Supplemental Materials and Methods
for Breitwieser et al., Genome Research 2019

Supplemental Fig. S1: Many bacterial and archaeal assemblies have more than one scaffold that matches to human sequences. Note that the x-axis is truncated at 50, but there are 26 additional genomes with up to 798 scaffolds that match the human genome or repeats (see Supplemental Table S3).

Supplemental Fig. S2: Many contigs and scaffolds from bacterial and archaeal assemblies contain multiple human repeats. For genomes whose assembly consisted entirely of contigs, contigs were counted as if they were scaffolds. (A) Larger scaffolds tend to contain more copies of the repeats. (B) The majority of scaffolds match less than 5 repeats. (C) ALR and HSATII often appear in multiple copies on the same scaffolds.
Supplemental Fig. S3: Cumulative proportion of genomic sequence based on scaffold lengths, for all bacterial and archaeal assemblies in RefSeq. Only 3.6% prokaryotic genomic sequences in RefSeq are in scaffolds that are 10 kbp or less in size, 1.8% are in scaffolds of size 5 kbp or less, and 0.34% are in scaffolds of size 1 kbp or less.

Supplemental Fig. S4: Number of reads, overall alignment rate to the genome, and average coverage in the back-alignment of raw sequencing runs from 427 assemblies. Red lines indicate the thresholds for the inclusion of samples in further analyses: at least 1 million read pairs, 90% overall alignment rate, and 20× or higher average coverage.
Supplemental Fig. S5: (A) The percentage of contaminated assemblies (yellow) has remained relatively stable over time, although the last four years saw a steady increase, from 0.7% in 2014 to 2.8% in 2018. (B) Assemblies that use long read-sequencing technologies have been increasing in recent years, however long read sequencing was rarely used in the assemblies that had contaminated scaffolds.

Supplemental Fig. S6: Sequencing technologies used in (A) all assemblies and (B) contaminated assemblies. The main diagonal shows the percent of genomes that used exactly one technology (e.g., Illumina-only in the upper left), and off-diagonal entries show the percent of genomes that use combinations of technologies. Contaminated assemblies are more often Illumina-only and employ only a single sequencing technology.
Analysis of large bacterial scaffolds containing human sequence

The vast majority of contaminated contigs and scaffolds found in our study were shorter than 10 kbp. However, we found 20 scaffolds and chromosomes larger than 100 kbp in 19 distinct genome assemblies that contain one or multiple human repeats. We investigated each of these matches, which are listed in Supplemental Table S4, to assess whether they represent true insertions—i.e., horizontal gene transfer from humans to bacteria—or instead represent errors or contamination. The assembly status of the 19 genomes included (according to RefSeq), ‘Complete genome’ (5), ‘Chromosome’ (5), and ‘Scaffold’ or ‘Contig’ (9). Two of the assemblies, Tetrashaera japonica and Streptomyces viridosporus, are considered ‘representative genomes’, meaning they are considered high-quality genomes which are annotated with non-redundant RefSeq protein accessions.

Four of the five complete genome were probably draft assemblies that were maybe inadvertently uploaded as complete genomes. Two of the complete genomes are strains of Klebsiella pneumoniae that were sequenced with Illumina technology (Sabirova et al. 2016), assembled into contigs, and it seems erroneously deposited in GenBank as complete when in fact they are draft genomes. These genomes contain the Alu repeats that are not present in any other Klebsiella genomes, although they are flanked by sequences that are present in other Klebsiellas. Even though we could not determine the contig boundaries, it seems nearly certain that the Alu repeats are contaminants. Another complete genome that hits a human Alu repeat element is the Paracoccus mutanolyticus assembly GCF_003285265.1 - the only assembly for the species and described in 2018 (Amrutha et al. 2018). As the paper states, an initial Illumina assembly yielded 694 contigs that were then aligned to the Paracoccus yeei reference genome. However, the contaminated stretch of the genome that we found is not contained in either the Paracoccus yeei reference genome nor in any of the other 100 Paracoccus assemblies in RefSeq, but matches very well to many human sequences in the nt database. The reference-guided assembly process probably inadvertently integrated a human contig. A fourth supposedly complete assembly is also described in a manuscript (Lean et al. 2015) which only describes a draft assembly with 102 contigs. The fifth complete assembly is from PacBio sequencing of an isolate of N. gonorrhoeae which is discussed further below.

For further analysis of all the genomes, we split the scaffolds into contigs by breaking them at locations with one or multiple ‘N’ bases. We assessed the distance of the human repeats to the closest end of the contigs and scaffolds containing them. The idea behind this is that if the human sequence was caused by a mis-assembly, it is much more likely that it got integrated into the start or the end of a contig. For 13 out of the 20 scaffolds, the human repeat appears within 10 kbp of the start or stop of a contig, and for another 3 assemblies within 20 kbp (see Supplemental Table S4), strongly suggesting that these are mis-assemblies. Only in two N. gonorrhoeae assemblies the human sequence appears more than 100 kbp into the contig sequence.
Only five of the 19 assemblies had raw sequencing data linked to their records via the Sequence Read Archive (SRA). We manually investigated these data sets to determine whether the integration of a human repeat in the genome might be credible:

- **Bacillus cereus** GCF_000161355.1: 454 GS FLX Titanium runs; 641,465 spots, 181.1M bases; SRA Accessions: SRS265360, SRX098627
- **Bacillus cereus** GCF_001584095.1: Illumina MiSeq; 1.1M spots, 215.5 M bases; SRA accessions: SRS1095654, SRX1297516
- **Paracoccus mutanolyticus** RSP-02 GCF_003285265.1: Illumina NextSeq 500; 12.9M spots, 3.9G bases; SRA Accessions: SRS3704233, SRX4598341
- **Neisseria gonorrhoeae** GCF_001047255.1: PacBio RS sequencing; 163,482 spots, 2.3G bases; SRA Accessions: ERS627822, ERX930413
- **Neisseria gonnorrhoeae** GCF_900087875.1: PacBio RS sequencing; 2 runs, 163,482 spots, 90.4M bases; SRA Accessions: ERS433507, ERX1518421, ERX1518410

**Bacillus cereus** AH676, accession GCF_000161355.1
Scaffold NZ_CM000738.1 of this assembly contains an AluY repeat in positions 5307534 to 5307718. Using SRABlast, we found only two reads that support this sequence: SRR346799.115686.2 (246 bp) and SRR346799.287780.2 (244 bp). In contrast, the assembly project had an average sequencing depth of 12.9. These two reads appear to be human contaminants that were erroneously assembled into a larger contig.

**Bacillus cereus**, accession GCF_001584095.1
Scaffold NZ_LONH01000293.1 of this assembly contains an AluSx repeat at the very end of the scaffold. Using SRABlast, many reads with 100% identity appear prior to the insertion site, but none cover the whole insertion. This appears to be a mis-assembly.

**Paracoccus mutanolyticus** RSP-02 GCF_003285265.1
Scaffold NZ_CP030239.1 of this assembly contains an AluSg repeat in positions 2991199 to 2991287. Using SRABlast, we found over 100 reads supporting this region, though all but one of the reads had a percent identity below 95%. As this is Illumina data, we would expect much higher percent identity (~99.9%). Indeed, when searching another region with the same length (10000 to 10288), all the reads we found had an identity of 100%. For the possibly contaminated part of the assembly, however, we found only one read with 100% identity. It covered 37% of the sequence. That same read pair (SRR7742086.4573711) is 100% identical to human sequences. This appears to be a mis-assembly.

**Neisseria gonorrhoeae** strains GCF_001047255.1 and GCF_900087875.1
Each of these genomes was assembled into a single contig from Pacific Biosciences long read data (Abrams et al. 2015). This appears to represent an extremely rare event: horizontal gene transfer of a human repeat sequence...
into a bacterial genome. In this case, a 685-bp portion of a human L1 repeat element was previously reported as present in seven different \textit{N. gonorrhoeae} strains (Anderson and Seifert 2011). Anderson et al. conducted multiple validation tests to demonstrate that the human sequence, which was 100\% identical across the seven bacterial strains (including the two complete genomes here plus five draft genomes), was genuinely present and integrated at the same location in each of the genomes.

Our alignment results on the FA6140 data confirmed this finding and extended it to include eleven sequenced strains of \textit{N. gonorrhoeae}, all of them containing the identical 685-bp human sequence and the same flanking IS1595 transposase sequence. The 685-bp L1 sequence is 100\% identical to some copies present in the human genome (see \textbf{Supplemental Fig. S7}). To validate the assembly of \textit{N. gonorrhoeae} WHO L at the site of the human insertion (genome coordinates 637948 to 638632), we used Blasr (Chaisson and Tesler 2012) to align the PacBio reads (accession SAMEA2448463) from a WHO L strain to the complete assembly of the same strain (accession LT591901). We then extracted the PacBio reads whose alignments extended at least 1000bp before and after the insertion site. As shown in Supplementary Figure S7, these 21 reads completely span the insertion and confirm that the assembly is correct. We also compared the eleven strains to one another, and found that they share many other long (> 685bp) sequences at 100\% identity, suggesting that (a) the horizontal transfer event happened very recently, and (b) that a single event explains the presence of the human sequence in all eleven genomes. This example represents an important exception and suggests that these strains of \textit{N. gonorrhoeae} should possibly be excluded from metagenomics databases, since they might easily cause false positive matches.

\textbf{Supplemental Fig. S7:} An IGV view of the alignments of 21 PacBio reads to the assembled genome of \textit{Neisseria gonorrhoeae} strain WHO L (accession LT591901). Each of the reads extend at least 1000bp both before and after the 685-bp human line element that was inserted into this genome. The human insertion is highlighted by the blue box. Blue or pink coloring of the aligned reads reflects the strand, and indels are shown by purple “[“ symbols.
Supplemental Fig. S8: The human sequence annotated as "Nef attachable protein" (protein accession BAA95214.1, mRNA accession AB015434.1) aligns to the human alpha satellite repeat ALRa. The sequence has been discontinued from some databases, but persists in others. More importantly, the annotation persists in spurious bacterial proteins that derive from human contaminant sequences.
Supplemental Fig. S9: Taxonomic profile computed by running a translated search with 19 million 160-bp human DNA sequences against the nr protein database. The human sequences were each 160bp long and collectively spanned the entire genome; approximately 411,000 reads matched with e-values below $10^{-7}$. The thickness of the flow (and the numbers above the boxes) represent the number of reads that matched to different phylogenetic groups; e.g., ~275,000 reads matched Eukaryota and ~86,500 matched Bacteria. Visualized with Pavian.