species levels. Notably, only so-called "redundant" contamination was reliably estimated. Our findings underscore the necessity of employing a combination of tools, including Kraken2, for accurate contamination level assessment. We also demonstrate that none of the assayed tools confused contamination and horizontal gene transfer. Finally, we release CRACOT, a freely accessible contamination simulation framework, which holds promise in evaluating the efficacy of future algorithms.

Keywords: genomic contamination; contaminant levels; contamination simulations; horizontal gene transfer simulations; metagenomics

check for updates

Citation: Cornet, L.; Lupo, V.; Declerck, S.; Baurain, D. Evaluation of Genomic Contamination Detection Tools and Influence of Horizontal Gene Transfer on Their Efficiency through Contamination Simulations at Various Taxonomic Ranks. *Appl. Microbiol.* **2024**, *4*, 124–132. https://doi.org/10.3390/applmicrobiol4010009

Academic Editor: Ian Connerton

Received: 11 December 2023
Revised: 7 January 2024
Accepted: 9 January 2024
Published: 10 January 2024

1. Introduction

Genomic contamination is a well-known, albeit recurrent, problem in genomics. It appears when a genome—often, but not always, a Metagenome-Assembled Genome (MAG)—contains DNA sequences that do not belong to the expected organism. This umbrella concept actually masks different sources of DNA mis-affiliation [1]. Hence, contaminant sequences can arise from three sources, namely biological, experimental, and computational issues. Nevertheless, regardless of these origins, at the sequence level, genomic contamination can be categorized into three main types, redundant, replaced, and single, with the last two being non-redundant contaminations [1]. Besides sequences artifactually introducing chimerism into genomes, there exists a naturally occurring form of "contamination", i.e., horizontal gene transfer (HGT), when two bacteria exchange genetic material (via transduction, transformation, or conjugation) without being descendants of one other. Genomes are the basis of numerous studies, and it is no longer necessary to

demonstrate that genomic contamination is a cause for errors, notably in phylogenomic inference [2–4]. Consequently, the detection of contaminants is a topic that has attracted the attention of scientists, with the development of numerous detection tools and an increasing rate of publication in recent years. Although all these tools ultimately report a quantified level of contamination, they are based on various algorithms and do not measure the same information [1,5]. Indeed, among the most popular tools, two major categories can be distinguished: those relying on the presence of multiple marker genes (e.g., CheckM [6] and BUSCO [7]) and those based on whole-genome surveys (e.g., GUNC [8], Physeter [5], and Kraken2 [9]). Because of these differences in algorithms, Cornet et al. (2018) [10] and Lupo et al. (2021) [5] have reported the difficulty of meaningfully comparing these tools, as well as computing statistically sound correlations between their estimates. Furthermore, with the notable exception of the recently released GUNC, the impact of HGT on the detection of genomic contamination has never been assessed. In the present study, we utilize orthologous gene inference (homologous genes that have undergone a speciation event; Walter and Fitch 1970 [11]) to define shared orthologous groups (sets of genes orthologous to each other) among organisms for the simulation of contamination events and horizontal transfer by orthologous gene insertion and/or deletion. We then compare the detection performance of six of the most used tools (CheckM [6], BUSCO [7], GUNC [8], Physeter [5], Kraken2 [9], and CheckM2 [12]) in order to assess their efficiency. To do so, we use simulations at multiple taxonomic ranks, while varying the contamination scenarios. In all cases, we know the exact amount of contaminant sequences introduced in the simulated genomic sequences.

2. Materials and Methods

2.1. Contamination Simulations (Overview of CRACOT)

The simulations were performed with the newly developed Nextflow script CRACOT, standing for "Critical Assessment of Genomic Contamination Detection at Several Taxonomic Ranks", freely available at https://github.com/Lcornet/GENERA/wiki/20.-CRACOT (accessed on 6 January 2024).

Seven hundred and five (705) high-quality genomes belonging either to class *Clostridia* (e.g., *Clostridium)* or class *Bacilli* (e.g., *Lactobacillus*) were selected as input for CRACOT. These genomes were selected based on the GUNC [8] clade separation score (CSS), which measures the chimerism of genome contigs. Furthermore, we required these genomes to have no more than five contigs and no "N" within contigs. The contamination values of these genomes for the six tools are available in Table S1. The median contamination level was 0.02% for GUNC V1.0.5 [8], 0.45% for CheckM V1.2.1 [6], 0.87% for BUSCO V5.4.3 [7], 2.45% for Kraken2 V2.1.2 [9], 8% for CheckM2 V0.1.3 [12], and 22.3% for Physeter V0.213470 [5]. These contamination levels reflected the initial state of contamination of the input genomes.

The first step of CRACOT, as shown in Figure 1, was to create random genome pairs, one genome being considered hereafter as the main "expected" organism and the second as the slave "contaminant" organism. The pairing, based on the NCBI Taxonomy [13,14] associated with the genomes, handled through Bio-MUST-Core V0212670 (https://metacpan.org/dist/Bio-MUST-Core, accessed on 6 January 2024), was achieved for one specific taxonomic rank, ranging from phylum to species. For a given rank, the two genomes should belong to the same taxon at this rank but have a different taxonomy starting with the next (lower) rank. For instance, analysis at the phylum rank (e.g., *Firmicutes*) implies that the two genomes to be mixed indeed belong to the same phylum but are not part of the same class (e.g., if one genome belongs to Bacilli, the other genome belongs to Clostridia).

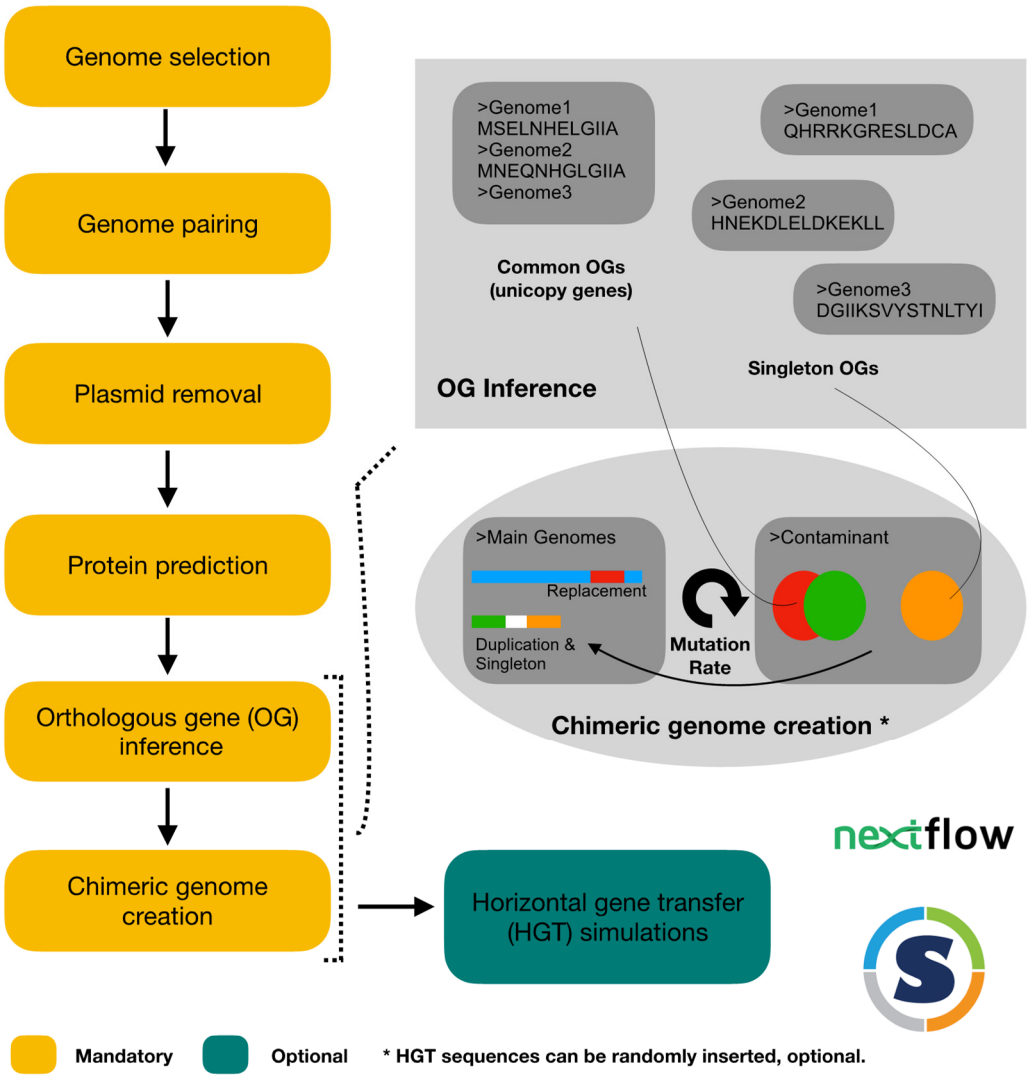

Figure 1. Flowchart of CRACOT. CRACOT is a Nextflow workflow, supported by a Singularity container. It is a six-step program. The first step is genome selection according to a user-specified list. The second step is the association of genomes, by pairs of the same taxonomic group. Steps 3 to 5 correspond to the removal of plasmids, protein prediction, and orthology inference. Finally, genome contamination simulations are based on the information produced during the orthology inference step, with common genes being used for "redundant" and "replaced" contamination events, while singletons are used for "single" contamination events. Optionally, a determined mutation rate can be enabled for each of these three basic event types to simulate horizontal gene transfer.

The plasmids of the selected genomes were removed after the pairing step, to not Interfere with the detection of contamination. Removal was performed with PlasmidPicker (https://github.com/haradama/PlasmidPicker, accessed on 6 January 2024), run with default settings. Proteins were then predicted with Prodigal V2.6.3 [15], used with default settings. Finally, OrthoFinder V2.5.4 [16], run with default settings, was used for orthology inference.

The three types of contamination were simulated based on the common and singleton protein orthogroups (oGs). Common proteins were defined as proteins present in only one copy for both the main and the slave genome in the OG, while single proteins were singletons of the slave genome. "Duplicated" contamination events were fished from the pool of common oGs, and the corresponding gene sequences of the slave genome were added to the end of the last contig of the main genome. Among the six tools analyzed, three (CheckM [6], CheckM2 [12], BUSCO [7], GUNC [8]) rely on Prodigal [15] for protein

prediction, and the addition of the duplicated genes at the end of the contigs had no influence on Prodigal as the prediction start codons are also translocated at the end of the contig. The two genome-wide tools tested in this study (Kraken2 and Physeter) classify sequences based either on the long kmers of 31 nucleotides (Kraken2) or on the pseudo-reads of 150 nucleotides created with a sliding window (Physeter), which makes the classification independent of the genomic position of the duplicated genes. "Replaced" contamination events were also fished from the pool of common oGs but slave genes replaced the genuine genes within the main genome. "Single" contamination events were fished from the pool of singletons of the slave organism, and the corresponding gene sequences were added to the end of the last contig of the main genome, as above. The number of events of each type is a user-specified option. For each simulation, 150 chimeric genomes were specified as CRACOT output, but the real output number depended on the number of available common and single-protein OGs. The numbers of simulated genomes used in this study are given in Table S2, while the chimeric levels of the simulations are indicated in Table S3. CRACOT was employed to simulate contamination events, encompassing not only the "redundant", "replaced", or "single" types individually but also their combined occurrence in more complex scenarios, allowing for the presence of the three types of contamination simultaneously.

To estimate the impact of HGT on genomic contamination detection, we added an optional mutation rate to our simulations to mimic real HGT. The idea is that HGT is akin to an old contamination that would have diverged from the orthologous sequence still present in the donor organism. In this context, varying the mutation rate is a way to simulate horizontal transfer events of different ages. HGT can be simulated for each of the three contamination types. Mutations were simulated at a given rate with HgtSIM [17], with the rate option set at 1-0-1-1, so that a mutation rate in DNA sequences corresponds to the same simulation rate in the proteins [17]. Hence, two HGT simulations for a combination of the three contamination types, with a mutation rate of either 10% or 25%, were generated.

2.2. Genomic Contamination Estimation

Genomic contaminants were estimated using the Nextflow workflow GENcontams (https://github.com/Lcornet/GENERA/wiki/09.-Genome-quality-assessment, accessed on 6 January 2024) from the GENERA project [18]. CheckM V1.2.1 [6] was used with the "lineage_wf" option and the provided database. GUNC V1.0.5 [8] was used with default settings and the Progenomes 2.1 database [19]. BUSCO V5.4.3 [7] was used in "auto-lineage" mode and the provided database. BUSCO's number of duplicated markers was used as a proxy for the contamination level. Physeter V0.213470 was used with the "auto-detect" option and the database provided in Lupo et al. (2021) [5]. Kraken2 V2.1.2 [9] was used with default settings and the database "PlusFP", downloaded from https://benlangmead.github.io/aws-indexes/k2 (accessed on 6 January 2024). Kraken2 levels of contamination were computed with the Physeter parser with the "auto-detect" option set to "count_first". The list of taxa used by the Physeter parser was automatically produced by the create-labeler.pl script using the list of genera found in the "nodes.dmp" file from the local mirror of NCBI Taxonomy. CheckM2 V0.1.3 [12] was used with default settings and the provided database.

2.3. Correlation and Violin Plot Creation

Spearman correlations between the contamination level estimates of the tools and the simulated levels of contaminants, as created by CRACOT, were computed with R [20]. Violin plots were created with ggplot [21]. The R code for the creation of these plots is available at https://github.com/Lcornet/GENERA/blob/main/Supplemental-scripts/CRACOT.R (accessed on 6 January 2024).

3. Results and Discussion

Regardless of the contaminant source, it is established that it can be summarized into three main types at the genomic sequence level (Figure 1) [1,8]. The first type is redundant contamination, which occurs when the contaminant sequence is redundant with a homologous genomic sequence of the expected organism [1]. The second type is replaced contamination, which is similar to the first one, but with the genuine sequence of the expected organism lacking from its genome [1]. The third type is single contamination, which occurs when the contaminant sequence has naturally no homologous sequence within the genome of the expected organism [1]. To mimic these three situations, we selected 705 high-quality reference genomes belonging either to class *Clostridia* (e.g., *Clostridium*) or class *Bacilli* (e.g., *Lactobacillus*) and simulated contamination events of the three types (Figure 1). The contamination detection discussed below depends on the initial contamination state of the 705 genomes used. Deviations from the simulations are minimal for four tools, with a low median initial contamination level of 0.02% for GUNC, 0.45% for CheckM, 0.87% for BUSCO, and 2.45% for Kraken2. However, they are higher for the last two tools tested, with 8% for CheckM2 and 22.3% for Physeter. Cornet et al., 2018 [10] and Lupo et al., 2021 [5] have shown that it is complicated to compute meaningful correlations of the contamination levels between different tools. These differences in the initial state of contamination in very high-quality genomes are therefore not surprising. Unfortunately, it turns out that to obtain enough genomes for simulations, it is necessary to incorporate genomes with a non-zero contamination level for at least one or two of the tools. Our simulations were performed at six different taxonomic ranks, from intra-phylum to intra-species. The rationale behind this strategy is twofold. Firstly, from a contamination perspective, especially for their in silico origin, it is easier to confuse sequences from a lower taxonomic rank, and, conversely, it is more challenging to distinguish them at the contamination detection level.

Surprisingly, our results reveal that, except for Kraken2, none of the tested tools was able to accurately estimate the contamination level (CL) of our combined scenarios, when the three different contamination types were mixed (Figure 2). Separate simulations are available in the Supplementary Materials for "redundant" (Figure S1), "replaced" (Figure S2), and "single" (Figure S3) events. CheckM, based on the duplication of gene markers [6], overestimated the redundant CL (Figure S1), but, quite logically, did not detect replaced (Figure S2) or single (Figure S3) contamination events. Like its main metric, CheckM's complementary metric used for genetically close contaminants ("strain heterogeneity") also overestimated CL, but at the genus and species ranks (Figure 2). BUSCO, which is also based on marker duplication [7], largely overestimated the redundant CL (Figure S1) at all ranks and, as for CheckM, under-detected replaced (Figure S2) and single (Figure S3) contamination events. GUNC, which searches for sequence chimerism [8], presented a pattern of both over- and underestimation at four ranks (phylum, class, order, and family) (Figure 2), with a minimum of 59% of underestimation (see Table S4 for the percentage of underestimation of each tool at each taxonomic rank). At the genus and species ranks, GUNC only underestimated CL (Figure 2), notably for replaced events, where it detected nothing (Figure S2). Physeter, which is based on the Lowest Common Inference (LCA) of DIAMOND blastx [22] hits [5], overestimates CL at all ranks for all types of contaminants (Figure 2). In contrast, Kraken2, which takes advantage of exact long k-mer matching [9], showed the best estimation of CL, fitting well to the simulations, except for the species rank, at which it was largely underestimated (see Table S4). It is noteworthy that the genomes used in our simulations were included in the Kraken2 database. Owing to its exact k-mer matching algorithm [23], one cannot exclude that Kraken2 would perform poorly on rare genomes, compared to our simulations. CheckM2, which uses a machine learning approach based on genomic contamination simulations (gradient boost model) without relying on taxonomic information [12], largely overestimated redundant CL (Figure S1), especially at the genus and species ranks. Replaced (Figure S2) and single (Figure S3) CL were underestimated at all ranks, apart from the single type at the genus and species ranks. The percentages of underestimation (Table S4) show that CheckM2 underestimated CL in

more than 97% of the cases for both replacement and single events, while it never under-detected the redundant type (with the exception of the species rank in 1.3% of the cases). To overcome the impossibility of directly correlating the performance of the different tools (due to their algorithmic differences [1]), we computed the correlation of each tool with the expected CL of our simulations. All the tools, not including Kraken2, correlated poorly, often negatively, with the simulated CL, with the correlation coefficient (R^2) never reaching beyond 0.37 (Figures 2 and S1–S3). All the tools used in this study, except for Kraken2, have the initial goal of calculating genomic contamination and informing researchers about the quality of the input genomes. The answer to this question should be similar for all algorithms because they all aim to detect the same phenomenon. CheckM and BUSCO have algorithms specifically created to estimate redundant contaminations and naturally disregard other types of contamination. The other four tools (GUNC, Kraken2, Physeter, and CheckM2) should detect all types of contamination, and, indeed, this is the case, but never precisely, except for Kraken2. GUNC exhibits both patterns of under- and over-detection, especially in non-redundant contaminations, which are under-detected by CheckM2. Physeter over-detects at all levels. It might be surprising that the best tool in these comparisons is Kraken2, which was not initially designed for contamination detection but for the classification of reads.

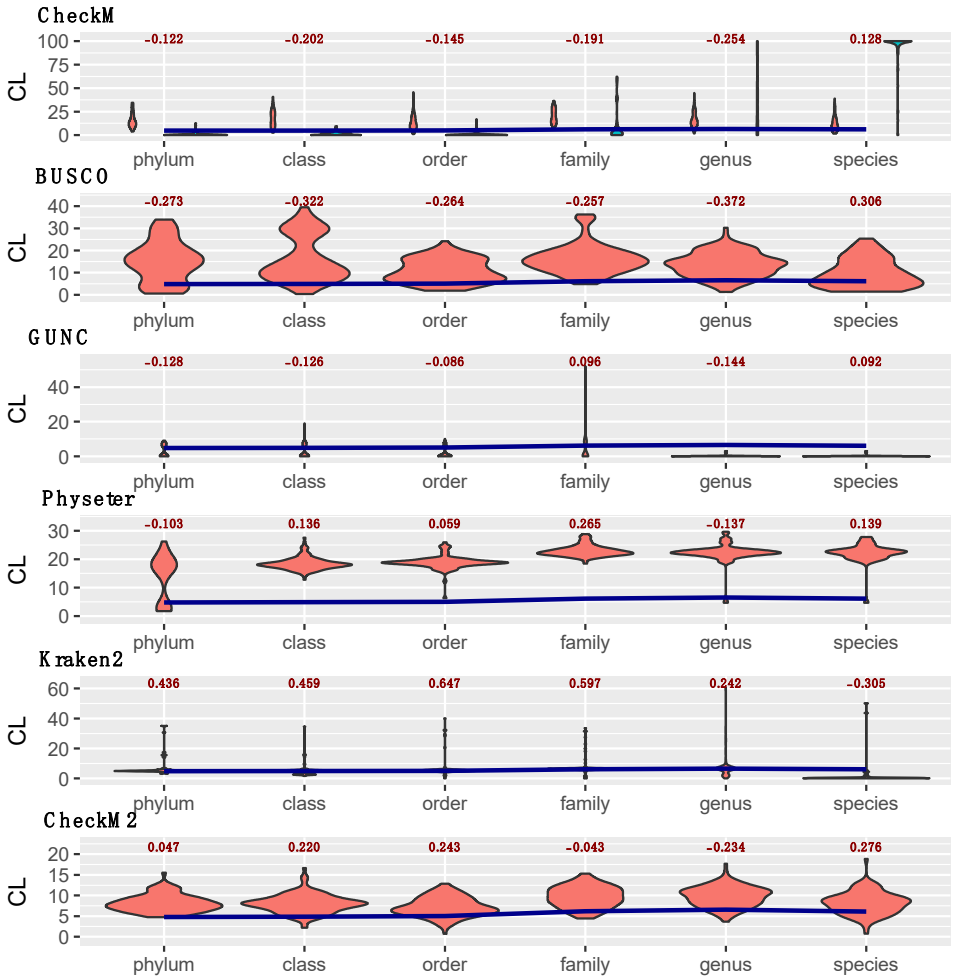

Figure 2. Contamination estimation, at six taxonomic ranks, of the combined types of contamination. Simulations were performed with a combination of the three contamination types (redundant, replaced, single). The median values of the contamination level (% CL) of these simulations are indicated by the blue line, while the CL estimated by the six tools are summarized by the violin plots. Spearman correlation values between the estimates of each tool and the simulated levels of contamination are indicated in red.

Besides genomic contamination, another type of genomic exchange naturally affects genomes: horizontal gene transfer (HGT). One of the major differences between HGT and contamination is that the first one accumulates mutations in the receiver (and donor) organisms [24] after transfer, whereas contamination occurs shortly before or after genome sequencing; hence, contaminant sequences are exact matches between donor and receiver genomes [1]. To investigate the effect of HGT on the detection performance, a non-null mutation rate was optionally enabled during the simulations (see Section 2), either at 10% (Figure S4) or 25% (Figure S5). None of the tools (with the exception of Physeter) confused contamination and HGT, which suggests that HGT events should not increase CL on real data. While reassuring, a possible drawback is that if the "contaminant" sequence is the result of a HGT, it has a low likelihood of being detected. This can be damaging since HGT frequently occurs in bacteria [25–29]. Somewhat ironically, the inability of Physeter to differentiate between HGT and genomic contamination indicates that LCA algorithms would be useful in such a case, even if probably too conservative due to their inclination towards over-detection.

4. Conclusions

We conducted this comparative study of contamination detection tools because no systematic benchmark, despite the availability of 18 programs, has been published to date, raising the question, "which tool should we use?" Our results show that CL is frequently overestimated, resulting in the unwarranted removal of sometimes precious (e.g., rare) genomes. Nevertheless, especially at the genus and species ranks, the odds of underestimation are always significant. This is a matter of concern because the risk of contamination by closely related taxa is higher when dealing with MAGs [30]. We also show that the replaced and single contamination types suffer less from underestimation compared to redundant events. The results of this study are quite surprising, as our simulations were rather simple. Furthermore, simulations were conducted with only one contaminant genome, at low CL, while contamination by more than one foreign taxon, at high CL, regularly occurs in public repositories [5,10]. Our conclusion is that, given the current algorithmic state of the field, which requires more innovation, users should use a combination of tools to estimate CL, and one of these tools should be Kraken2. Our contamination simulation framework, CRACOT, is freely available as a Nextflow workflow [31], sustained by a Singularity container [32], at https://github.com/Lcornet/GENERA/wiki/20.-CRACOT (accessed on 6 January 2024). It might be useful in future projects—for example, to estimate the accuracy of new tools under development.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/applmicrobiol4010009/s1, Figure S1: Contamination estimation, at six taxonomic ranks, of the redundant type of contamination; Figure S2: Contamination estimation, at six taxonomic ranks, of the replaced type of contamination; Figure S3: Contamination estimation, at six taxonomic ranks, of the single type of contamination; Figure S4: Contamination estimation, at six taxonomic ranks and with a mutation rate of 10%, of the combined types of contamination. Figure S5: Contamination estimation, at six taxonomic ranks and with a mutation rate of 25%, of the combined types of contamination. Table S1: Level of contamination estimation in reference genomes for the six tools; Table S2: Number of simulations used; Table S3: Chimeric levels of the simulations; Table S4: Under-detection of the tools.

Author Contributions: L.C. and D.B. conceived the study. L.C. developed CRACOT and performed all analyses. V.L. developed the parser of Kraken2, used for contamination level estimation. L.C. and D.B. wrote the manuscript with the help of V.L. and S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a research grant (no. B2/191/P2/BCCM GEN-ERA) financed by the Belgian State–Federal Public Planning Science Policy Office (BELSPO). Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI) funded by

the F.R.S.-FNRS (2.5020.11), and through two research grants to D.B.: B2/191/P2/BCCM GEN-ERA (Belgian Science Policy Office—BELSPO) and CDR J.0008.20 (F.R.S.-FNRS).

Data Availability Statement: CRACOT is freely available at https://github.com/Lcornet/GENERA/wiki/20.-CRACOT (accessed on 6 January 2024).

Conflicts of Interest: The authors declare no competing interests.

References

1. Cornet, L.; Baurain, D. Contamination Detection in Genomic Data: More Is Not Enough. *Genome Biol.* **2022**, *23*, 60. [CrossRef] [PubMed]
2. Schierwater, B.; Eitel, M.; Jakob, W.; Osigus, H.-J.; Hadrys, H.; Dellaporta, S.L.; Kolokotronis, S.-O.; DeSalle, R. Concatenated Analysis Sheds Light on Early Metazoan Evolution and Fuels a Modern "Urmetazoon" Hypothesis. *PLoS Biol.* **2009**, *7*, e20. [CrossRef] [PubMed]
3. Philippe, H.; Brinkmann, H.; Lavrov, D.V.; Littlewood, D.T.J.; Manuel, M.; Wörheide, G.; Baurain, D. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biol.* **2011**, *9*, e1000602. [CrossRef] [PubMed]
4. Laurin-Lemay, S.; Brinkmann, H.; Philippe, H. Origin of Land Plants Revisited in the Light of Sequence Contamination and Missing Data. *Curr. Biol.* **2012**, *22*, R593–R594. [CrossRef] [PubMed]
5. Lupo, V.; Van Vlierberghe, M.; Vanderschuren, H.; Kerff, F.; Baurain, D.; Cornet, L. Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. *Front. Microbiol.* **2021**, *12*, 755101. [CrossRef]
6. Parks, D.H.; Imelfort, M.; Skennerton, C.T.; Hugenholtz, P.; Tyson, G.W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Res.* **2015**, *25*, 1043–1055. [CrossRef]
7. Manni, M.; Berkeley, M.R.; Seppey, M.; Simao, F.A.; Zdobnov, E.M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *arXiv* **2021**, arXiv:2106.11799. [CrossRef]
8. Orakov, A.; Fullam, A.; Coelho, L.P.; Khedkar, S.; Szklarczyk, D.; Mende, D.R.; Schmidt, T.S.B.; Bork, P. GUNC: Detection of Chimerism and Contamination in Prokaryotic Genomes. *Genome Biol.* **2021**, *22*, 178. [CrossRef]
9. Wood, D.E.; Lu, J.; Langmead, B. Improved Metagenomic Analysis with Kraken 2. *Genome Biol.* **2019**, *20*, 257. [CrossRef]
10. Cornet, L.; Meunier, L.; Vlierberghe, M.V.; Léonard, R.R.; Durieu, B.; Lara, Y.; Misztak, A.; Sirjacobs, D.; Javaux, E.J.; Philippe, H.; et al. Consensus Assessment of the Contamination Level of Publicly Available Cyanobacterial Genomes. *PLoS ONE* **2018**, *13*, e0200323. [CrossRef]
11. Fitch, W.M. Distinguishing Homologous from Analogous Proteins. *Syst. Biol.* **1970**, *19*, 99–113. [CrossRef]
12. Chklovski, A.; Parks, D.H.; Woodcroft, B.J.; Tyson, G.W. CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning. *Nat. Methods* **2022**, *20*, 1203–1212. [CrossRef] [PubMed]
13. Federhen, S. The NCBI Taxonomy Database. *Nucleic Acids Res.* **2012**, *40*, D136–D143. [CrossRef] [PubMed]
14. Schoch, C.L.; Ciufo, S.; Domrachev, M.; Hotton, C.L.; Kannan, S.; Khovanskaya, R.; Leipe, D.; Mcveigh, R.; O'Neill, K.; Robbertse, B.; et al. NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools. *Database* **2020**, *2020*, baaa062. [CrossRef] [PubMed]
15. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef]
16. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics. *Genome Biol.* **2019**, *20*, 238. [CrossRef]
17. Song, W.; Steensen, K.; Thomas, T. HgtSIM: A Simulator for Horizontal Gene Transfer (HGT) in Microbial Communities. *PeerJ* **2017**, *5*, e4015. [CrossRef]
18. Cornet, L.; Durieu, B.; Baert, F.; D'hooge, E.; Colignon, D.; Meunier, L.; Lupo, V.; Cleenwerck, I.; Daniel, H.-M.; Rigouts, L.; et al. The GEN-ERA Toolbox: Unified and Reproducible Workflows for Research in Microbial Genomics. *GigaScience* **2023**, *12*, giad022. [CrossRef]
19. Mende, D.R.; Letunic, I.; Maistrenko, O.M.; Schmidt, T.S.B.; Milanese, A.; Paoli, L.; Hernández-Plaza, A.; Orakov, A.N.; Forslund, S.K.; Sunagawa, S.; et al. proGenomes2: An Improved Database for Accurate and Consistent Habitat, Taxonomic and Functional Annotations of Prokaryotic Genomes. *Nucleic Acids Res.* **2020**, *48*, D621–D625. [CrossRef]
20. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2014.
21. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016; ISBN 978-3-319-24277-4.
22. Buchfink, B.; Xie, C.; Huson, D.H. Fast and Sensitive Protein Alignment Using DIAMOND. *Nat. Methods* **2015**, *12*, 59–60. [CrossRef]
23. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments. *Genome Biol.* **2014**, *15*, R46. [CrossRef] [PubMed]
24. Arnold, B.J.; Huang, I.-T.; Hanage, W.P. Horizontal Gene Transfer and Adaptive Evolution in Bacteria. *Nat. Rev. Microbiol.* **2021**, *20*, 206–218. [CrossRef] [PubMed]
25. Zhaxybayeva, O.; Gogarten, J.P.; Charlebois, R.L.; Doolittle, W.F.; Papke, R.T. Phylogenetic Analyses of Cyanobacterial Genomes: Quantification of Horizontal Gene Transfer Events. *Genome Res.* **2006**, *16*, 1099–1108. [CrossRef]

26. Dagan, T.; Artzy-Randrup, Y.; Martin, W. Modular Networks and Cumulative Impact of Lateral Transfer in Prokaryote Genome Evolution. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 10039–10044. [CrossRef] [PubMed]
27. Dagan, T.; Martin, W. Ancestral Genome Sizes Specify the Minimum Rate of Lateral Gene Transfer during Prokaryote Evolution. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 870–875. [CrossRef]
28. Bohr, L.L.; Mortimer, T.D.; Pepperell, C.S. Lateral Gene Transfer Shapes Diversity of *Gardnerella* spp. *Front. Cell. Infect. Microbiol.* **2020**, *10*, 293. [CrossRef]
29. Frazão, N.; Sousa, A.; Lässig, M.; Gordo, I. Horizontal Gene Transfer Overrides Mutation in Escherichia Coli Colonizing the Mammalian Gut. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 17906–17915. [CrossRef]
30. Chen, L.-X.; Anantharaman, K.; Shaiber, A.; Eren, A.M.; Banfield, J.F. Accurate and Complete Genomes from Metagenomes. *Genome Res.* **2020**, *30*, 315–333. [CrossRef]
31. Di Tommaso, P.; Chatzou, M.; Floden, E.W.; Barja, P.P.; Palumbo, E.; Notredame, C. Nextflow Enables Reproducible Computational Workflows. *Nat. Biotechnol.* **2017**, *35*, 316–319. [CrossRef]
32. Kurtzer, G.M.; Sochat, V.; Bauer, M.W. Singularity: Scientific Containers for Mobility of Compute. *PLoS ONE* **2017**, *12*, e0177459. [CrossRef]