Steinegger and Salzberg *Genome Biology* (2020) 21:115
https://doi.org/10.1186/s13059-020-02023-1

Genome Biology

METHOD

Open Access

Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank

Martin Steinegger[1,2,3*] and Steven L. Salzberg[2,4,5]

*Correspondence:
martin.steinegger@snu.ac.kr
[1]School of Biological Sciences, Seoul National University, Seoul, 08826, South Korea
[2]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, 21218 Baltimore, Maryland, USA
Full list of author information is available at the end of the article

Abstract

Genomic analyses are sensitive to contamination in public databases caused by incorrectly labeled reference sequences. Here, we describe Conterminator, an efficient method to detect and remove incorrectly labeled sequences by an exhaustive all-against-all sequence comparison. Our analysis reports contamination of 2,161,746, 114,035, and 14,148 sequences in the RefSeq, GenBank, and NR databases, respectively, spanning the whole range from draft to "complete" model organism genomes. Our method scales linearly with input size and can process 3.3 TB in 12 days on a 32-core computer. Conterminator can help ensure the quality of reference databases. Source code (GPLv3): https://github.com/martin-steinegger/conterminator

Keywords: Genomes, Contamination, Software, RefSeq, GenBank

Introduction

The number of genomes in public and private repositories has been skyrocketing for at least the past decade, primarily due to the rapidly dropping costs of sequencing. The public genome database GenBank, which is regularly synchronized with the EMBL and DDBJ databases, has been doubling in size roughly every 18 months [1]. These genomics databases provide a vital worldwide resource that has been driving new findings in biotechnology and medicine for nearly three decades.

Draft genomes consisting of hundreds to thousands of unordered DNA sequence fragments represent a large fraction of the over 500,000 genomes stored in GenBank [2]. Some of these fragments contain foreign DNA due to contamination from reagents, laboratory materials, sample processing artifacts, or cross-contamination from multiplexed sequencing runs (Fig. 1a). These contaminating sequences may cause a variety of problems, including incorrect labels on sequences in metagenomic studies [3], faulty conclusions about horizontal gene transfer [4, 5], or poor annotation quality of genomes [6].

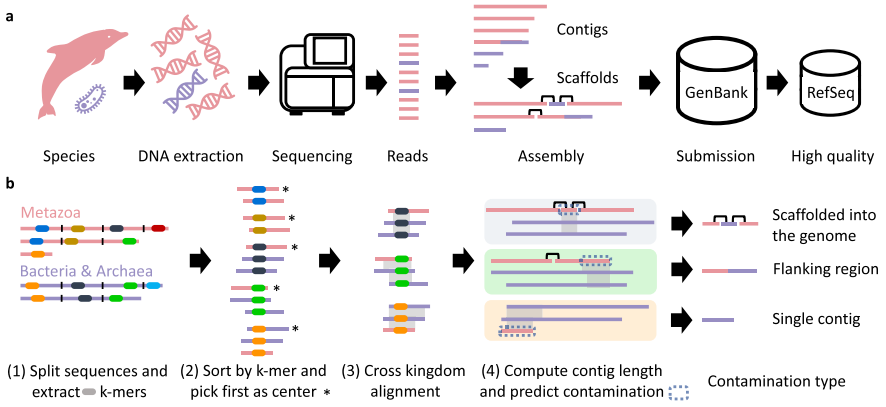

Fig. 1 How contamination occurs and how Conterminator detects it. **a** DNA extraction from an organism (red) is imperfect and often introduces contamination by other species (violet). DNA sequencing then generates short reads that are assembled into longer contigs. Contaminated DNA is typically assembled into separate, small contigs, but sometimes is erroneously included in the same contigs as DNA from the source organism. Contigs may also be linked by scaffolding, which can produce scaffolds containing a mixture of different species. Final assemblies are submitted to GenBank, and higher-quality assemblies are entered in RefSeq. **b** Conterminator detects contamination in proteins and nucleotide sequences across kingdoms, e.g., bacterial contaminants in plant genomes. The following describes the nucleotide contamination detection workflow. (1) We take taxonomically labeled input sequences and cut them into non-overlapping segments of length 1000 and extract a subset of *k*-mers. (2) We group the *k*-mers by sorting them and compute ungapped alignments between the first and all succeeding sequences per group. (3) We extract each region of the first sequence that has an alignment to other kingdoms that is longer than 100 nucleotides with a sequence identity greater than 90 %. We perform an exhaustive alignment of the input sequence segments against the multi-kingdom regions. (4) We reconstruct contig lengths within scaffolds by searching for the scaffold breakpoints (indicated by `N` characters in the DNA sequence) on the left and right side from the alignment start and end position. We predict that contamination is present if an alignment hits a contig that is shorter than 20 kb that aligns to a different kingdom with an contig length longer than 20 kb

To combat the contamination issue, NCBI (the home of GenBank) applies two filtering protocols for detection of contaminated fragments. First, VecScreen [7] is used to detect synthetic sequences (vectors, adapters, linkers, primers, etc.), and second, BLAST [8] alignments against common contaminants identify a broader array of contaminating sequences. Despite these filters, contamination still occurs, and its detection remains challenging [9, 10].

Because humans are always present in sequencing labs, *Homo sapiens* continues to be a major source of contamination for genome projects. Contaminating pieces of human DNA occasionally remain in published genomes [11] despite automated searches. A recent study, for example, showed that thousands of human DNA fragments can be found in draft bacterial genomes and that many of these have been erroneously translated and annotated as proteins [10]. However, many other species [12–16] also cause contamination. Systematic approaches to detect contamination are limited by computational costs of comparing every submitted genome against all other known genomes. For example, a BLAST all-against-all comparison of the RefSeq database [17], which has a size of 1.5 Tb, would take $\approx$ 30,000 CPU years. Faster alignment methods such as Minimap2 [18] or Bowtie2 [19] will take less time, but will still suffer from the quadratic complexity of this comparison. Other fast methods such as Mash [20] and sourmash [21] can compare genomes more quickly, but are not suited for finding small contaminating sequencing within a larger genome.

We present Conterminator (Fig. 1b), a fast method for detecting contamination in nucleotide and protein databases by computing local alignments across taxonomic kingdoms. It utilizes the linear-time all-against-all comparison algorithm from Linclust [22] followed by exhaustive alignments using MMseqs2 [23]. This enables us to process huge nucleotide and protein sequence sets on a single server. We applied this method to quantify the current state of contamination in the nucleotide databases Genbank [1] and RefSeq [17], and in the comprehensive NR protein database [1].

Results

Figure 2 summarizes the contamination found by Conterminator in RefSeq (Fig. 2a, b) and GenBank (Fig. 2c, d). Processing the 1.5 and 3.3 TB in RefSeq and GenBank took 5 and 12 days on a single 32-core machine with 2 TB of main memory. Conterminator reported 114,035 and 2,161,746 contaminated sequences affecting 2767 and 6795 species in RefSeq and GenBank, respectively. Identifiers of the contaminated sequences are available

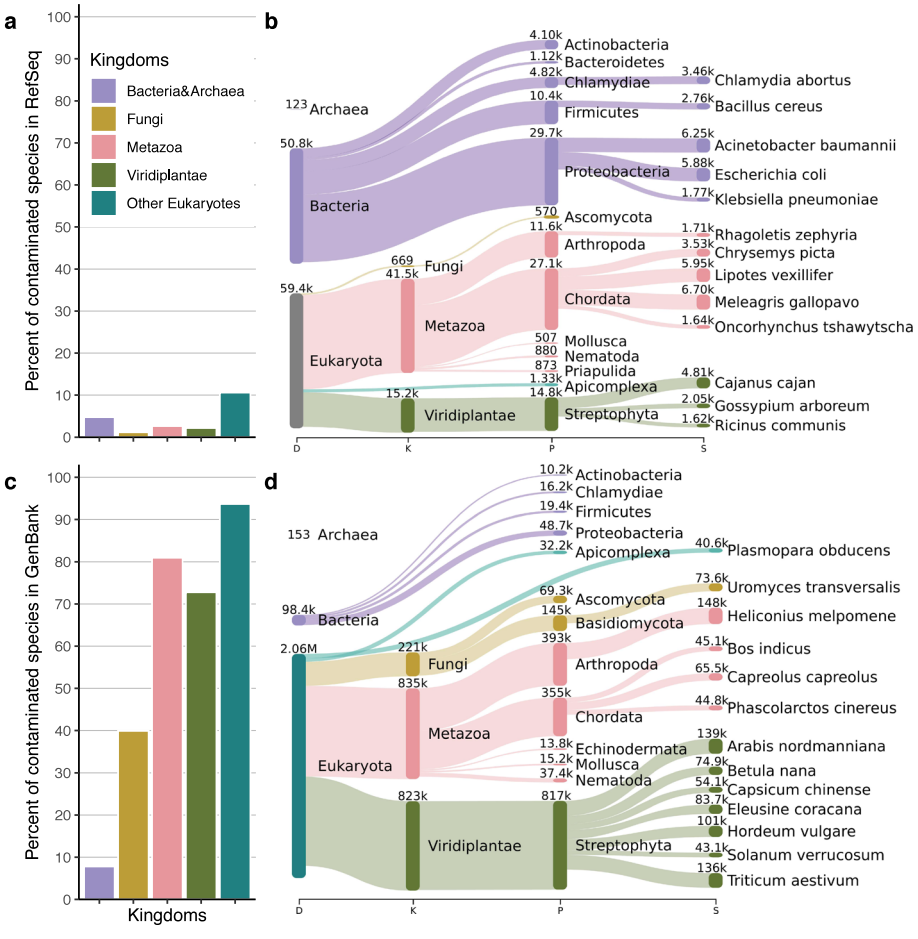

Fig. 2 Results of contamination within the RefSeq and GenBank. **a** Distribution of contaminated species in RefSeq across five kingdoms: Bacteria and Archaea (violet), Fungi (yellow), Metazoa (red), Viridiplantae (green) and other Eukaryotes (turquoise). **b** Sankey plot of the top 13 contaminated species in RefSeq. We show the taxonomic ranks domain, kingdom, phylum, and species. Numbers shown above each taxonomic node indicate the total number of contaminated sequences. The tree uses the same color code for kingdoms as in **a**. **c**, **d** Same as **a**, **b** but for GenBank

in Additional files 1 and 2: Listing S1. In GenBank, over 95 % of contamination occurred in eukaryotic genomes. Eukaryotic genomes tend to be much more fragmented due to their larger genome sizes and higher repetitive content (as compared to prokaryotes), and many of the smaller contigs in eukaryotic genome assemblies suffer from contamination.

In RefSeq, only 52 % of the contamination occurred in eukaryotic genomes. One likely reason for this is the more stringent filters used to determine which GenBank genomes are included in RefSeq; these filters reject genomes with very low contig sizes or genomes that were flagged as contaminated. The number of species identified as contaminants (i.e., the species causing contamination) in RefSeq was 2881, and in GenBank, the number was 13,981. The leading contaminant species are *Homo sapiens*, *Saccharomyces cerevisiae*, *Stenotrophomonas maltophilia*, and *Serratia marcescens* (see Additional file 3: Figure S1).

Contamination in high-quality genomes

We expected that well-studied model organisms would have the highest-quality genomes and that these genomes would have very little, if any, contamination. We also expected very little contamination in finished microbial genomes. Therefore, we created a control set of high-quality genomes consisting of 928 genomes from FDA-ARGOS, a curated set of complete microbial genomes [24], plus genomes for model organisms *Saccharomyces cerevisiae*, *Danio rerio*, *Mus musculus*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Homo sapiens*. We searched for (presumably) false positive predictions by scanning our RefSeq results for contaminants in these high-quality genomes. Initially, our method did not report any contamination for any of these genomes, in part because by default it only reports contamination when the target sequence is shorter than 20 kb (see the "Methods" section). We then considered alignments to sequences longer than 20 kb in this high-quality genome set. In this additional scan, we found alignments between bacterial sequences and two eukaryotes: (1) *Acidithiobacillus thiooxidans* in *Homo sapiens* and (2) *E. coli* in *Caenorhabditis elegans*, shown in Fig. 3.

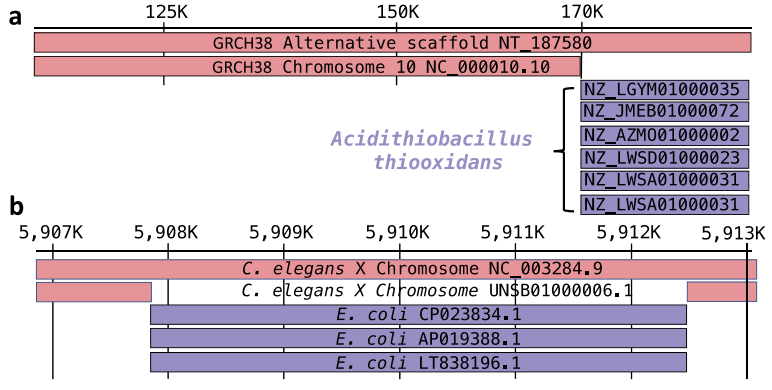

Fig. 3 Contamination in the reference genomes of *Homo Sapiens* and *Caenorhabditis elegans*. **a** Alignment of *Homo sapiens* alternative scaffold `NT_187580` of chromosome 10 against RefSeq. Chromosome 10 (`NC_000011.10`) aligns with 100 % sequence identity from position 1 to 169918. The remaining 18,397 residues of `NT_187580` align only to *Acidithiobacillus thiooxidans* at 98 % sequence identity. Shown are only 6 out of 15 alignments to *Acidithiobacillus thiooxidans*. **b** The X chromosome of *Caenorhabditis elegans* `NC_003284.9` aligns on the left and right flanking position around 5907856 until 5912458. *E. coli* genomes aligns from 5907856 to 5912087, a total of 4231 residues. Shown are only 3 out of 8199 alignments to *E. coli*

A. thiooxidans in human genomic sequence

The human reference genome (currently GRCh38) consists of chromosomal scaffolds, unplaced scaffolds, and "alternate" scaffolds. The last group is included in the reference genome to represent sequences that are divergent from the primary chromosome sequence. In `NT_187580`, an alternate scaffold on chromosome 10 in GRCh38.p13, we detected a sequence matching *Acidithiobacillus thiooxidans* that spans positions 169,917–188,315 of the human scaffold (Fig. 3a), which has a length of 188,315 bp. Fifteen different *A. thiooxidans* genomes align to the contaminated portion of the scaffold. The primary sequence of human chromosome 10 aligns perfectly from positions 1 to 169,918 on the alternate scaffold, but that alignment stops at the region that aligns to *A. thiooxidans*. Thus, the last ∼ 18 kb of this human alternative scaffold appears to be bacterial.

E. coli in the *C. elegans* reference genome

Our method also detected a bacterial contaminant in the *C. elegans* reference genome, in chromosome X (GenBank accession `NC_003284.9`). A segment spanning positions 5907856–5912087 of the *C. elegans* sequence aligns perfectly to multiple strains of *E. coli* (Fig. 3b). To check whether this might be a false positive reported by our method, we downloaded the raw Illumina reads used for a more-recent assembly of the same *C. elegans* strain (`SRR003808` and `SRR003809`) and aligned them against the chromosome X assembly (`NC_003284.9`) using Bowtie2 [19]. Only six reads (30 bp each) aligned in this region. In contrast, the average coverage over the rest of the chromosome was ∼ 99.8. This indicates that the *E. coli* sequence was indeed a contaminant. To corroborate the contamination further, we looked at a recent assembly of *C. elegans* that used a combination of long and short reads [25]. We aligned their assembly of chromosome X (GenBank accession UNSB01000006.1) against the current reference and found that in this newer assembly, the *E. coli* region is not present. This strongly suggests that the *C. elegans* reference genome contains a ∼ 4-kb insertion of *E. coli* contamination.

Meleagris gallopavo genome cleanup

The most contaminated genome in RefSeq, on our initial scan, was the turkey genome, *Meleagris gallopavo* [26]. The contaminants included 6698 small, unplaced scaffolds with a total size of 2,655,271 bases. More than half of the contaminations were caused by *Achromobacter xylosoxidans* and *Serratia marcescens*. We contacted the original authors of that assembly to communicate our findings, and they subsequently removed all contaminated fragments, plus an additional 39,413 contigs that were shorter than 300 bp. The new version of the assembly, `Turkey_5.1`, has no contaminants and is available in GenBank as accession `GCA_000146605.4`.

Proteins in contaminated RefSeq contigs

We detected that 19.4 % of the contaminated RefSeq contigs contain protein annotations and encode a total of 47,943 proteins. A previous study [10] reported 3437 spurious bacterial proteins that originate from human repeats that have contaminated bacterial genome assemblies. We aligned these sequences against our set using MMseqs2, enforcing a 80 % alignment coverage of the shorter sequence (`--comp-bias-corr 0 -mask 0 --cov-mode 5 -c 0.8`), and discovered that our set contains 62 % of the previously reported proteins.

We clustered the proteins using MMseqs2 at a 95 % sequence identity, enforcing a bi-directional coverage of 95 % (`cluster --min-seq-id 0.95 -c 0.95`). This resulted in 3339 clusters that covered 12,494 sequences. The remaining sequences were singleton clusters. The largest cluster consists of 185 bacterial proteins, all of which are located on contigs shorter than 1 kb, and the proteins are widely spread among multiple phyla in the bacterial kingdom. Despite the long evolutionary distance, 166 of the sequences are 100 % identical to each other and the remaining are at least 95 % identical, suggesting that all of them represent contaminants (see Fig. 4). All short bacterial contigs containing the 185 proteins align to multiple positions in the domestic sheep *Ovis aries* genome; the fragments align to chromosome 15 (`NC_040266.1`) with a sequence identity greater than 94 % with nearly complete coverage.

Contamination in the protein database NR

Conterminator can be used to analyze protein sequences. It clusters proteins [22] at 95 % sequence identity, while requiring at least 95 % sequence overlap. It reports clusters containing multiple kingdoms, using the same kingdom definition as for the nucleotide comparison. We predict that the kingdom with fewer members in the cluster is contaminated, e.g., if a cluster contains 100 proteins, and 99 represent animals while 1 represents bacteria, then the bacterial protein likely originates from a contaminated genome.

We analyzed the NCBI NR protein sequence [1] database using this procedure. We predicted 14,148 proteins to represent contaminants (Additional file 4: Listing S1), out of

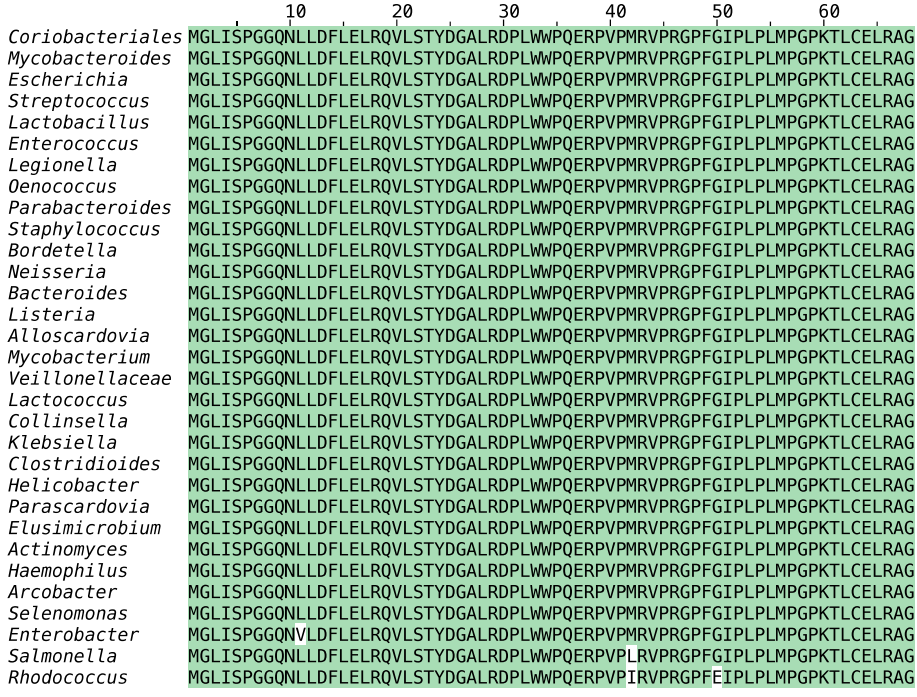

Fig. 4 Multiple sequence alignment of 31 spurious bacterial proteins encoded on short contaminated contigs. Shown here are 31 out of 185 spurious proteins from bacterial genomes. A majority of the sequences are 100 % identical. The only differing residues are highlighted in white. This highly conserved "protein" is conserved on across different bacterial phyla, suggesting it is likely a contaminant that has been erroneously translated as part of automated annotation procedures. The respective short contigs (< 1 kb) encoding these spurious proteins align with high sequence identity and coverage to the *Ovis aries* genome

which 7359 are also present in the Uniprot database [27]. The majority of these proteins (70.46 %) are eukaryotic, and the remaining 29.34 % are bacterial (see Additional file 3: Figure S2). Over 6114 contaminant proteins originate from the phylum Arthropoda, and out of this 2401 are from the species *Trichonephila clavipes*, the golden silk orb weaver spider, which contributes overall the most contamination. We next take a closer look at this organism.

Contaminated proteins in *Trichonephila clavipes*

The *T. clavipes* proteins identified as possible contaminants are distributed across 504 contigs with a total length of 24,152,567 residues. One hundred sixty-three of the contigs are longer than 20 kb. The longest contaminated contig from the assembly [28] is `MWRG01000001`, which spans 1,655,743 bases and encodes 490 proteins. We found 219 of these proteins in contaminated protein clusters, a majority of them matching the bacterium *Gemmobacter* sp. *YJ-T1-11*. We aligned the contig against the assembly of *Gemmobacter* sp. *YJ-T1-11* and found that it is nearly 90 % covered at a 97.76 % identity (see Additional file 3: Figure S3). Thus, this clearly appears to be a bacterial contig mistakenly included in the assembly of the *T. clavipes* spider.

Discussion

We present two complementary approaches to detect contamination, first using nucleotide information to detect short contaminated fragments and second using protein-based analysis to reveal long contaminated contigs. We used a conservative approach that only considered a sequence to be a contaminant if it had a near-identical match to a species in an entirely different kingdom from the source, e.g., a bacterial sequence found in an animal genome or vice versa. However, our software has a parameter (`--kingdom`) that allows one to identify contamination across phyla or other taxonomic levels as well. Our method can efficiently detect contamination in large reference databases, and we found that a substantial fraction of the genome sequences in both GenBank and RefSeq (0.54 % and 0.34 % of entries, respectively) appear to be contaminated. Contamination occurs mostly as short contigs, flanking regions on longer contigs, or regions of larger scaffolds flanked by `Ns`, but we also observed a few longer sequences with contamination.

Note that to simplify our analysis, we merged Bacteria and Archaea, which are two distinct kingdoms, into one group. Thus, we do not detect contamination between these two kingdoms. Also note that we excluded environmental (metagenomic) samples from our analysis. We excluded viruses because some of them can integrate their genomes into other organisms, making it hard to distinguish contamination from genuine artifacts of viral integration.

Contamination can be transferred into other databases that are built from GenBank, such as the protein databases NR and Uniprot. Methods that rely on taxonomical classification, particularly metagenomics analyses, are strongly affected by cross-kingdom contamination because they often rely on subsets, e.g., microbial sequences extracted from a larger database. This makes it more difficult for such methods to detect contamination of the type reported here.

With the rapid and ongoing increase in the number of novel genomes sequenced every year, the number and variety of contaminating sequences continue to increase as well,

presenting challenges for alignment-based methods to detect contamination. Conterminator's efficiency means that it can be used routinely to detect new contamination, even on the largest databases.

Methods

Conterminator detects cross-kingdom alignments and predicts contamination. It builds upon existing modules of MMseqs2 [23], which it extends for use in contamination detection.

Detection of cross-kingdom alignments

Conterminator identifies regions in genome sequences that align to genomes from other kingdoms with a minimum length of at least 100 nucleotides and a sequence identity threshold of at least 90 %. With very few exceptions, DNA sequences from different kingdoms should not be aligned at all, and sequences that match at this level of identity are strong candidates for contaminants. Exceptions to this rule include recent horizontal gene transfer events, but these are very rare.

Because modern sequence databases are very large, we cannot use a naive all-against-all alignment, which would entail a quadratic number of comparisons. Therefore, we used a similar strategy to Linclust [22] to reduce the computational cost to a linear number of comparisons. We reworked the algorithm to support nucleotide sequences, since Linclust was originally built to cluster protein sequences.

Conterminator first cuts all sequences into fragments of length 1000 and records their start positions. For each fragment, we extract m canonical k-mers (default $m = 100$) of length 24 with the lowest hash value and write them into an array. (We use the hash function defined in [22]; see Supplementary Figure 5.) We store the k-mer in 8 bytes, with the most significant bit indicating whether the k-mer is reversed, sequence identifiers (4 bytes), its length (2 bytes), and its position j in the genomic sequence (2 bytes). We sort the array by k-mer, length, and sequence identifiers. For each k-mer group, we assign all sequences to the longest sequence with the lowest sequence identifier c by overwriting their k-mer with the identifier of c and their position with the diagonal $i - j$ respecting the strand directionality. We sort the array again with the previous criteria so that all sequences with same assignment are in a consecutive block. We write each block's central sequence identifier, assigned sequence, strand, and diagonal to hard disk while only keeping the diagonal with the most k-mer matches per sequence.

We perform a one-dimensional dynamic programming ungapped alignment (using the MMseqs2 command "`rescorediagonal --rescore-mode 2`") on each diagonal. We assign matches a score of 2 and mismatches a score of -3 bits and compute an E value using ALP [29]. We compute the sequence identity by dividing the number of identical positions by the number of aligned positions. We filter out all hits that are shorter than 100 bases, or that have a sequence identity below 90 %, or that have an E value above 10^{-3}. We compute the alignment start positions by adding the start position of the fragment to the alignment coordinates (MMseqs2 command "offsetalignment").

Based on the alignment, we extract the sequence intervals from c that are overlapped by different kingdoms. We define five "kingdoms" based on the NCBI Taxonomy [30]: (1) Bacteria and Archaea (taxonomy IDs 2 and 2157), (2) Fungi (4751), (3) Metazoa (33208), (4) Viridiplantae (33090), and (5) all other eukaryotes. We ignored sequences

from Viruses (10239), unclassified sequences (12908), other sequences (28384), artificial sequences (81077), and environmental samples from bacteria, archaea, and eukaryotes (61964, 48479, 48510).

Gather all alignments by exhaustive alignment

Our detection method might miss alignments because it does not extract all *k*-mers and because it uses 24-mers rather than shorter *k*-mers. After the previous steps, we perform an exhaustive alignment of the sequence fragments against the extracted potential contamination sequences (and their respective reverse complements) using MMseqs2. The search is performed by using the two modules `prefilter` and `rescorediagonal`. The `prefilter` program masks out low complexity regions and short tandem repeats in the potential contaminants using `tantan` [31] and detects all consecutive double 15-mer diagonal hits. We rescore the detected diagonals again with the `rescorediagonal` module, enforcing a minimal alignment length of 100 and a minimal sequence identity of $\sqrt{0.9}$. The square root of the sequence identity ensures that no pair of sequences is greater than 90 % different from each other.

Predict contig length by finding scaffolding boundaries

Genome assembly programs create scaffolds by ordering and orienting contigs using a variety of types of linking information, such as paired-end reads. A scaffold thus consists of a sequence of contigs, usually separated (in many GenBank entries) by `N`s to indicate the scaffolding boundaries. Some of the contaminants that we identified appear as short contigs in the midst of a longer scaffold, and we can identify these by finding the flanking `N`s. It is important for our contamination detection to know the real length of each contig in a scaffold. A naïve approach to determining contig length would be to search for the closest `N` upstream and downstream from each alignment start and end. However, this is inefficient because many sequences contain million of bases without any `N`s. We therefore indexed all `N`s for each sequence. We store the position of the first `N` per block in an array associated with the sequence. The `N` positions are sorted in ascending order, which enables us to perform a binary search to detect the closest `N` efficiently.

Predict the source of contamination

A large majority of contamination occurs as small contigs (see Fig. 2 in Breitweiser et al. [10]). Conterminator uses this property to help it identify contamination based on the length distribution of sequences from each kingdom. By default, it only calls a sequence a contaminant if the sequence is shorter than 20 kb and if it aligns to a sequence in another kingdom that is longer than 20 kb. Note that in the rare cases where a contaminating sequence is longer than 20 kb, our method will fail to identify it. However, this prevents us from labeling recent horizontal gene transfer events as contamination.

Predicting contamination in protein databases

Conterminator can also detect protein sequence contamination using cross-kingdom analysis. It clusters proteins using Linclust [22] with a bidirectional length overlap of 95 % and a sequence identity of 95 % (`--min-seq-id 0.95 -c 0.95 --cov-mode 0 -a`). It reports every cluster with cross-kingdom members. For each contaminated cluster, it counts how often each kingdom occurs and reports the least abundant kingdom as

Table 1 List of software used in this paper

Resource	Version
Conterminator	fa1c80
MMseqs2 [23]	333546
KrakenUniq [32]	5c0019
Pavian [33]	81d784
RefSeq [17]	July 2019
GenBank [1]	Dec 2018
Jalview [34]	2.11.0

The versions for MMseqs2, KrakenUniq, and Pavian are the first 6 characters of the git commit. For databases, we list the date at which the data was downloaded

the one that is contaminated. It also reports kingdoms with equal abundance; however, in those cases, it cannot predict the contaminated entry. Using only abundance without concern for length may lead to incorrect directionality calls. For example, human repeats cause contamination in multiple bacterial genomes [2]. In this case, abundance-based directionality prediction would wrongly call the human genome to be contaminated.

Data visualization

We created the Sankey plots using the `krakenuniq-report` tool from KrakenUniq [32] to create a Kraken-style report from our predicted contaminations. The visualization was done using Pavian [33] extracted as SVG and colored by Inkscape. The multiple alignment was created by MMseqs2 `result2msa` and visualized using Jalview [34].

Software and database versions

Table 1 lists the softwares and their corresponding version.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13059-020-02023-1.

Additional file 1: List of contaminated RefSeq identifier.
Additional file 2: List of contaminated GenBank identifier.
Additional file 3: Supplementary materials. Contains Figures S1–S3 and Listing S1.
Additional file 4: List of contaminated NR identifier.
Additional file 5: Review history.

Peer review information
Barbara Cheifet was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Acknowledgements
We thank Milot Mirdita, Florian Breitwieser, and Jennifer Lu for fruitful discussions. We thank the NCBI for their answering our questions timely and Aleksey Zimin for cleaning up the Turkey genome.

Review history
The review history is available as Additional file 5.

Authors' contributions
MS and SS designed the research. MS developed the code and performed the analyses. MS and SS wrote the manuscript. Both authors read and approved the final manuscript.

Funding
This work was supported in part by NIH grants R35-GM130151 and R01-HG006677, and by NSF grant IOS-1744309 to SLS.

Availability of data and materials
Conterminator [35] is implemented in C++ and its open source licensed as GPLv3 and available at https://github.com/martin-steinegger/conterminator. The version to reproduce the results is available under https://doi.org/10.5281/zenodo.3750825 [36]. Commands to rerun the analysis of RefSeq and NR are in Additional file 3: Listing S1. The list of contamination for GenBank (`genbank.gz`), NR (`nr.gz`)), and RefSeq (`refseq.gz`) are available at: ftp://ftp.ccb.jhu.edu/pub/data/conterminator/ and https://figshare.com/projects/Conterminator/77346 [37].

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Biological Sciences, Seoul National University, Seoul, 08826, South Korea. [2]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, 21218 Baltimore, Maryland, USA. [3]Institute of Molecular Biology and Genetics, Seoul National University, Seoul, 08826, South Korea. [4]Department of Biomedical Engineering, Johns Hopkins University, 21218 Baltimore, Maryland, USA. [5]Departments of Computer Science and Biostatistics, Johns Hopkins University, 21218 Baltimore, Maryland, USA.

Received: 27 January 2020 Accepted: 16 April 2020
Published online: 12 May 2020

References
1. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2019;47(D1):94–99.
2. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. Brief Bioinform. 2019;20(4):1125–36.
3. Kirstahler P, Bjerrum SS, Friis-Møller A, la Cour M, Aarestrup FM, Westh H, Pamp SJ. Genomics-based identification of microorganisms in human ocular body fluid. Sci Rep. 2018;8(1):4126.
4. Arakawa K. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc Natl Acad Sci USA. 2016;113(22):3057.
5. Salzberg SL. Horizontal gene transfer is not a hallmark of the human genome. Genome Biol. 2017;18(1):85.
6. Poptsova MS, Gogarten JP. Using comparative genome analysis to identify problems in annotated microbial genomes. Microbiology. 2010;156(Pt 7):1909–17.
7. Schäffer AA, Nawrocki EP, Choi Y, Kitts PA, Karsch-Mizrachi I, McVeigh R. VecScreen_plus_taxonomy: imposing a tax(onomy) increase on vector contamination screening. Bioinformatics. 2018;34(5):755–9.
8. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.
9. De Simone G, Pasquadibisceglie A, Proietto R, Polticelli F, Aime S, JM Op den Camp H, Ascenzi P. Contaminations in (meta) genome data: an open issue for the scientific community. IUBMB Life. 2019;72:698–705.
10. Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. Genome Res. 2019;29(6):954–60.
11. Longo MS, O'Neill MJ, O'Neill RJ. Abundant human DNA contamination identified in non-primate genome databases. PLoS ONE. 2011;6(2):16410.
12. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ. 2014;2:675.
13. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS ONE. 2014;9(5):97876.
14. Orosz F. Two recently sequenced vertebrate genomes are contaminated with apicomplexan species of the sarcocystidae family. Int J Parasitol. 2015;45(13):871–8.
15. Mukherjee S, Huntemann M, Ivanova N, Kyrpides NC, Pati A. Large-scale contamination of microbial isolate genomes by illumina PhiX control. Stand Genomic Sci. 2015;10:18.
16. Reiter T, Titus Brown C. Microbial contamination in the genome of the domesticated olive. 2018. https://doi.org/10.1101/499541.
17. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):733–45.
18. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.
19. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9.
20. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132.

21. Pierce NT, Irber L, Reiter T, Brooks P, Brown CT. Large-scale sequence comparisons with *sourmash*. F1000Res. 2019;8:1006.
22. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9(1):2542.
23. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35(11):1026–8.
24. Sichtig H, Minogue T, Yan Y, Stefan C, Hall A, Tallon L, Sadzewicz L, Nadendla S, Klimke W, Hatcher E, Shumway M, Aldea DL, Allen J, Koehler J, Slezak T, Lovell S, Schoepp R, Scherf U. FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science. Nat Commun. 2019;10(1):3313.
25. Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, Smith CL, Edgley ML, Rougvie AE, Fire AZ, Morishita S, Schwarz EM. Recompleting the caenorhabditis elegans genome. Genome Res. 2019;29(6):1009–22.
26. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg LA, Bouffard P, Burt DW, Crasta O, Crooijmans RPMA, Cooper K, Coulombe RA, De S, Delany ME, Dodgson JB, Dong JJ, Evans C, Frederickson KM, Flicek P, Florea L, Folkerts O, Groenen MAM, Harkins TT, Herrero J, Hoffmann S, Megens H-J, Jiang A, de Jong P, Kaiser P, Kim H, Kim K-W, Kim S, Langenberger D, Lee M-K, Lee T, Mane S, Marcais G, Marz M, McElroy AP, Modise T, Nefedov M, Notredame C, Paton IR, Payne WS, Pertea G, Prickett D, Puiu D, Qioa D, Raineri E, Ruffier M, Salzberg SL, Schatz MC, Scheuring C, Schmidt CJ, Schroeder S, Searle SMJ, Smith EJ, Smith J, Sonstegard TS, Stadler PF, Tafer H, Tu ZJ, Van Tassell CP, Vilella AJ, Williams KP, Yorke JA, Zhang L, Zhang H-B, Zhang X, Zhang Y, Reed KM. Multi-platform next-generation sequencing of the domestic turkey (meleagris gallopavo): genome assembly and analysis. PLoS Biol. 2010;8(9):e1000475.
27. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):506–15.
28. Babb PL, Lahens NF, Correa-Garhwal SM, Nicholson DN, Kim EJ, Hogenesch JB, Kuntner M, Higgins L, Hayashi CY, Agnarsson I, Voight BF. The nephila clavipes genome highlights the diversity of spider silk genes and their complex expression. Nat Genet. 2017;49(6):895–903.
29. Sheetlin S, Park Y, Frith MC, Spouge JL. ALP & FALP: C++ libraries for pairwise local alignment e-values. Bioinformatics. 2016;32(2):304–5.
30. Federhen S. The NCBI taxonomy database. Nucleic Acids Res. 2012;40(Database issue):136–43.
31. Frith MC. A new repeat-masking method enables specific detection of homologous sequences. Nucleic Acids Res. 2011;39(4):23.
32. Breitwieser FP, Baker DN, Salzberg SL. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Genome Biol. 2018;19(1):198.
33. Breitwieser FP, Salzberg SL. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. Bioinformatics. 2019. https://doi.org/10.1093/bioinformatics/btz715.
34. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91.
35. Steinegger M, Salzberg SL. Github repository of Conterminator https://github.com/martin-steinegger/conterminator. Accessed 14 Apr 2020.
36. Steinegger M, Salzberg SL. Zenodo source of Conterminator https://zenodo.org/record/3750825. Accessed 14 Apr 2020.
37. Steinegger M, Salzberg SL. Figshare data repository for Conterminator https://figshare.com/projects/Conterminator/77346. Accessed 18 Mar 2020.

Publisher's Note