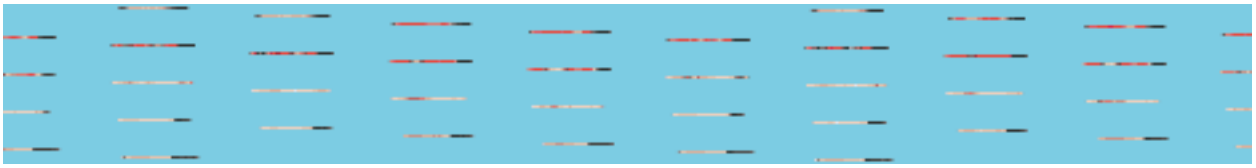
Human contamination in bacterial genomes has created thousands of spurious proteins

Florian P Breitwieser, Mihaela Pertea, Aleksey Zimin, et al.

Genome Res. published online May 7, 2019
Access the most recent version at doi:10.1101/gr.245373.118



Human contamination in bacterial genomes has created thousands of spurious proteins

Florian P. Breitwieser[1,*,] Mihaela Pertea[1,2], Aleksey V. Zimin[1,3], and Steven L. Salzberg[1,2,3,4,*]

1 Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD

2 Department of Computer Science, Whiting School of Engineering, Johns Hopkins University

3 Department of Biomedical Engineering, Johns Hopkins University

4 Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University

*To whom correspondence should be addressed: florian.bw@gmail.com, salzberg@jhu.edu.

Abstract

Contaminant sequences that appear in published genomes can cause numerous problems for downstream analyses, particularly for evolutionary studies and metagenomics projects. Our large-scale scan of complete and draft bacterial and archaeal genomes in the NCBI RefSeq database reveals that 2250 genomes are contaminated by human sequence. The contaminant sequences derive primarily from high-copy human repeat regions, which themselves are not adequately represented in the current human reference genome, GRCh38. The absence of the sequences from the human assembly offers a likely explanation for their presence in bacterial assemblies. In some cases, the contaminating contigs have been erroneously annotated as containing protein-coding sequences, which over time have propagated to create spurious protein “families” across multiple prokaryotic and eukaryotic genomes. As a result, 3437 spurious protein entries are currently present in the widely-used nr and TrEMBL protein databases. We report here an extensive list of contaminant sequences in bacterial genome assemblies and the proteins associated with them. We found that nearly all contaminants occurred on small contigs in draft genomes, which suggests that filtering out small contigs from draft genome assemblies may mitigate the issue of contamination while still keeping nearly all of the genuine genomic sequences.

Introduction

Over the past two decades, the number of publicly available genomes has grown from just a handful of species to well over 100,000 genomes today. These genomes are pivotal resources for countless biomedical research questions, including microbiome studies that use them to identify species in complex samples (Breitwieser et al. 2017). Ideally, all genomes in reference databases would be complete and accurate (Fraser et al. 2002), but for practical reasons, the vast majority of genomes available today are still “drafts.” A draft genome consists of multiple contigs or scaffolds that are typically unordered and not assigned into chromosomes (Ghurye et al. 2016). A genome is not truly complete or "finished" until every base pair has been determined for every chromosome and organelle, end-to-end, with no gaps. Even the human genome, although far more complete than most other animal genomes, is still unfinished: the current human assembly, GRCh38.p13 (released February 28, 2019), has 473 scaffolds that contain 875 internal gaps. While most of the human sequence has been placed on chromosomes, some highly repetitive regions are under-represented (Altemose et al. 2014), leading to problems that we discuss below. Draft genomes of other species vary widely in quality as well as contiguity, with some having thousands of contigs and others having a much smaller number.

Contamination of genome assemblies with sequences from other species is not uncommon, especially in draft genomes (Longo et al. 2011; Merchant et al. 2014; Delmont and Eren 2016; Kryukov and Imanishi 2016; Lu and Salzberg 2018). In 2011, researchers reported that over 10% of selected non-primate assemblies in the NCBI and UCSC Genome Browser databases were contaminated with the primate-specific AluY repeat (Longo et al. 2011). Although validation pipelines have improved substantially since then (Tatusova et al. 2016; Haft et al. 2018), some contaminants still remain, as we describe below. Furthermore, when open reading frames (ORFs) in the contaminated contigs get annotated as protein-coding genes, their protein sequence may be added to other databases. Once in those databases, these spurious proteins may in turn be used in future annotation, leading to the so-called “transitive catastrophe” problem where errors are propagated widely (Karp 1998; Salzberg 2007; Danchin et al.

2018). Indeed, one study found that the percentage of mis-annotated entries in the NCBI non-redundant (nr) protein collection, which is used for thousands of BLAST searches every day, has been increasing over time (Schnoes et al. 2009).

Contamination of genomic sequences can be particularly problematic for metagenomic studies. For example, if a genome labelled as species X contains fragments of the human genome, then any sample containing human DNA might erroneously be identified as also containing species X. Since human DNA is virtually always present in the environment of sequencing laboratories, human contamination is very common in sequencing experiments of all types. Contamination of laboratory reagents with DNA from other organisms can also lead to serious misinterpretations, such as the supposed detection of the novel virus NIH-CQV in hepatitis patients, which was ultimately determined to be a contaminant of nucleic acid extraction kits (Smuts et al. 2014).

In the process of assembling a non-human genome, any fragments of human DNA present in the data will typically remain un-assembled or will form separate, relatively small contigs. These contigs or the raw reads can be filtered out by alignment to the human genome, using fast methods such as Bowtie 2 (Langmead and Salzberg 2012) or BWA (Li and Durbin 2009), or the slower but more sensitive BLAST aligner (Camacho et al. 2009). This type of filtering procedure is very effective, but it still fails when the human genome assembly does not contain the human sequence that one is trying to remove.

In order to identify human contamination in publicly available microbial genomes, we undertook a systematic search of the bacterial and archaeal sections of the NCBI RefSeq genome database (O'Leary et al. 2016). By employing profile Hidden Markov Models of human repeats, we were able to detect repeats that were more divergent than other aligners might detect. We further searched for and found numerous erroneous protein entries in the NCBI nr database and in the TrEMBL protein database (The UniProt Consortium 2017), the vast majority of them labelled as bacterial or archaeal, that originated from human repeat sequences.

Results

Variants of high-copy numbers repeats are not fully represented in the human reference genome

The current assembly of the human genome, (currently GRCh38), although far more complete and contiguous than earlier versions, is still missing some sequences, particularly the repeat-rich centromeres and pericentromeric regions (Miga et al. 2015). Although most of these high-copy repeats, such as HSATII (human satellite II), have been well-characterized and widely studied (Prosser et al. 1986; Garrido-Ramos 2017; Hall et al. 2017), the human reference genome contains only a limited number of copies of them (Altemose et al. 2014). Due to variation among the many copies of these repeats, some of which occur thousands of times, some repeats do not match the reference genome very well. *This phenomenon appears to have contributed substantially to the ongoing presence of human contamination in draft genomes.*

As an illustration, consider Figure 1, which shows the alignment of a whole-genome shotgun data set from one of the Simons Genome Diversity Project genomes (Mallick et al. 2016) to GRCh38. The region shown is near the centromere of Chromosome 1 (Chr 1: 125,179,927 - 125,180,401), where GRCh38 contains several tandemly repeated copies of HSATII. The average depth of coverage of this sample was 100X, meaning that most locations on the genome have ~100 reads covering them. In the region shown, though, the depth of coverage is over 157,000×. Because each read was aligned to the best-matching location in GRCh38, this suggests that the reads from over 1,500 copies of the HSATII sequence have aligned to this one location, because other (better-matching) copies are simply not present in the genome assembly. As the figure shows, many of the reads have substantial numbers of differences, including 5-10bp deletions, with respect to the reference genome.

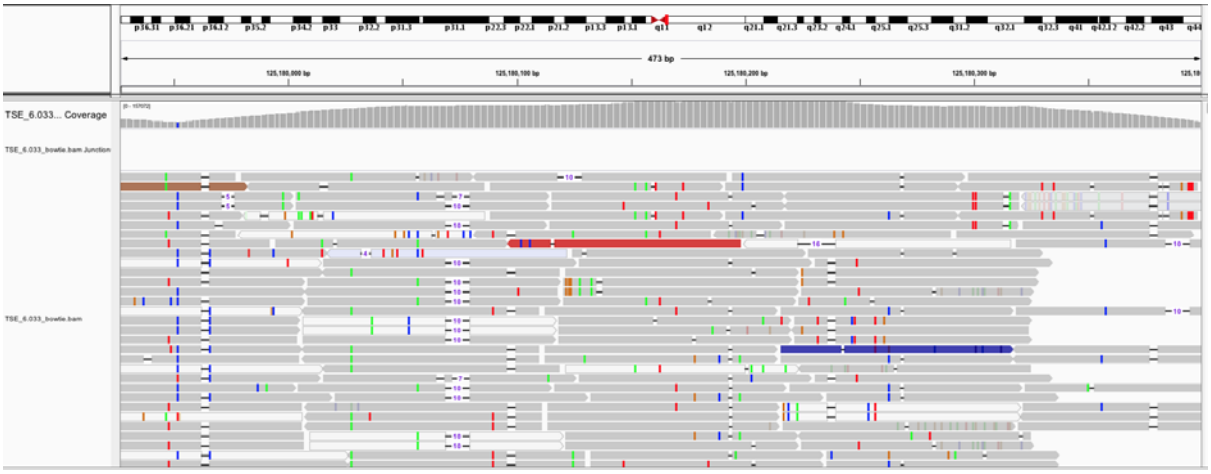


Figure 1: Alignment of a human whole-genome shotgun sequencing data set to GRCh38 shown in the Integrated Genome Viewer. This region, which contains a copy of the HSATII repeat, is covered extremely deeply, over 1500-fold deeper than the rest of the genome. The region at the top shows a schematic of chromosome 1, and below that is a histogram showing the depth of coverage, which peaks at 157,072. Individual reads in their aligned positions are shown as grey rectangles in the bottom portion of the figure. Mismatches are shown by red, blue, green, or brown marks, and gaps indicated by breaks in the grey rectangles connected with a thin black line. The numerous gaps and mismatches suggest that GRCh38 is missing many other copies of the HSATII repeat, some of which would provide a better match.

One consequence of the missing repeats in the human reference genome is that attempts to filter human reads from genome sequencing projects of bacteria and other non-human organisms may miss these repeats; e.g., if a read does not match the human genome well enough, it will not be recognized as human. Thus, reads that slip through these filters are disproportionately likely to come from high-copy repeat regions such as the one shown in Figure 1. Note that, because more than half of the human genome is covered by repeats (de Koning et al. 2011), a random human read is more likely to originate from one of these regions than from a non-repetitive region.

Over one thousand bacterial and archaeal genomes in RefSeq contain scaffolds mapping to human repeats

To identify human repeats in non-human genomes, we searched profile HMMs from the Dfam database of eukaryotic repeats (Hubley et al. 2016) against all archaeal and bacterial genomes in RefSeq release v90 (Sept 17, 2018). In addition, we also screened the same genomes against the complete human reference genome using KrakenUniq and MUMmer. In total, this release contains 749 archaeal and 129,090 bacterial genomes, of which 264 and 10,639, respectively, were labelled as complete and the remainder were draft assemblies.

In total, 2250 bacterial and archaeal assemblies in RefSeq have scaffolds that align to human repeat profiles or the human reference genome (see **Table 1**). All but 6 of the contaminated assemblies were draft genomes, and 99.7% of the matching scaffolds were shorter than 10 kilobases (kbp) (see next section). 49 of the contaminated draft genomes are in the category ‘representative genome’ which indicates that they are considered high-quality representatives of specific species or strains (O'Leary et al. 2016). **Supplemental Tables S1** and **S2** contains details on all contaminating scaffolds, the bacterial and archaeal assemblies in which they appear, the human repeat profile they match, and the best matching sequence in the human genome.

Profile	**Description**	**Profile length [bp]**	**Genomes**		
			Complete	**Draft**	**Total**
LINEs	Long interspersed nuclear elements, > 15% of human genome	~6000	2	1066	1068
***Alu* family**	Most abundant SINEs, about 10% of human genome	~300	3	746	749
Satellites	Satellite repeats ALR, BSR,	~170	0	910	910

	HSATII				
LTRs	Long terminal repeats from endogenous retroviruses	~200 - 5000	0	228	228
DNA Transposon	Tigger1 DNA transposon	2418	1	20	21
Other	Matching the human reference genome	-	0	373	373
Total			**5**	**2245**	**2250**

Table 1: Summary of human repeat elements found in bacterial and archaeal genomes. The last three columns show the number of distinct RefSeq genomes (complete, draft and total) containing each of the different human repeat types. As some genomes match more than one type of repeat, the total number of distinct genomes containing human sequences (last row), is not simply the sum of the cell values.

1068 prokaryotic genomes contain sequences that match repeats of the LINE (long interspersed nuclear elements) group, which also account for about 1/6th of the human genome overall (Sheen et al. 2000). The short interspersed nuclear element (SINE) family of *Alu* repeats, which accounts for about 10% of the human genome (Batzer and Deininger 2002), appears in 746 draft and 3 complete genomes. Note that a previous study found AluY contamination in 11 of 94 NCBI assemblies and 11 of 42 UCSC assemblies (Longo et al. 2011). In the family of repeats of the satellite regions of the human DNA, we found 910 assemblies that matched either ACRO1, ALR, BSR or HSATII (Prosser et al. 1986; Vissel and Choo 1987). In addition, 228 genomes contained a Long Terminal Repeat (LTR) retrotransposon sequence.

Over half of the contaminated bacterial and archaeal genome assemblies contain two or more scaffolds mapping to human repeats (see **Supplemental Fig. S1**). 26 assemblies have more than 50 contaminating contigs with up to 798 scaffolds mapping to the human genome or repeats (see **Supplemental Table S3**). The majority of contigs or scaffolds that map to a repeat only map to one copy of a human repeat element

(see **Supplemental Fig. S2B**). Some contigs and scaffolds, however, map to multiple copies of the same or different repeat profiles. For example, a 4.4 kbp sequence (accession NZ_CMHF01000052.1) in the *Streptococcus pneumoniae* assembly GCF_001116085.1, which was isolated from a human nasopharynx, contains 26 copies of the human ALR repeat, and a 6.8 kbp sequence (NZ_FTZV01000200.1) in the *Shigella sonnei* assembly GCF_900159525.1, which was isolated from human stool, contains 26 copies of the HSATII repeat. There is some relationship between the number of repeats per sequence and the sequence length (see **Supplemental Fig. S2A**). ALR and HSATII are the only repeats that are found more than 10 times in a single contig (see **Supplemental Figs. S2B** and **S2C**).

We investigated whether the date or the type of sequencing technology was associated with contamination in assemblies. We found a steady increase in the proportion of contaminated assemblies over the last four years, from 0.7% in 2015 to 2.8% in 2018 (**Supplemental Fig. S5A)**. In the same time frame, there has been an increase in the number of assemblies using Oxford Nanopore and Pacific Biosciences technology (from 2.9% in 2015 to 8.4% in 2018); however, over 99% of the contaminated assemblies were generated using exclusively short-read Illumina sequences (**Supplemental Fig. S5B**). Genome assemblies that employed multiple sequencing technologies also had less contamination (**Supplemental Fig. S6**).

Human contamination is found almost exclusively on small, low-coverage contigs

We expected that contigs resulting from human contamination would be short, for several reasons. First, because reads from bacteria and humans do not overlap, assembly algorithms should not integrate contaminants into the bacterial genome sequence. Second, assuming the amount of contaminating human DNA is small, the coverage of the human genome will be very low, which in turn means that only high-copy repeats are likely to assemble into contigs. For the same reason, any human reads that do assemble are likely to form relatively small contigs.

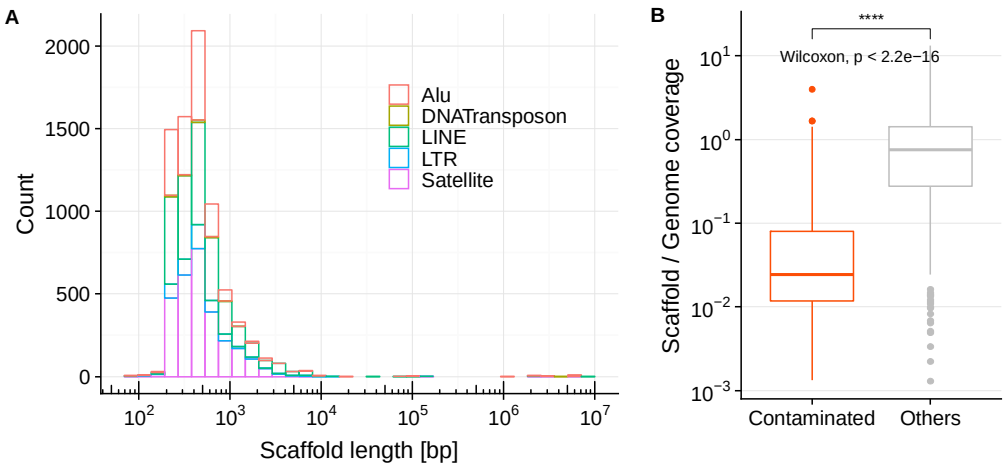


Figure 2: Lengths of scaffolds in prokaryotic genomes that contain or consist entirely of human repeats. (A) Histogram showing the number of scaffolds of a given length that contain human repeats. (B) The coverage depth of contaminant scaffolds is on average 30 times lower than the average genome coverage (red box). Similar-sized scaffolds in the same assemblies do not show the same trend in this extend (gray box, Wilcoxon signed rank test $p < 2.2 \times 10^{-16}$).

Consistent with this expectation, we found that 99.7% of contaminated contigs and scaffolds are shorter than 10 kbp, 99.3% are below 5 kbp, and 92% are below 1 kbp (**Figure 2A**). The median length of contigs containing human repeat sequences is 356 bp. At the same time, only 0.34% of the total sequence of those assemblies is in scaffolds smaller than 1 kbp, 1.8% of sequence is in scaffolds smaller than 5 kbp, and 3.6% of the total bacterial and archaeal sequence in RefSeq is in contigs that are less than 10 kbp in size (**Supplemental Fig. S3**). Just 19 genomes had scaffolds longer than 100 kbp with matches to human repeats (see **Supplemental Table S4**). We examined all of these and found that most are probably mis-assembled (though only five assemblies provide sequencing data); see **Supplemental Materials** for details. For several strains of *Neisseria gonorrhoeae*, however, a human repeat appears in the middle of a correctly-assembled genomic sequence which seems to represent a genuine case of horizontal gene transfer from humans to bacteria. This extremely unusual case was previously reported (Anderson and

Seifert 2011), and our results confirm and extend that finding to include eleven sequenced genomes of *N. gonorrhoeae* (see **Supplemental Materials** and **Supplemental Fig. S7** for further details).

We further explored whether sequence coverage may be used in the assembly process to filter out contaminant sequences. When sequencing reads are available, they can be mapped back to the assembled scaffolds, and the average coverage of the scaffolds can be computed. Assuming that only a small amount of human contamination is present, we would expect that any assembled contaminants would have lower coverage than the target genome. We retrieved the raw Illumina sequencing data for 427 contaminated assemblies from the Sequence Read Archive (Leinonen et al. 2011), and aligned the reads back to their assemblies (see **Methods**). We selected 219 high-quality samples for further analysis, choosing those with at least 20× coverage (see **Supplemental Fig. S4**), and found that contaminated scaffolds had significantly lower coverage than the genome-wide average (**Figure 2B**, red box). To ensure that this difference was not an artefact of the small size of the contaminating scaffolds, we also compared the coverage of non-contaminant scaffolds that were similar in size (**Figure 2B**, gray box). We found that even compared to scaffolds of the same size, contaminated scaffolds have significantly lower coverage.

Bacterial "proteins" that derive from human repeat contamination made their way into protein databases

Some human repeats contain open-reading frames (ORFs) that are long enough to be considered as possible protein-coding genes. When automated annotation methods erroneously identify these ORFs as proteins, they may subsequently be stored in databases as bacterial proteins. To identify proteins in the nr and TrEMBL protein databases that derive from human repeat sequences, we extracted all the nucleotide regions identified in the previous section and matched them against those databases using the fast translated search implemented in PLAST (Nguyen and Lavenier 2009).

In total, we found 3473 distinct protein entries in nr and TrEMBL that derive from human repeats. (See **Supplemental Tables S5 and S6** for the query results and **Supplemental Files S1 and S2** for the protein sequences.) These 3437 entries contain 2245 unique protein sequences, 2009 of which were found in nr (1866 bacterial, 5 archaeal, and 138 eukaryotic, including 10 entries that have been very recently suppressed) and 888 of which were found in TrEMBL (530 bacterial, 2 archaeal, and 264 eukaryotic, including 92 since deleted). Merging these two sets and removing identical matches yielded 2245 unique proteins. Note that we only identified eukaryotic proteins as spurious if they were found in non-vertebrates and were near-identical to human repeat sequences. A large fraction (113) of the spurious eukaryotic proteins were found in a single genome, *Plasmodium ovale wallikeri*.

Spurious proteins that derive from the same human repeat sequence are, as expected, nearly identical; see **Figure 3** for an alignment of spurious HSATII-derived proteins annotated in bacteria. Because RefSeq combines redundant protein sequences into “Identical Protein Groups” (Haft et al. 2018), some of the matches in nr cover many species. For example, accession WP_016831114.1, which hits HSATII, contains 21 assemblies and 14 proteins from various bacteria. NCBI assigns a taxonomic class to these identical protein groups using the lowest common ancestor; in this case the assigned taxon is Bacteria because the group contains organisms across diverse bacterial phyla including Proteobacteria, Firmicutes, and Actinobacteria. The largest such group we found was accession WP_021666093.1, which contains 62 different coding regions in 59 different assemblies. Because identical proteins have been collapsed into these groups, the 2009 spurious non-redundant protein entries that we identified in nr encompass a total of 3473 distinct proteins.

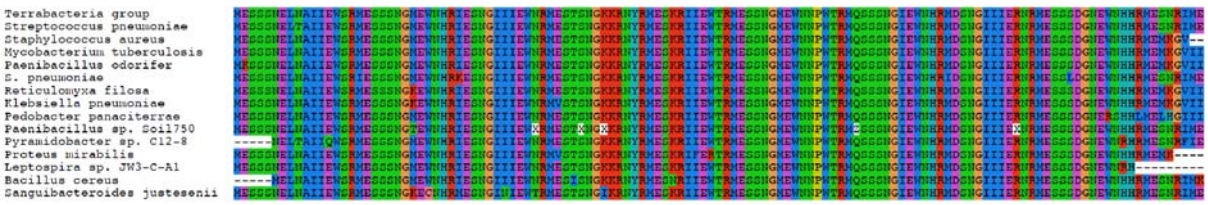


Figure 3: Human repeat element HSATII-derived proteins annotated in bacteria are nearly identical to one another, as shown in this multiple alignment, despite the large evolutionary distances separating the species in which they were reported. Visualized with SeaView (Gouy et al. 2010).

While most of the human-derived “proteins” that we found in bacterial genomes were annotated as ‘hypothetical protein’ or ‘uncharacterized protein’, some were assigned other names. For example, we found 183 proteins that were annotated as ‘Nef attachable domain protein,’ most of them in the bacterium *Chlamydia psittaci*. The origin of this annotation is a human protein entry for ‘Nef attachable protein’ (GenBank accessions BAA95214.1 and AB015434.1; UniProtKB accession Q9P2Y3), which in turn derives from the 172 base-pair long ALRa repeat (see **Supplemental Fig. S8**). The entry for the human protein-coding gene has been removed: the NCBI database contains a note that “LocusID 55866 was defined by AB015434.1 and NM_018483.1 which do not appear to represent a protein-coding gene.” However, both the human protein and mRNA entries, as well as the *Chlamydia psittaci* entries, are still present in GenBank, nr, and TrEMBL. Another example is a 10,174-bp contig (accession FPIH01000010.1) in a *Chlamydia abortus* assembly that is a mouse contaminant (and that also matches the human LINE-1 element). The seven genes annotated on this contig, all of which derive from contamination, have names that include “Exodeoxyribonuclease,” “exonuclease III,” and "L1 transposable element."

We also identified a handful of additional spurious proteins that fell below our 95% identity threshold but that are nonetheless clearly human contaminants. Two examples are the “putative sulfurtransferase” and a "centlein-like protein (CNTLN)" (protein IDs AIF19795 and AIF19796) in an assembly of a marine archaeon. Both proteins are part of a mis-assembly that was created by erroneously concatenating a human *Alu* sequence onto the end of a 35-kbp contig (KF901147.1). Protein AIF19795 is a false chimera, spanning the point in the mis-assembly where bacteria and human DNA were concatenated together, and its first 100 amino acids (out of 139) represent a genuine bacterial sulfurtransferase), while the remaining

amino acids are a translation of the human *Alu* repeat. The following and final protein on the contig, AIF19796, is a translation of an entirely human repeat sequence.

To quantify the impact of protein database contamination on metagenomics searches, we conducted translated searches of genuine human sequences against the *nr* protein database. In sequencing experiments, methods like PLAST (Nguyen and Lavenier 2009), diamond (Buchfink et al. 2015) or MMSeqs2 (Steinegger and Soding 2017) may be used to map sequencing reads to proteins and compute a taxonomic profile of the results. When all the reads are human, the taxonomic profile should shows only human or primate entries. We generated 19 million simulated reads that covered the human genome (see Methods), of which ~411,000 mapped to nr proteins with an e-value below 10^{-7} (see **Supplemental Fig. S9**). As expected, the majority of these high-scoring reads matched primates (56.78%). However, > 21% of the reads matched to various bacteria including *Mycobacterium tuberculosis* (2.42%), *Bacillus cereus* (1.15%) and *Klebsiella pneumoniae* (0.65%). A portion of reads also went to the eukaryotic human pathogen *Plasmodium ovale* (2.36%) and various nematodes (0.72%). Our searches of human repeats, described above, found 2029 spurious bacterial proteins in the nr database; these new searches identified 1050 spurious bacterial proteins, of which 28% were not found by the earlier repeat-based searches.

Discussion

We demonstrated that human contamination has made its way into 1731 publicly available microbial genomes, primarily bacteria but also archaea and some eukaryotes. In turn, erroneous translations of these contaminants have generated more than 3000 annotated proteins, which now form highly conserved but spurious protein families spanning a broad range of bacterial phyla and some eukaryotic species. All of these genomes and proteins appear in at least one if not several widely-used sequence databases. It is possible that additional contaminants might be present, because we did not screen for all possible sources

of contamination, such as other human genomic regions, fragments of DNA from non-human host organisms, environmental sources, and laboratory vectors.

This widespread contamination creates serious problems for many types of scientific analyses that depend on genome and protein databases. One example where this problem is most acute is the use of metagenomic sequencing to diagnose infections, a rapidly growing clinical application in which human tissues are sequenced to identify a potential pathogen (Wilson et al. 2014; Naccache et al. 2015; Berger and Wilson 2016; Salzberg et al. 2016). In these samples, where the dominant species is human, contamination of even a small fraction of the bacterial genomes in the database will cause numerous false positives, as human reads may appear, incorrectly, to represent bacterial organisms.

Another issue that is confounded by contamination is horizontal gene transfer. When fragments of the wrong species appear in a genome, they can be mistaken for genuine horizontal gene transfer, leading to claims (e.g., (Boothby et al. 2015; Crisp et al. 2015)) that may later be shown to be incorrect once the contamination is discovered (Arakawa 2016; Salzberg 2017) .

Simply cleaning up the existing contaminated genomes will not be sufficient to correct this problem, because many proteins now exist as entries in separate databases, as we have described here. Both genome and protein databases need to be corrected, and new controls need to be established to avoid re-contamination in the future. As was pointed out more than 15 years ago, one solution is to finish as many genomes as possible; i.e., to fill in all the gaps and ensure that each chromosome is correctly assembled in one piece (Fraser et al. 2002). Admittedly, finishing every genome remains costly. A simpler alternative strategy, taking advantage of the fact that most contamination appears on small, low-coverage contigs, is to exclude those contigs when building any database containing draft genomes. Prior to releasing a genome to the public, we also recommend running sensitive searches against human repeat profiles, as we have done here. In addition, if a microbial genome was isolated from a host whose genome is available, the microbial assembly should be carefully screened against that genome as well.

Methods

We downloaded the archaeal and bacterial sequences from the NCBI RefSeq database (O'Leary et al. 2016) release 90 (9 Oct 2018). For bacteria, this included 10639 complete assemblies, 1651 chromosome-level assemblies, 53057 scaffold-level assemblies, and 63627 contig-level assemblies. The archaea comprised 264 complete genomes, 14 chromosome-level assemblies, 164 scaffold-level assemblies, and 304 contig-level assemblies. Note that NCBI characterizes assemblies that include chromosomes, scaffolds and contigs as "chromosome-level," assemblies that include scaffolds and contigs as "scaffold-level," and assemblies that only contain contigs as "contig-level." Except for those labeled as "complete," all other genomes are considered drafts. We downloaded the nr database, a non-redundant collection of protein sequences from multiple sources, as well as the SWISS-PROT database from NCBI on July 16, 2018 (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). We downloaded the TrEMBL sequences from UniProt on Sep 27, 2018. Using HMMer 3.1b2 (Eddy 2011), we mapped the entire Dfam database of eukaryotic repeats (release 2.0, 23 September 2015) (Hubley et al. 2016) against a random subset of 5000 prokaryotic genomes to find repeats that occur in multiple incomplete genomes. We then mapped selected human repeats (LINE family: L1HS_3end, L1HS_5end, L1MC4_3end, L1P1_orf2, L1PBa_5end, L2; *Alu* family: AluJo, AluSg, AluSx, AluSz, AluY, BC200, FRAM; Satellites: ACRO1, ALR, BSR, HSATII; LTR EVRs: ERVL, MER5A, MIR, MIRb, MST-int, MSTB, THE1-int; DNATransposons: Tigger1) against all prokaryotic genomes. Only mappings with an e-value below 10^{-10} were analyzed further. For the whole genome mappings, we matched all prokaryotic genomes against the human reference genome using first KrakenUniq (Breitwieser et al. 2018) and then NUCmer (Delcher et al. 2002). We only considered scaffolds that were >95% identical to the reference genome over at least ≥90% of their lengths. We extracted the matching regions of the genomic sequences using seqtk subseq, and mapped them to the human genome GRCh38.p12 (GCF_000001405.38) with BLASTN v2.7.1+ (Camacho et al. 2009) (parameters -max_hsps 1 -max_target_seqs 100 -dust no -soft_masking false) and to the nr, SWISS-PROT and TREMBL protein sequence databases with PLASTX (Nguyen and Lavenier 2009).

For the translated PLASTX search, we kept all results with an e-value below 10-7 and minimum percent identity of 95%. We queried protein and nucleotide information using NCBI's e-utilities.

The whole-genome shotgun data used in Figure 1 was from a collection of publicly-available human genomes collected by the Simons Genome Diversity Project (Mallick et al. 2016). We aligned the reads to GRCh38 with Bowtie v2.3.4.3 (Langmead and Salzberg 2012) using default settings. The alignments were processed with SAMtools (Li et al. 2009) and visualized with the Integrative Genomics Viewer (IGV) (Thorvaldsdottir et al. 2013).

To compare the coverage depth of spurious scaffolds with the coverage of the rest of the genome, we downloaded and aligned the raw sequencing data from 413 out of the 1731 assemblies with spurious contigs. The assembly records do not directly link to raw sequencing data. Instead, NCBI assembly records link to Biosamples, which may have a link to an SRA record, which is not guaranteed to be the data used for the assembly. Using the Entrez API, we found 731 assemblies had Biosamples with a link to an SRA record. 413 of these had publicly accessible Illumina data, which we downloaded using fasterq-dump and aligned to the assembled genomes using Bowtie v2.3.4.3 (Langmead and Salzberg 2012) using default settings. The median sample in this set had 2.5 million read pairs, 96.6% overall alignment rate and 122.04× average coverage. For further analyses we selected the 217 samples that had at least 1 million reads, 90% overall alignment rate and 20× average coverage (see **Supplemental Fig. S4**). We calculated the average genome coverage, as well as the coverage of the contaminated contig and the coverage of a non-contaminated contig that was closest in size to the contaminated contig using bamcov (https://github.com/fbreitwieser/bamcov).

The multiple alignment shown in Figure 3 was created with MUSCLE (Edgar 2004) and visualized in SeaView (Gouy et al. 2010). Additional data analysis and validation was done with the R statistical software (R Core Team 2017) and ggplot2 (Wickham 2016).

To perform the translated search of human sequences against the nr database, we split the human reference genome GRCh38.p12 into 19 million 160 bp synthetic reads, and ran translated searches using MMSeqs2 (Steinegger and Soding 2017). In total, 411,340 of the "reads" matched proteins with an e-value below 10^{-7}. In case of identical bitscores across multiple proteins, the lowest common taxonomic ancestor of the proteins was assigned. The taxonomic profile of the hits was computed and visualized using Pavian (Breitwieser and Salzberg 2016).

Acknowledgments

This work was supported in part by grants R01-HG006677, R01-GM083873, and R35-GM130151 from the U.S. National Institutes of Health, grants IOS-1744309, DBI-1458178 and DBI-1759518 from the National Science Foundation, grant 2018-67015-28199 from the U. S. Department of Agriculture NIFA, and grant W911NF-14-1-0490 from the U. S. Army Research Office.

References

Altemose N, Miga KH, Maggioni M, Willard HF. 2014. Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Comput Biol* **10**: e1003628.

Anderson MT, Seifert HS. 2011. Neisseria gonorrhoeae and humans perform an evolutionary LINE dance. *Mob Genet Elements* **1**: 85-87.

Arakawa K. 2016. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America* **113**: E3057.

Batzer MA, Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370-379.

Berger JR, Wilson MR. 2016. Next-generation sequencing of tissue: A logical extension. *Neurology(R) neuroimmunology & neuroinflammation* **3**: e261.

Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Nishimura EO, Tintori SC, Li Q, Jones CD, Yandell M et al. 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 15976-15981.

Breitwieser FP, Baker DN, Salzberg SL. 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome biology* **19**: 198.

Breitwieser FP, Lu J, Salzberg SL. 2017. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* doi:10.1093/bib/bbx120.

Breitwieser FP, Salzberg SL. 2016. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *bioRxiv*.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59-60.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G. 2015. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome biology* **16**: 50.

Danchin A, Ouzounis C, Tokuyasu T, Zucker JD. 2018. No wisdom in the crowd: genome annotation in the era of big data - current status and future prospects. *Microbial Biotechnology* **11**: 588-605.

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**: e1002384.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**: 2478-2483.

Delmont TO, Eren AM. 2016. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* **4**: e1839.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**: e1002195.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Fraser CM, Eisen JA, Nelson KE, Paulsen IT, Salzberg SL. 2002. The value of complete microbial genome sequencing (you get what you pay for). *J Bacteriol* **184**: 6403-6405; discusion 6405.

Garrido-Ramos MA. 2017. Satellite DNA: An Evolving Topic. *Genes (Basel)* **8**.

Ghurye JS, Cepeda-Espinoza V, Pop M. 2016. Metagenomic Assembly: Overview, Challenges and Applications. *The Yale journal of biology and medicine* **89**: 353-362.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**: 221-224.

Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR et al. 2018. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res* **46**: D851-D860.

Hall LL, Byron M, Carone DM, Whitfield TW, Pouliot GP, Fischer A, Jones P, Lawrence JB. 2017. Demethylated HSATII DNA and HSATII RNA Foci Sequester PRC1 and MeCP2 into Cancer-Specific Nuclear Bodies. *Cell Rep* **18**: 2943-2956.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AF, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**: D81-89.

Karp PD. 1998. What we do not know about sequence analysis and sequence databases. *Bioinformatics* **14**: 753-754.

Kryukov K, Imanishi T. 2016. Human Contamination in Public Genome Assemblies. *PloS one* **11**: e0162424.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. 2011. The sequence read archive. *Nucleic Acids Res* **39**: D19-21.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Longo MS, O'Neill MJ, O'Neill RJ. 2011. Abundant human DNA contamination identified in non-primate genome databases. *PloS one* **6**: e16410.

Lu J, Salzberg SL. 2018. Removing contaminants from databases of draft genomes. *PLoS Comput Biol* **14**: e1006277.

Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A et al. 2016. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**: 201-206.

Merchant S, Wood DE, Salzberg SL. 2014. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* **2**: e675.

Miga KH, Eisenhart C, Kent WJ. 2015. Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res* **43**: e133.

Naccache SN, Peggs KS, Mattes FM, Phadke R, Garson JA, Grant P, Samayoa E, Federman S, Miller S, Lunn MP et al. 2015. Diagnosis of neuroinvasive astrovirus infection in an immunocompromised adult with encephalitis by unbiased next-generation sequencing. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **60**: 919-923.

Nguyen VH, Lavenier D. 2009. PLAST: parallel local alignment search tool for database comparison. *BMC Bioinformatics* **10**: 329.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733-745.

Prosser J, Frommer M, Paul C, Vincent PC. 1986. Sequence relationships of three human satellite DNAs. *Journal of molecular biology* **187**: 145-155.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. Vienna, Austria.

Salzberg SL. 2007. Genome re-annotation: a wiki solution? *Genome biology* **8**: 102.

Salzberg SL. 2017. Horizontal gene transfer is not a hallmark of the human genome. *Genome biology* **18**: 85.

Salzberg SL, Breitwieser FP, Kumar A, Hao H, Burger P, Rodriguez FJ, Lim M, Quinones-Hinojosa A, Gallia GL, Tornheim JA et al. 2016. Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system. *Neurol Neuroimmunol Neuroinflamm* **3**: e251.

Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* **5**: e1000605.

Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD. 2000. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res* **10**: 1496-1508.

Smuts H, Kew M, Khan A, Korsman S. 2014. Novel hybrid parvovirus-like virus, NIH-CQV/PHV, contaminants in silica column-based nucleic acid extraction kits. *Journal of virology* **88**: 1398.

Steinegger M, Soding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**: 1026-1028.

Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* **44**: 6614-6624.

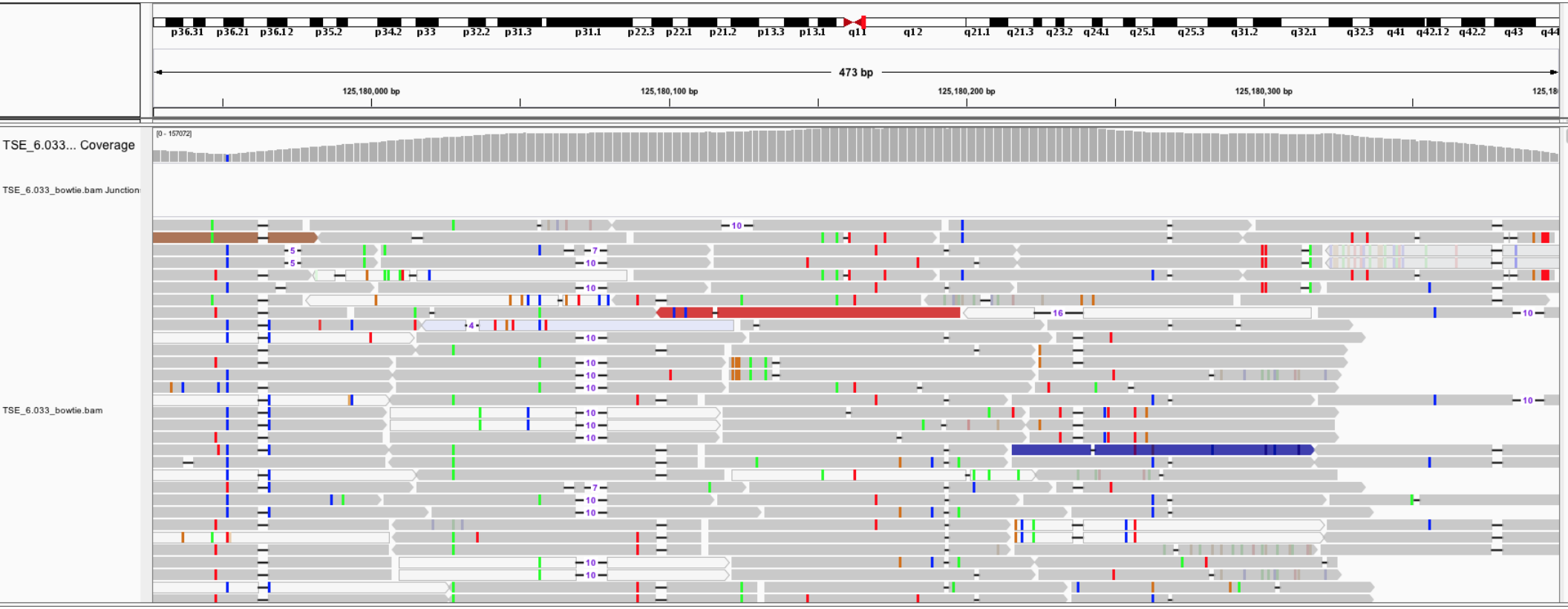
The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**: D158-D169.

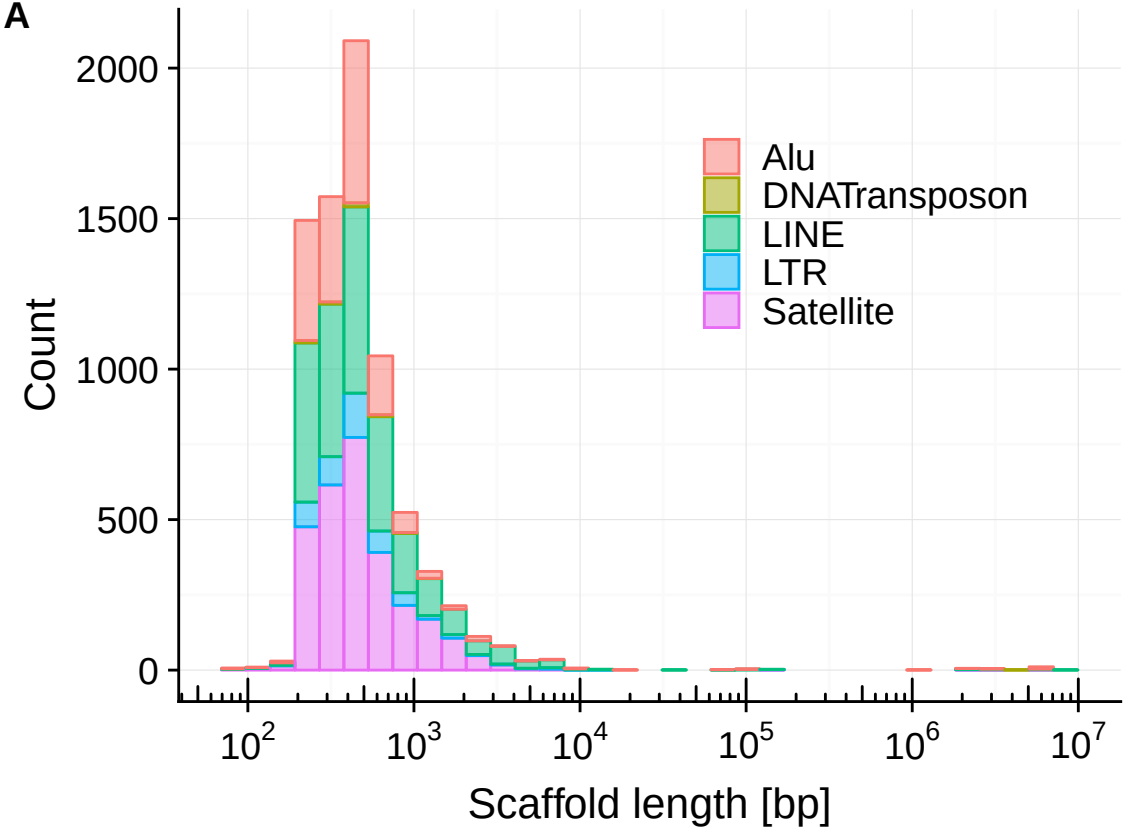
Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178-192.

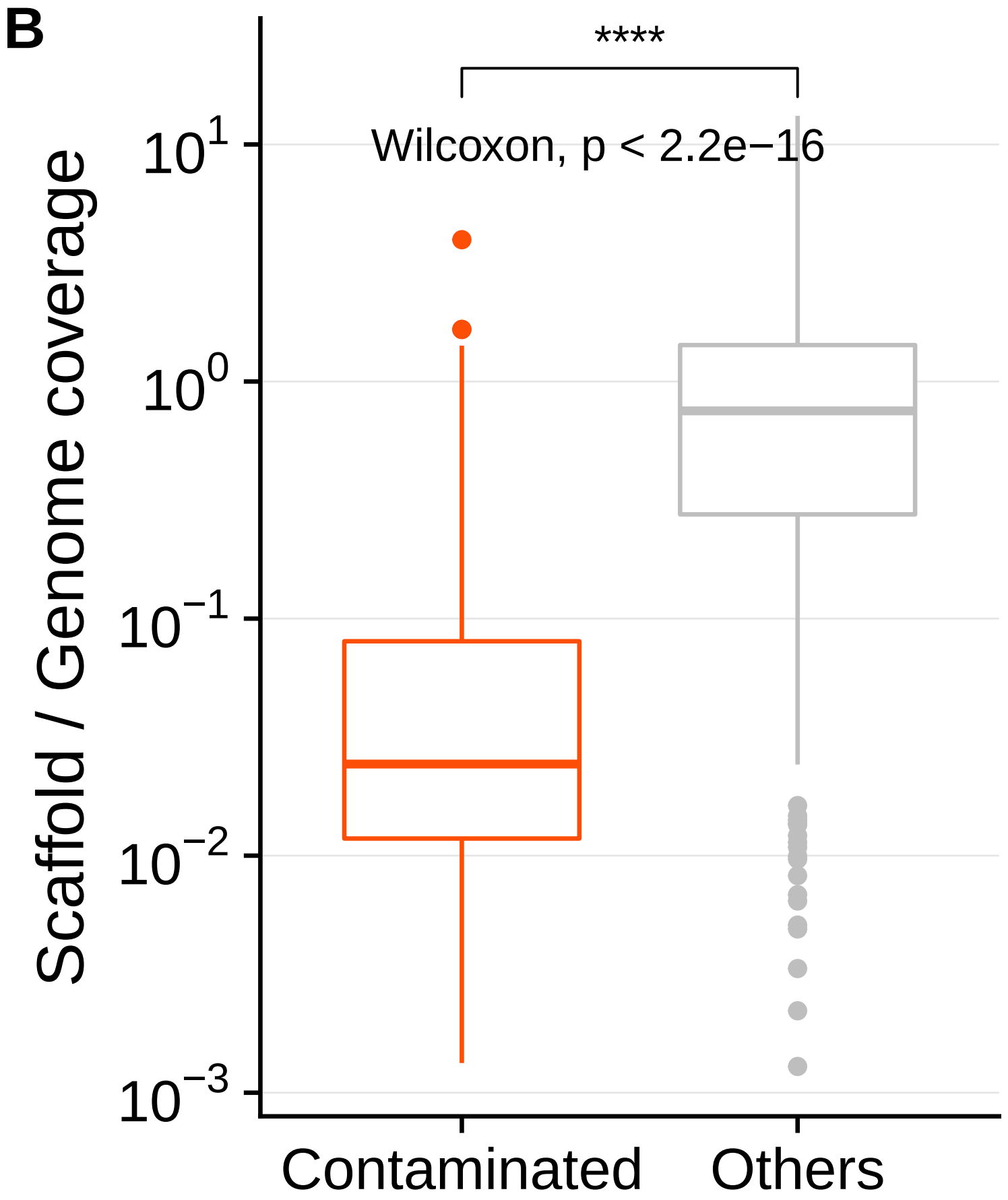
Vissel B, Choo KH. 1987. Human alpha satellite DNA--consensus sequence and conserved regions. *Nucleic Acids Res* **15**: 6751-6752.

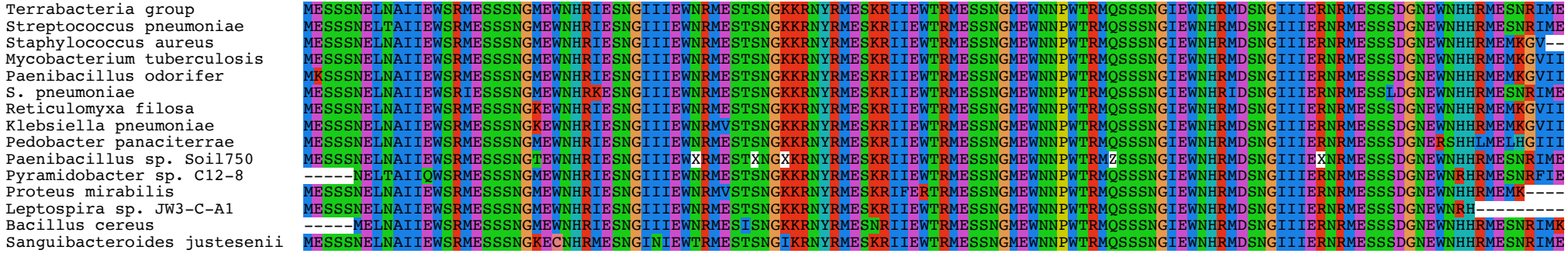
Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York.

Wilson MR, Naccache SN, Samayoa E, Biagtan M, Bashir H, Yu G, Salamat SM, Somasekar S, Federman S, Miller S et al. 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med* **370**: 2408-2417.


p36.31 p36.21 p36.12 p35.2 p34.2 p33 p32.2 p31.3 p31.1 p22.3 p22.1 p21.2 p13.3 p13.1 q11 q12 q21.1 q21.3 q23.2 q24.1 q25.1 q25.3 q31.2 q32.1 q32.3 q41 q42.12 q42.2 q43 q44
473 bp
125,180,000 bp
125,180,100 bp
125,180,200 bp
125,180,300 bp
TSE_6.033... Coverage
[0 - 157072]
TSE_6.033_bowtie.bam Junction
TSE_6.033_bowtie.bam


A
2000
1500
1000
500
0
Count
Alu
DNATransposon
LINE
LTR
Satellite
10^2
10^3
10^4
10^5
10^6
10^7
Scaffold length [bp]


B
****
Wilcoxon, p < 2.2e−16
Scaffold / Genome coverage
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
Contaminated
Others


Terrabacteria group
Streptococcus pneumoniae
Staphylococcus aureus
Mycobacterium tuberculosis
Paenibacillus odorifer
S. pneumoniae
Reticulomyxa filosa
Klebsiella pneumoniae
Pedobacter panaciterrae
Paenibacillus sp. Soil750
Pyramidobacter sp. C12-8
Proteus mirabilis
Leptospira sp. JW3-C-A1
Bacillus cereus
Sanguibacteroides justesenii