Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Recurrent human-sequence contamination and recoverable genetic information in bacterial sequencing archives

Mahdi Moosa

mahdi.moosa@gmail.com

Brac University https://orcid.org/0000-0002-2823-3178

Article

Keywords:

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11693708/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

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

Recurrent human-sequence contamination and recoverable genetic information in bacterial sequencing archives

Mahdi Muhammad Moosa

Department of Biotechnology, BRAC University, Dhaka, Bangladesh

Correspondence: mahdi.moosa@gmail.com; mahdi.moosa@bracu.ac.bd

Abstract

Human-derived nucleic acids introduced during sample collection or laboratory handling are a longstanding concern in genomics. Sensitive, high-throughput sequencing can capture trace human nucleic acids, potentially exposing genetic information in nominally host-free microbial datasets. We systematically examined bacterial isolate sequencing archives and found frequent human-sequence contamination, with recurrence within projects and marked heterogeneity among libraries, including those sharing sequencing lanes. Human fragments spanned chromosomes and retained population-informative alleles, concordant partial allele states and clinically annotated variants. Computational depletion removed more than 99.99% of strict human fragments, eliminating measured population-affinity and clinically annotated variant signals while largely preserving bacterial assembly features. Human contamination was also detected in bacterial RNA-seq, with both genome-alignable reads and splice-junction evidence of processed human transcripts. Together, these findings show how incidental human material can become recoverable genetic information in public microbial archives and underscore the need for screening before data deposition.

Introduction

Declining sequencing costs and increasing throughput have expanded genomic and transcriptomic research across diverse organisms, generating rapidly growing collections of publicly archived nucleotide sequences[1;2]. Despite differences in biological targets, nucleic-acid sequencing workflows share multiple stages of specimen handling, extraction, library preparation and sequencing, each creating opportunities for unintended material to enter the resulting datasets (Fig. 1a). Human DNA has been traced to individuals involved in handling ancient specimens[3], while human-derived sequences have been identified in genomic resources spanning bacteria, plants and other non-primate organisms[4;5].

A prominent focus of contamination research has been its consequences for genome integrity and biological interpretation. Human-derived fragments incorporated into bacterial and archaeal assemblies have generated erroneous gene annotations and thousands of spurious protein entries in reference databases[6]. Contaminating reads can also distort bacterial variant analysis, producing false-positive and false-negative polymorphism calls[7]. These concerns have motivated systematic approaches to detecting and removing foreign sequences from assembled genomes[1]. However, assembly-level screening does not necessarily address the underlying raw sequencing reads, which are frequently deposited separately in public archives.

Raw-read archives present a different concern. Human fragments that are too sparse to contribute to a microbial genome assembly may nevertheless remain accessible in deposited reads, retaining information about human genetic variation. Bacterial isolate sequencing provides a useful setting in which to investigate this problem: human nucleic acids are not an intended component of these libraries, and the evolutionary distance between bacterial and human genomes generally facilitates discrimination between their sequences. Bush and colleagues demonstrated the significance of this exposure by recovering human polymorphism observations, including variants of potential clinical significance, from public bacterial sequencing datasets[8]. Nevertheless, the distribution of human-compatible sequences across projects and technical contexts, the breadth of recoverable genetic information, and the extent to which depletion reduces this information while preserving microbial data remain incompletely characterized.

Here we investigated human-reference-compatible sequences in bacterial isolate whole-genome sequencing using an exploratory cohort and three separately selected, contamination-blind replication cohorts. We exam-

ined contamination prevalence, project-level recurrence, technical heterogeneity and recoverable human genomic information, including chromosomal breadth, population-reference affinity and read-derived allele observations. Matched host-read depletion experiments assessed reductions in human sequence and information-related endpoints alongside the preservation of bacterial assemblies. We extended this framework to bacterial RNA sequencing, distinguishing genome-aligned human signal from splice-junction evidence of processed transcripts and evaluating their differential persistence after depletion. Together, these analyses examine the distribution, informational consequences and mitigation of unintended human sequence in microbial sequencing archives.

Results

Human-sequence contamination is prevalent and structured across bacterial isolate sequencing projects

We first examined the prevalence of human-reference-compatible sequence in an exploratory cohort (E) comprising 1,713 bacterial isolate WGS runs from 426 BioProjects and 1,703 BioSamples. The cohort was selected without reference to human-contamination outcomes and prioritized predominantly axenic bacterial culture contexts, excluding selected obligate intracellular taxa and culture conditions with plausible exposure to exogenous nucleic acids. To identify putative human sequences, we mapped sequencing reads against the telomere-to-telomere human reference genome (T2T-CHM13 v2.0)[9]. We defined *strict human fragments* using high-confidence alignments (mapping quality $\geq$ 30, query coverage $\geq$ 80% and sequence identity $\geq$ 98%), additionally requiring both mates of paired-end fragments to map concordantly as proper pairs to the same human reference sequence. Runs containing at least one qualifying fragment or single-end read were considered *strict-positive*, yielding 1,178 of 1,713 runs (68.77%). To distinguish putative human contamination from bacterial sequences that might also align to the human reference, we further evaluated candidate reads using taxonomic classification and alignment against bacterial genome assemblies. Following this bacterial-sequence adjudication, 1,169 runs (68.24%) remained *adjudicated-positive*. In 1,017 runs (59.37% of the full cohort), the human-sequence interpretation was further supported by a successfully completed comparison against an exact BioSample-matched bacterial assembly, with no qualifying competing bacterial alignment (Extended Data Fig. 1a). Thus, human-reference-compatible sequence remained common even after accounting for potential bacterial sequence similarity.

To assess whether the observed human-sequence positivity was specific to the more selectively defined exploratory cohort or persisted under broader sampling criteria, we analyzed three separately selected systematic replication cohorts (R1–R3), each initially comprising 900 runs selected without knowledge of human-contamination outcomes (Fig. 1b–d). Strict human-sequence positivity was detected in 677 of 883 evaluable runs (76.67%) in R1, 628 of 888 (70.72%) in R2 and 632 of 889 (71.09%) in R3. Following bacterial-sequence adjudication, 663 of 883 (75.08%), 621 of 888 (69.93%) and 623 of 888 (70.16%) runs remained positive, respectively. One R3 run had an unresolved adjudication and was excluded from that denominator. Across the exploratory and replication cohorts, human-sequence burden varied over several orders of magnitude, although median burdens among adjudicated-positive runs were similar (4.15–4.91 per million sequencing units; Fig. 1b; Supplementary Table 1).

The marked variation in human-sequence burden across runs prompted us to examine whether contamination occurred independently among sequencing runs or showed recurrence within BioProjects. To assess this, we constructed a contamination-blind, balanced panel of 214 BioProjects, each represented by three runs (642 runs), and compared the observed distribution of adjudicated-positive runs per project against expectations under run-level independence. Of these projects, 22 contained no positive runs and 114 were positive in all three, compared with approximately four and 84, respectively, expected under independence (Fig. 1c; Extended Data Fig. 1b). We further quantified project-associated clustering using a Bayesian random-intercept model, which yielded a BioProject intraclass correlation coefficient (ICC) posterior mean of 0.566 in E. Separately fitted models in the replication cohorts yielded ICC estimates of 0.600, 0.644 and 0.646 in R1–R3, respectively (Fig. 1d; Extended Data Fig. 1c,d). Together, these findings demonstrate pronounced BioProject-associated recurrence of human-sequence positivity, rather than independent occurrence across sequencing runs.

The observed BioProject-level clustering prompted us to investigate whether contamination was also associated with more specific sequencing contexts. We focused these analyses on Illumina datasets, for which original se-

quencing metadata can encode instrument, run, flowcell and lane identities. In particular, read-group metadata preserved in original Sanger Institute CRAM files enabled reconstruction of physical sequencing episodes and shared-lane contexts. Among six BioProjects comprising 296 runs with multiple archive-submission cohorts, submission-associated variation in human-sequence positivity and burden was primarily driven by one focal BioProject (Extended Data Fig. 2a,b). In the focal BioProject comprising 230 runs, sequencing episodes reconstructed from original read-group metadata explained 9.5% of variation in contamination occurrence, 14.5% of all-run log-transformed burden and 12.9% of positive-only burden, indicating additional technical structure without establishing the stage of contamination introduction (Extended Data Fig. 2c). Substantial within-project heterogeneity was also observed between two sequencing lanes, with 206 of 225 runs (91.6%) positive in one lane compared with 31 of 50 (62.0%) in another (Extended Data Fig. 2d). Conversely, comparisons across four shared-lane sequencing contexts, encompassing 1,085 analyzed runs from six BioProjects, revealed between-project heterogeneity despite shared physical sequencing lanes (Extended Data Fig. 2e). A systematic, year-matched comparison of HiSeq X and NovaSeq 6000 involving 340 runs from nine BioProjects found no significant platform difference in the magnitude of lane-associated effects on contamination burden ($P = 0.47$) or occurrence ($P = 0.50$; Extended Data Fig. 2f). Together, these findings indicate that sequencing episodes capture some contamination-associated variation, but neither shared physical lanes nor platform-specific lane effects provide a general explanation for the observed BioProject-associated heterogeneity.

Sequencing platforms differ in their technical characteristics, which have also evolved across successive instrument generations. If platform-associated processes contributed substantially to human-sequence contamination, these differences might be reflected in platform-specific variation or temporal changes in positivity. To examine these possibilities, we analyzed project-balanced annual estimates by sequencing platform and archive release year in an expanded dataset of 2,109 runs from 558 BioProjects (Fig. 1e; Extended Data Fig. 3). Fitted temporal slopes were negative across all three platform groups: −1.60 percentage points per year for Illumina (2011–2026; 95% BioProject-cluster-bootstrap CI, −2.35 to −0.88), −0.60 for Ion Torrent (2013–2026; 95% CI, −2.37 to +1.10) and −3.33 for BGI/MGI (2024–2026; 95% CI, −12.82 to +6.50). However, only the Illumina estimate showed a decline with a confidence interval excluding zero; estimates for Ion Torrent and BGI/MGI were less conclusive. These descriptive trends were subject to differences in project composition and temporal coverage, with 2026 representing a partial year, and could not be attributed specifically to instrument generations. Overall, human-sequence contamination remained common across the examined cohorts, with pronounced project-associated recurrence and limited evidence for temporal declines across sequencing platforms. We next asked how much human genetic information these fragments retained.

Contaminating human sequence retains distributed genetic information

Having established the prevalence and project-associated recurrence of human-sequence contamination, we next investigated the extent of human genetic information retained in bacterial sequencing reads. We first examined whether human-reference-compatible fragments were confined to a limited set of genomic regions or distributed more broadly. Across the exploratory cohort (E), 30 of 1,713 runs contained strict human fragments mapping to all 25 human reference sequences (22 autosomes, X, Y and mitochondrial DNA), including seven Ion Torrent runs. Among 100 high-burden Illumina paired-end runs selected on the basis of absolute human-fragment counts, 23 showed this breadth of representation (Extended Data Fig. 4a,b). Both the total number of recovered human fragments and the number of runs representing individual autosomes increased with chromosome length (R^2 = 0.73 and R^2 = 0.94, respectively; Extended Data Fig. 4c–e). These patterns indicate that the recovered human signal extends across the genome rather than being restricted to a few recurrently mapped regions, although chromosomal representation does not imply complete sequence coverage.

We next asked whether these dispersed fragments retained population-informative human genetic variation. Using single-nucleotide polymorphism (SNP) markers from the 1000 Genomes Project (1KGP)[10], selected by allele-frequency differences across five reference superpopulations (AFR, AMR, EAS, EUR and SAS), we evaluated read-derived allele states against the same five populations. Affinity was assessed using allele-frequency likelihoods, with support evaluated by resampling genomic 1-Mb blocks. Of the adjudication-evaluable runs in the systematic replication cohorts, 197 of 883 in R1, 169 of 888 in R2 and 159 of 888 in R3 met the marker-eligibility criteria. Reference-population affinity reached bootstrap support of at least 0.90 in 62, 53 and 70 runs, respectively, and at least 0.99 in 35, 30 and 38 runs (Fig. 2a). These counts use all adjudication-evaluable runs as denominators, including those without sufficient marker observations for scoring. Thus, a subset of bacterial sequencing datasets retained enough informative human sequence for reproducible affinity to broad reference-population allele-frequency profiles, although such affinity does not establish the ancestry or geographic origin

of an individual contributor.

To examine the potential for more extensive recovery of human genetic variation, we analyzed a separately selected set of 346 runs from 177 BioProjects using the five-population-ascertained SNP resource. Runs were selected conditionally on the presence of at least 20 clean autosomal SNP markers, with no more than five runs retained per BioProject. Within this set, 28 runs contained at least one classifiable recovered SNP-site configuration under more stringent per-site read-count criteria, distinguishing REF-only, ALT-only and two-allele-supported observations (Fig. 2b). Three selected high-recovery runs from a single BioProject contained 23,269–25,871 classifiable recovered 1KGP SNP-site configurations each, distributed across all 22 autosomes and 2,451–2,489 genomic 1-Mb blocks (Fig. 2c). These findings demonstrate that selected bacterial sequencing libraries can retain substantial, widely distributed human allele information, although the marker-conditioned selection precludes extrapolating these recovery frequencies to the wider bacterial sequencing archive.

We further examined whether human allele observations were concordant between sequencing runs belonging to the same BioProject. Across 21 within-project comparisons involving 26 runs from ten BioProjects, shared high-quality SNP loci allowed direct comparison of read-derived reference and alternate allele states (Fig. 2d). The strongest comparison involved 85,262 shared loci, of which 77,451 (90.84%) had matching allele states. Such concordance indicates that selected bacterial sequencing datasets can retain strongly overlapping partial human genetic profiles.

One possible explanation for recurrent human genetic signals across bacterial libraries is the redistribution of human-containing fragments during pooled sequencing, through index misassignment or other lane-associated carryover. We therefore examined whether libraries sharing a final sequencing lane exhibited similar contamination patterns and human reference-population affinities. In one HiSeq X Ten lane shared by three BioProjects, 127 of 188 libraries from one project were human-sequence-positive, compared with only 2 of 62 from the other two projects (Extended Data Fig. 5a,b). In a second shared-lane context comprising 348 runs from one BioProject, positivity differed across three bacterial specimen source-country groups (177 of 288, 37 of 46 and 14 of 14), while the marker-informative subsets exhibited differences in reference-population affinity across groups and marker thresholds (Extended Data Fig. 5c,d). These observations argue against a homogeneous lane-wide contamination signal, but neither exclude nonuniform cross-library carryover nor establish whether multiple runs contained human material from the same individual.

Finally, we investigated whether contaminating reads retained clinically annotated human variants. We selected 100 high-burden Illumina paired-end runs from cohort E on the basis of absolute strict human-fragment counts and screened their read-derived allele observations against ClinVar. High-quality human-sequence observations at ClinVar loci were detected in 96 runs, including alternate alleles in 61. Six runs contained candidate pathogenic or likely pathogenic (P/LP) alternate alleles, and all six also contained P/LP alternate-allele observations with computational read-pair support (Fig. 2e; Supplementary Note 2). These observations represent sequence-level evidence at clinically annotated loci, rather than validated individual genotypes or evidence of a clinical diagnosis.

Together, these analyses demonstrate that unintended human sequence in bacterial sequencing archives can retain information spanning genomic distribution, population-informative variation, partial allele profiles and clinically annotated variants. The extent of recovery varied among the analyzed datasets and was particularly pronounced in selected high-burden libraries. We therefore next examined whether human-read depletion could reduce these signals while preserving the microbial sequence information for which the libraries were generated.

Human-read depletion reduces recoverable human information while preserving bacterial assemblies

We next investigated whether the human genetic information recovered from bacterial sequencing libraries could be reduced through computational read depletion without substantially compromising bacterial genome analysis. We applied Hostile human-read depletion[11] to the same 100 high-burden Illumina paired-end libraries examined for clinically annotated variants and evaluated the original (PRE) and filtered (POST) datasets using identical human-sequence detection and genetic-information endpoints (Fig. 3a). Across the matched libraries, strict human fragments decreased from 899,063 PRE to 60 POST, corresponding to a reduction of more than 99.99%. Eighty of 100 POST libraries contained no residual strict human fragments (Fig. 3b).

We then examined whether this reduction was sufficient to prevent recovery of the human genetic-information

signals characterized above. Before depletion, 99% of libraries contained human fragments representing at least 22 chromosomes, 39% supported recoverable reference-population affinity, 96% contained high-quality observations at ClinVar loci and 61% contained ClinVar alternate alleles. Six libraries additionally contained candidate pathogenic or likely pathogenic (P/LP) alternate alleles with computational read-pair support. After human-read depletion with Hostile, none of these higher-level endpoints remained detectable (Fig. 3c). Median human chromosome breadth decreased from 24 to 0, while median population-informative marker counts declined from 68 to 0 (Fig. 3d). Although seven POST libraries retained at least one population-informative marker site, none retained sufficient information for recoverable reference-population affinity. Thus, substantial reductions in human-fragment burden were accompanied by the loss of all measured higher-level human genetic-information endpoints.

We further investigated the small amount of human sequence remaining after depletion. Residual fragments were concentrated in libraries with the greatest initial human-sequence burden: 50 of the 60 POST strict fragments occurred among the 25 highest-burden PRE libraries (Extended Data Fig. 6a,b). Nevertheless, none of the 20 libraries retaining strict human fragments met the measured population-affinity or clinically annotated variant endpoints (Extended Data Fig. 6c). To examine whether these findings extended beyond Illumina sequencing, we independently evaluated 10 high-burden BGI/MGI libraries and 10 Ion Torrent libraries. Strict human fragments decreased from 193,081 to 14 in BGI/MGI datasets and from 115,549 to 29 in Ion Torrent datasets, with no POST library retaining recoverable population affinity or ClinVar alternate-allele observations (Extended Data Fig. 6d). These analyses demonstrate substantial depletion across the tested sequencing platforms, while showing that sparse human-reference-compatible fragments can remain after filtering.

Having established the reduction in recoverable human information, we next assessed whether depletion preserved the bacterial sequence information for which the libraries were generated. We assembled paired PRE and POST datasets from ten high-burden Illumina libraries and compared assembly sequence and gene annotations. Median POST-to-PRE assembly alignment was 99.9%, average nucleotide identity (ANI) exceeded 99.99%, coding-sequence (CDS) retention was 99.7% and non-hypothetical CDS retention exceeded 99.9% (Fig. 3e). Regions represented in PRE but not POST assemblies accounted for a median of 0.8% of PRE assembly sequence, and approximately half of these PRE-only regions aligned to the human reference. No annotated bacterial CDS overlapped a PRE-only region of at least 500 bp (Extended Data Fig. 7a,b). Eight of ten POST assemblies contained no human-reference-aligned sequence, while long-scaffold sequence was largely retained and assembly N50 was equal to or greater than its PRE value in all ten comparisons (Extended Data Fig. 7c–e). BUSCO completeness was unchanged in eight assemblies and increased in two; further investigation showed that these apparent improvements reflected changes in automated genome assessment rather than recovery of previously missing bacterial genes (Extended Data Fig. 7f; Extended Data Fig. 8). Together, these findings indicate that human-read depletion preferentially removed a human-enriched assembly component while preserving the principal bacterial assembly and annotation features in the tested libraries.

Collectively, these experiments show that computational human-read depletion can sharply reduce recoverable human genetic information while largely preserving bacterial assembly and annotation in the tested WGS datasets. We next examined whether this mitigation approach extended to bacterial RNA-seq, where processed human transcripts present an additional source of potentially recoverable human sequence.

Bacterial RNA-seq contains genome-aligned and processed human RNA signals

We next investigated whether unintended human sequence was also detectable in bacterial transcriptomic datasets, where human RNA could introduce an additional class of contaminating sequence. We examined 153 bacterial RNA-seq runs representing 123 SRA Experiments, 117 BioSamples and 14 BioProjects. To distinguish different forms of human-sequence evidence, we assessed both reads compatible with the human genomic reference and reads supporting stringent human exon–exon splice junctions. Among the 123 Experiments, 74 contained genome-aligned human signal alone, 31 contained both genome-aligned and splice-junction signals, one contained splice-junction signal alone and 17 contained neither. Overall, genome-aligned human sequence was detected in 105 Experiments (85.4%), stringent splice-junction evidence in 32 (26.0%) and at least one signal class in 106 (86.2%; Fig. 4a). Genome-aligned human-sequence burden varied over several orders of magnitude, while splice-positive Experiments occurred across the burden distribution (Fig. 4b). Importantly, genome-aligned reads do not necessarily originate from genomic DNA, as genome-compatible human RNA may also contribute to this signal.

To establish whether the splice-junction observations reflected reproducible evidence of processed human RNA

rather than annotation-dependent alignment artifacts, we independently remapped reads from 32 splice-positive runs without supplying transcript annotations or known splice sites. Among 363 pair-testable splice-supporting read identifiers, 330 (90.9%) recovered at least one original target junction, while 343 of 373 (92.0%) testable run-by-junction observations were recovered exactly (Fig. 4c). High annotation-free recovery was observed across individual runs, supporting the reproducibility of the stringent splice-junction signal (Extended Data Fig. 10a,b). These findings indicate that bacterial RNA-seq datasets can retain human sequence derived from processed RNA in addition to genome-compatible human sequence.

We further examined whether these observations were concentrated in a small number of sequencing projects. Genome-aligned human sequence was detected in at least one Experiment from each of the 14 surveyed BioProjects, while stringent splice-junction evidence occurred in seven (Extended Data Fig. 9a,b). Human-sequence burden varied both between projects and among Experiments within the same project, with corresponding differences in the relative contributions of the two signal classes (Extended Data Fig. 9c,d). Thus, the observed human RNA-seq signal was distributed across multiple projects rather than being driven by a single unusually affected dataset.

Having established the presence of both signal classes, we next asked whether human-read depletion would remove them equally effectively. We applied the same genome-reference-based depletion approach to a selected 18-run bacterial RNA-seq stress-test cohort and independently assessed genome-aligned human information before (PRE) and after (POST) filtering. Strict human fragments decreased from 43,169 to just two, with only one POST run retaining detectable strict human-fragment evidence (Fig. 4d; Extended Data Fig. 10c). Recoverable reference-population affinity decreased from nine of 18 runs to none, while ClinVar alternate-allele observations and the measured pathogenic or likely pathogenic variant endpoints also fell to zero after depletion (Fig. 4d). Thus, human-read depletion reduced genome-aligned human sequence by more than 99.99% and eliminated the measured genetic-information endpoints in this stress-test cohort. Across these 18 libraries, human-read depletion retained 99.98% of the 383.9 million input read pairs (median per-library removal, 0.012%; range, 0.0007–0.116%).

Processed-RNA evidence, however, responded differently. Thirteen of the 18 stress-test runs contained stringent splice-junction evidence before depletion. Across these runs, splice-supporting read-pair observations decreased from 202 PRE to 63 POST, while stringent splice-junction observations declined from 207 to 87 (Fig. 4e). Despite these reductions, all 13 originally splice-positive runs remained splice-positive after filtering. Every surviving POST splice-supporting read pair and junction was also present in the corresponding PRE data, with no newly detected splice observations (Extended Data Fig. 10d–f). Thus, genome-reference-based depletion substantially reduced but did not eliminate the measured processed-RNA signal.

Together, these results extend the detection of unintended human sequence from bacterial WGS to bacterial RNA-seq and identify an important limitation of genome-reference-based depletion. Although filtering nearly eliminated genome-aligned human fragments and the measured genetic-information endpoints, stringent splice-junction evidence persisted in the tested RNA-seq libraries. Complementary transcriptome- or splice-aware screening may therefore be needed to more comprehensively minimize unintended human sequence in bacterial transcriptomic datasets.

Discussion

Our analyses show that unintended human sequence in bacterial isolate whole-genome sequencing is not confined to isolated datasets but recurs across projects and separately selected sequencing cohorts. Previous studies established distinct consequences of human-sequence contamination: human-derived fragments incorporated into microbial genome assemblies can propagate into erroneous protein annotations[6], whereas human SNPs, including clinically annotated variants, can be recovered directly from publicly archived bacterial sequencing reads[8]. Building on these observations, our study characterizes the project-associated structure and reproducibility of this exposure, its recoverable genetic information and its response to computational depletion. Substantial human-sequence positivity was reproduced across three contamination-blind replication cohorts, supporting the conclusion that this phenomenon is recurrent within the bacterial isolate datasets examined. Although the sampling designs do not yield an unbiased prevalence estimate for the entire public archive, the consistency across separately selected cohorts strengthens the evidence that unintended human sequence is a systematic consideration for microbial sequencing quality control.

A particularly notable finding was the strong association of human-sequence positivity with BioProject member-

ship. Substantial project-level clustering persisted after accounting for sequencing depth and was reproduced across the separately selected cohorts, suggesting that contamination is influenced by shared experimental or technical contexts rather than arising solely as independent, sporadic events. However, BioProject membership encompasses multiple potentially shared processes, including specimen handling, nucleic-acid extraction, library preparation and sequencing. Detailed analyses of reconstructed sequencing episodes revealed additional heterogeneity, while differences among libraries sharing the same final sequencing lane argued against homogeneous lane-wide carryover as a general explanation. Although cross-library misassignment is a recognized source of sequencing contamination[12], our observations do not support attributing the overall phenomenon to a single mechanism. Instead, they are consistent with heterogeneous workflow-associated contributions whose precise origins cannot be resolved from the publicly available archive metadata.

The implications of these findings extend beyond the integrity of microbial genome assemblies. In selected libraries, human-reference-compatible fragments were distributed across multiple chromosomes and retained population-informative markers, partial allele profiles and clinically annotated alternate alleles. Such observations establish that incidental human sequence can carry genetically interpretable information even when the intended sequencing target is bacterial. Tomofuji and colleagues demonstrated that residual human reads from gut metagenomic sequencing could support individual re-identification against matched genotype datasets[13]. Although that study examined host-associated samples in which human DNA is biologically expected, it establishes an important precedent for the information potentially recoverable from sparse human sequence. In sequencing archives, partial genetic profiles could potentially be linked to external genetic datasets, while publicly available study metadata, including submitter identities and institutional affiliations, might provide additional contextual information. Although we deliberately did not attempt contributor identification or re-identification, the possibility of such linkage underscores the privacy implications of unintended human-sequence exposure.

Importantly, the information-exposure concern appears substantially amenable to computational mitigation. In the selected bacterial WGS libraries, human-read depletion reduced strict human-fragment counts by more than 99.99% and eliminated the measured population-affinity and clinically annotated variant endpoints. Matched bacterial assembly comparisons showed near-complete preservation of bacterial sequence and coding content, indicating that substantial human-sequence reduction need not compromise the principal microbial genomic features examined. These observations support incorporating human-read depletion into routine pre-deposition processing, although residual human-reference-compatible fragments remained in some filtered libraries. Our findings parallel observations in host-associated metagenomic datasets, where more comprehensive human-reference filtering prevented patient re-identification from residual human genetic information[14]. The disappearance of our measured information endpoints should therefore be interpreted as substantial risk reduction rather than proof of complete anonymization or elimination of all possible genetic inferences.

Our extension to bacterial RNA-seq revealed that unintended human sequence can also include processed transcripts, introducing a distinct consideration for detection and depletion. Stringent exon–exon splice-junction evidence established the presence of processed human RNA, which cannot be reliably distinguished from genomic DNA on the basis of genome-aligned reads alone. In the selected RNA-seq mitigation cohort, genome-reference-based filtering nearly eliminated strict human-fragment evidence and the measured genotype-derived information endpoints while retaining approximately 99.98% of input read pairs. Nevertheless, stringent splice-junction evidence persisted in all originally splice-positive stress-test libraries, demonstrating that highly effective depletion of genome-aligned human sequence does not necessarily remove processed-RNA signals. This distinction suggests that transcriptome- or splice-aware screening may be needed for more comprehensive removal of unintended human sequence from bacterial RNA-seq.

Together, these findings place unintended human sequence at the intersection of sequencing quality control and genetic-information stewardship. The recurrence of human-compatible sequence across bacterial sequencing projects, its capacity to retain interpretable human genetic information and its persistence in publicly accessible raw reads indicate that assembly-level quality assessment alone is insufficient to address this concern. At the same time, substantial reduction of measured human-information endpoints with minimal disruption to the examined bacterial genomic features demonstrates that effective mitigation is feasible. We therefore propose that screening for unintended human sequence, appropriate depletion and verification of residual signals should become routine considerations before public deposition of nominally host-free microbial sequencing data. Integrating these safeguards into existing sequencing workflows would help preserve the scientific value of open datasets while reducing unintended disclosure of human genetic information.

Methods

Study design and analysis populations

Publicly archived bacterial isolate whole-genome sequencing (WGS) datasets were examined where human DNA was not an intended experimental component. Eligibility for the exploratory cohort E excluded direct-specimen sequencing, culture-free target-enrichment studies, metagenomic and microbiome datasets, ambiguous records associated with multiple BioSamples, and host-cell-dependent workflows in which host DNA could be present. Eligibility and scope review of 1,733 candidate records yielded 1,713 runs from 426 BioProjects and 1,703 BioSamples. Cohort-specific sampling frames and eligibility procedures were used rather than a single common retrospective filter.

To assess reproducibility, three separately selected replication cohorts (R1–R3) were constructed using selection criteria independent of human-sequence detection outcomes. Each began with 900 selected run memberships from 300 BioProjects. Runs with unavailable, failed or unresolved core processing or assembly-based adjudication were excluded from the corresponding evaluable denominators rather than classified as negative. Cohort-specific results were analysed separately because selection designs differed and run or BioProject membership could overlap. In unique-run sensitivity analyses, duplicate memberships were assigned to the earliest core-evaluable occurrence in fixed R1, R2, R3 order, without using human-information outcomes.

A separate temporal and platform extension of the cohort E contributed 396 analysed runs from 399 selected records, yielding a combined population of 2,109 runs from 558 BioProjects for the extended E cohort. Selection targeted approximately ten BioProjects per platform family and archive-release year, generally with three runs per newly selected project, subject to archive availability. Archive `first_public` dates denote public sequence release rather than sequencing dates. Targeted sequencing-context studies were analysed separately and were not treated as replication or archive-wide prevalence cohorts.

Bacterial RNA-seq was evaluated separately. Of 157 completed runs from 15 BioProjects, four runs belonging to a project using animal-tissue-derived brain–heart infusion medium were excluded during scope review. The final survey comprised 153 runs, 123 SRA Experiments, 117 BioSamples and 14 BioProjects. The survey was not designed to estimate prevalence across all bacterial RNA-seq archives.

Human-reference alignment, fragment definition and adjudication

Sequencing reads were aligned to the T2T-CHM13 v2.0 human reference with a masked Y chromosome and the revised Cambridge Reference Sequence (rCRS) for mitochondrial DNA. Short-read alignment was performed using minimap2 with the `-ax sr` preset[15]. A strict paired-end human fragment required both mates to have primary alignments with mapping quality (MAPQ) $\geq$30, aligned query coverage $\geq$80% and sequence identity $\geq$98%, together with proper-pair placement on the same reference sequence. Unmapped, secondary and supplementary alignments were excluded. Query coverage was defined as the proportion of read bases participating in the alignment, including insertions, while sequence identity was calculated from the alignment span and edit distance (NM), accounting for mismatches and insertions/deletions. Qualifying paired fragments were deduplicated by reference sequence and outer alignment-coordinate span. Single-end libraries were evaluated using the corresponding strict single-read criteria and reference-coordinate deduplication. Human-sequence burden was expressed per million input read pairs for paired-end libraries and per million input reads for single-end libraries.

A run was classified as strict positive if it contained at least one qualifying human fragment or read. In the exploratory cohort E, candidate human reads were additionally classified using Kraken2[16] with the PlusPFP-16 database (9 December 2022 release) and evaluated against the corresponding exact bacterial assembly, where available. A qualifying bacterial match required $\geq$80% query coverage and $\geq$95% sequence identity for at least one mate or read. Under the exploratory-cohort adjudication definition, strict-positive runs with qualifying bacterial matches (`BACTERIAL_MATCH_PRESENT`) were excluded from the adjudicated-positive category; eight strict-positive runs with unsuccessful assembly checks remained classified as adjudicated positive. Exact-assembly-supported positivity required strict positivity and a completed `PASS` assembly challenge, indicating no qualifying bacterial match. Missing exact assemblies and unsuccessful challenges were not considered `PASS`. Replication cohorts retained their separately defined assembly-adjudication and evaluability criteria.

Project recurrence and clustering models

The E recurrence analysis used 214 BioProjects (642 runs) drawn from six sampling strata, each contributing exactly three eligible runs from distinct BioSamples on the same sequencing platform; a common bacterial species was not required. Selection of the three-run panel did not depend on human-sequence outcomes. Complete adjudication-evaluable triplets were used for descriptive recurrence analyses in R1 (293 projects, 879 runs), R2 (296 projects, 888 runs) and R3 (295 projects, 885 runs). For each cohort, the observed distribution of projects containing zero to three positive runs was compared with a binomial expectation assuming independent runs with a common cohort-wide positivity probability, irrespective of BioProject membership. Expected counts were calculated as $N\binom{3}{k}p^k(1-p)^{3-k}$, where k denotes the number of positive runs per project, p is the observed positivity fraction among complete triplets and N is the number of eligible projects.

BioProject clustering was estimated using separately fitted Bayesian Bernoulli-logit mixed-effects models with a BioProject random intercept. The E model used the 642-run panel. Replication models retained all adjudication-evaluable memberships, including incomplete triplets: 883 runs/295 projects (R1), 888/296 (R2) and 888/297 (R3). Models used Bambi v0.20.0, PyMC v6.3.1 and ArviZ v1.3.0, with four chains, 1,000 tuning iterations, 2,000 retained draws per chain and a target acceptance probability of 0.95. The effective E priors were Normal$(0, 1.5)$ for the intercept and HalfNormal(2.5) for the BioProject-level standard deviation; an E-only sensitivity used HalfNormal(1) for the latter. Divergent transitions, $\widehat{R}$ and effective sample sizes were examined.

For posterior project-standard-deviation draws σ, latent-scale intraclass correlation coefficients (ICCs) were calculated as $\sigma^2/(\sigma^2 + \pi^2/3)$ and median odds ratios (MORs) as $\exp\{\Phi^{-1}(0.75)\sqrt{2}\,\sigma\}$. Extended Data Fig. 1c summarizes ICCs using posterior means and 95% equal-tail intervals; Extended Data Fig. 1d reports posterior median MORs and equal-tail intervals. Model-summary tables separately report posterior means and 95% highest-density intervals, as labelled.

Depth-adjusted sensitivity analyses were restricted to the E panel. Sequencing depth was calculated as input read pairs for paired-end libraries or reads for single-end libraries, $\log_{10}$-transformed and standardized using the population standard deviation across the panel. Mean read length was similarly standardized. The final sensitivity model separated project-mean log-depth from within-project deviations and included these terms and mean read length as fixed effects alongside the BioProject random intercept.

Temporal and technical-context analyses

Temporal summaries used run-weighted or BioProject-balanced estimates as specified. Figure 1e fitted strict-positive project-year fractions with equal project-year weighting and obtained uncertainty intervals from 5,000 BioProject-cluster bootstrap resamples, retaining all observed years for each sampled project. Extended Data Fig. 3 used separate descriptive ordinary least-squares fits to equally weighted annual percentages over 2011–2026 (Illumina), 2013–2026 (Ion Torrent) and 2024–2026 (BGI/MGI). Partial-year 2026 data were not annualized.

For selected Sanger Institute datasets, available submitted CRAM read-group and platform-unit fields were used to reconstruct instrument, run, flowcell, lane, date-field and plex contexts. These were distinguished from archive-submission metadata. In the targeted shared-lane case study (Extended Data Fig. 2e), contamination results were available for 1,085 of 1,511 technically eligible libraries across four reconstructed contexts; libraries outside the analysed outcome snapshot were not counted as negative.

Sequencing-context comparisons used panel-specific tests. Extended Data Fig. 2a–c reported permutation-based R^2 contrasts (50,000 permutations in the retained specifications). Extended Data Fig. 2d used Fisher's exact and Mann–Whitney tests, including a plex-balanced sensitivity comparison. Extended Data Fig. 2e used Fisher's exact or chi-square tests and Mann–Whitney or Kruskal–Wallis comparisons as applicable; Holm correction was restricted to designated pairwise comparisons. Extended Data Fig. 2f used its specified permutation and false-discovery-rate procedures. Extended Data Fig. 5 included plex-stratified Cochran–Mantel–Haenszel comparisons. Corrections for multiple testing were applied only to the comparisons for which they were specified.

Chromosomal distribution and population-affinity endpoints

Human chromosomal representation required at least one qualifying strict fragment or read on a reference sequence, rather than a minimum fraction of bases covered. The analysis considered 25 reference sequences (22 autosomes, X, Y and mitochondrial DNA). Extended Data Fig. 4 calculated chromosome-level densities per non-N reference megabase and fitted reference-length regressions across the 22 autosomes only. Exploratory-cohort chromosome totals combined paired fragments and single reads and were not normalized to common read-layout or read-length units.

Population-informative SNPs in CHM13 coordinates were derived from 1000 Genomes Project reference data[10]. Biallelic autosomal SNPs were selected using the maximum pairwise allele-frequency difference (ΔAF) across five reference superpopulations (AFR, AMR, EAS, EUR and SAS). Markers with $\Delta AF \geq$0.10 were used for the replication-cohort population-affinity analysis (Fig. 2a), while thresholds of 0.10, 0.20 and 0.30 were evaluated for the sequencing-context comparisons (Extended Data Fig. 5d). Population affinity was scored against the same five reference populations using allele-frequency likelihoods. Clean REF or ALT observations required mapping quality (MAPQ) and base quality $\geq$30, with each locus contributing at most once to the likelihood calculation; mixed and other allele states were excluded. Allele frequencies were derived from reference-population allele counts and denominators, retaining genuine zero frequencies while excluding sites with undefined required frequencies. For likelihood scoring, frequencies were bounded within [0.005, 0.995]. Marker observations for the replication analysis were obtained through read remapping, whereas SRAFEGUARD used strict human alignments and the five-population marker subset with $\Delta AF \geq$0.30.

For Figure 2a, marker eligibility required at least 20 clean informative loci per run. Within each run, 1,000 bootstrap replicates resampled observed chromosome-qualified 1-Mb genomic blocks with replacement, preserving the number of blocks. Bootstrap computation required at least ten loci across two blocks, while the reported affinity endpoint additionally required the 20-locus eligibility criterion. Population support was defined as the largest fraction of bootstrap replicates favoring any one of the five reference populations, which did not necessarily correspond to the full-data maximum-likelihood population. Support thresholds of $\geq$0.90 and $\geq$0.99 were evaluated, with proportions reported relative to the adjudication-evaluable runs unless otherwise specified.

SRAFEGUARD population-affinity recoverability was defined separately using at least 20 clean informative loci per run and 1,000 within-run genomic-block bootstrap replicates. Recoverability required bootstrap support $\geq$0.90 for the reference population with the highest full-data likelihood. Extended Data Fig. 5d instead pooled qualifying observations by source-country cohort and used 5,000 bootstrap replicates resampling run-by-chromosome-qualified 1-Mb genomic blocks with replacement. No minimum of 20 informative loci per individual run was imposed for this analysis. Population-affinity scores describe resemblance to reference-population allele frequencies rather than individual ancestry, ethnicity, nationality or identity.

Recovery and comparison of human allele observations

To assess the recovery of human SNP information, a separate analytical set was assembled from previously processed WGS datasets spanning the exploratory cohort and targeted technical and pipeline-validation analyses. Candidate runs required at least 20 strict human-sequence units and available retained sequencing evidence, yielding 853 runs for marker extraction. Using the five-population SNP resource described above, 622 runs contained at least 20 clean autosomal SNP marker loci. Retaining no more than five eligible runs per BioProject, ranked by absolute strict-human-fragment count, yielded a final set of 346 runs from 177 BioProjects. Marker recovery was distinguished from a more stringent per-site SNP-configuration endpoint. Qualifying configurations required at least four REF observations and no ALT observations, at least four ALT observations and no REF observations, or at least two observations of each allele with an ALT fraction of 0.25–0.75. A separate sensitivity analysis applied minimum read depth of 10 and genotype quality of 30.

Within-project allele-state concordance was calculated among loci with clean REF or ALT states in both runs as the fraction with matching states, requiring at least 20 shared loci. Comparisons used read-derived allele states rather than validated diploid genotypes, and a likely technical-duplicate comparison was excluded from the displayed set.

ClinVar screening and read-pair-supported alternate alleles

Clinically annotated human variation was screened against a CHM13-coordinate ClinVar biallelic SNV resource dated 13 March 2022. Analysis was restricted to loci supported by strict human alignments with mapping quality (MAPQ) and base quality $\geq$30, applying a predefined genomic hotspot-exclusion mask. High-quality ClinVar locus observations, ClinVar alternate-allele (ALT) observations, candidate pathogenic or likely pathogenic (P/LP) ALT observations and read-pair-supported P/LP ALT observations were evaluated as separate endpoints. Candidate P/LP classification used the recorded ClinVar clinical-significance categories, excluding conflicting, uncertain, benign, risk-factor, protective and drug-response classifications. ClinVar review-star ratings were not used as an additional filter.

For paired-end libraries, read-pair-supported P/LP ALT evidence required an ALT-carrying read to retain the alternate allele at the expected genomic locus during physical-pair remapping, with MAPQ $\geq$30, proper-pair placement and the highest alignment score. The competing-alignment search considered up to 50 secondary alignments with a minimum score ratio of 0.5. No detected alignment at a distinct locus outside a 500-bp local start-coordinate window on the same chromosome could have an equal or higher score. This procedure evaluated the genomic placement of the observed allele rather than establishing a validated donor genotype, clinical diagnosis or exhaustive exclusion of alternative alignments. Read-pair support was assessed for paired-end Illumina and BGI/MGI libraries; Ion Torrent single-end libraries underwent the corresponding candidate single-read screen without a pair-supported endpoint.

Human-read depletion and bacterial assembly preservation

Human-read depletion was performed using Hostile v2.0.2[11] with the `human-t2t-hla` index and the recorded short-read filtering configuration. Paired-end WGS and RNA-seq libraries were processed using Bowtie2 paired-end alignment[17]. The depletion index was distinct from the masked-Y/rCRS human reference used for independent assessment of residual human sequence. The principal WGS mitigation set comprised the 100 paired-end Illumina libraries from cohort E with the highest absolute strict-human-fragment counts, representing 60 BioProjects. Additional stress tests included ten paired-end BGI/MGI libraries and ten single-end Ion Torrent libraries.

Original (PRE) and Hostile-filtered (POST) datasets were independently evaluated using SRAFEGUARD v0.1.0 with consistent endpoint definitions. Measurements included strict human-fragment burden, chromosomal representation, high-quality population-informative SNP observations, reference-population affinity, ClinVar ALT observations, candidate pathogenic or likely pathogenic (P/LP) ALT observations and, where applicable, read-pair-supported P/LP ALT observations. Population-affinity evaluation used the five-population SNP ascertainment and scoring procedures described above. SRAFEGUARD's independent detection and auditing procedures are detailed in Supplementary Note 1.

Bacterial assembly preservation was evaluated in ten high-burden paired-end Illumina libraries from six BioProjects, selected based on strict-human-fragment-count rank. PRE and POST assemblies were generated using SPAdes v4.3.0[18] with `--isolate --only-assembler`. PRE assemblies used the original paired-end reads, whereas POST assemblies used the corresponding Hostile-filtered reads.

Whole-assembly alignment was performed using dnadiff from the MUMmer suite[19], with the PRE assembly as reference and the POST assembly as query. Average nucleotide identity (ANI) was calculated separately using FastANI[20] in both PRE-to-POST and POST-to-PRE directions. The POST-to-PRE aligned fraction was normalized to total POST assembly bases. Assembly contiguity and N50 were calculated using scaffolds $\geq$500 bp, whereas long-sequence retention was calculated from total sequence on scaffolds $\geq$5 kb. Coding-sequence (CDS) count retention was calculated as $100 \times N_{\mathrm{POST}}/N_{\mathrm{PRE}}$, using independently annotated CDS counts rather than sequence-level matching of individual CDSs. Annotation was performed using Bakta v1.12.0[21], with the full database v6.0 (24 February 2025 release) and a minimum contig length of 500 bp. Genome completeness was assessed using BUSCO v6.1.0[22] with `bacteria_odb12.2`.

PRE-only intervals were defined as regions of complete PRE assemblies not covered by the corresponding POST assembly alignment. Intervals $\geq$500 bp were evaluated for human-reference alignment and overlap with PRE CDS annotations. For two assemblies showing apparent increases in BUSCO completeness after depletion, sensitivity analyses progressively excluded shorter scaffolds without reassembly and compared gene-prediction outcomes under genetic codes 4 and 11 using unchanged PRE sequences (Extended Data Fig. 8).

RNA-seq human-fragment and splice-junction analyses

Genome-alignable human fragments in bacterial RNA-seq were quantified using the paired-end strict-human-fragment procedure described above. An SRA Experiment was considered positive for a signal class if any associated run met the corresponding criteria. Experiment-level genomic burden was calculated as $10^6 \sum(\text{strict fragments}) / \sum(\text{input read pairs})$.

Processed human RNA was assessed using HISAT2 v2.2.3[23] with GRCh38 and known splice sites from GENCODE release 50[24]. Stringent junction evidence required primary split alignments with mapping quality (MAPQ) $\geq$30, adjacent CIGAR M/=/X segments of at least 20 bases on each side of the intron, aligned-query coverage $\geq$80% and sequence identity $\geq$98%, calculated from aligned query length and the `NM` edit-distance tag. Splice-boundary coordinates were required to match an annotated junction, with agreement of the HISAT2 `XS` strand where present; missing `XS` was permitted. Proper-pair status was not required. Read identifiers (QNAMEs), distinct junctions and run-by-junction observations were quantified separately.

Survey runs with stringent splice-junction evidence underwent annotation-free validation using a separately constructed genome-only HISAT2 index, paired-end reads and no supplied known splice sites. Validation was restricted to read identifiers for which the original read pair was available. Exact junction recovery required matching reference chromosome and left and right splice-boundary coordinates, without requiring matching strand tags.

An RNA-seq depletion stress-test cohort was selected from genome-positive libraries with recoverable population affinity or ClinVar ALT evidence. Libraries with PRE splice-junction evidence were reassessed after Hostile filtering using the annotation-guided HISAT2 and junction-quality-control procedures. POST read identifiers and junction-coordinate observations were compared with their corresponding PRE observations within each run.

RNA-seq read-pair retention was calculated from Hostile filtering records, converting counts of individual reads to read pairs after verifying paired-read filtering semantics and balanced mate counts. PRE input-pair counts were cross-checked against the RNA-seq survey records. Per-run read removal was calculated as $100 \times (1 - N_{\mathsf{POST}}/N_{\mathsf{PRE}})$, with aggregate removal calculated from the corresponding sums across runs. Read-pair retention was not interpreted as evidence of preserved bacterial gene-expression profiles.

Statistical reporting

Counts and proportions were summarized at the run, SRA Experiment or BioProject level as specified for each analysis, with missing and unevaluable observations excluded from their relevant denominators. Prevalence intervals were calculated using Wilson or BioProject-bootstrap methods as indicated for each cohort. Posterior intervals, temporal cluster-bootstrap intervals and population-affinity bootstrap support were reported according to their respective definitions above. Complete-triplet independence expectations were descriptive. All sequencing-context analyses were observational and did not establish the causal stage of human-sequence introduction.

Data and code availability

The bacterial sequencing datasets analyzed in this study are publicly available through the NCBI Sequence Read Archive (SRA) and European Nucleotide Archive (ENA), subject to the availability and withdrawal status of individual records. Aggregate results and plot-level measurements are provided in the accompanying Source Data and Supplementary Information.

The computational pipeline for identifying, quantifying and characterizing human-sequence contamination in bacterial sequencing archives is publicly available at https://github.com/Mahdi-Moosa/wgs-conta-pipeline. SRAFEGUARD, the tool for auditing population-informative and clinically annotated genetic information in contaminating human sequences, is publicly available at https://github.com/Mahdi-Moosa/srafeguard. Both repositories provide source code, installation instructions, software dependencies and methodological documentation.

To reduce direct linkage of unintended human genetic information to public sequencing records, original privacy-sensitive run/BioProject selection rosters, random seeds, accession-to-pseudonym mappings and run-level

genetic observations are retained privately. The documented configurable selection and analysis procedures are intended to support independently sampled, outcome-blind cohorts, rather than exact reconstruction of privately held historical cohort memberships.

Use of generative AI tools

Generative AI tools, including ChatGPT and Codex (OpenAI), were used to assist with computational workflow development, code generation and debugging, methodological exploration, and manuscript drafting and editing. AI-assisted computational procedures were reviewed and validated through software testing and comparison of analytical outputs with source data and independently generated summaries. All reported results were generated using computational procedures described in the Methods, with analytical definitions, validation criteria and software implementations documented to support reproducibility. Study design, scientific interpretation, final analytical decisions and responsibility for the manuscript remained with the author.

Author contributions

M.M.M. conceived and designed the study, developed and validated computational workflows, performed and interpreted the analyses, prepared the figures, and wrote and revised the manuscript.

Competing interests

The author declares no competing interests.

Ethics declaration

Not applicable.

Funding

This research received no specific external grant funding. Partial support for computational expenses was provided by BRAC University through an RSGI grant (BRACURSGI25029).

Acknowledgements

M.M.M. acknowledges the Department of Biotechnology, BRAC University, for providing course releases during the Summer and Fall 2026 semesters.

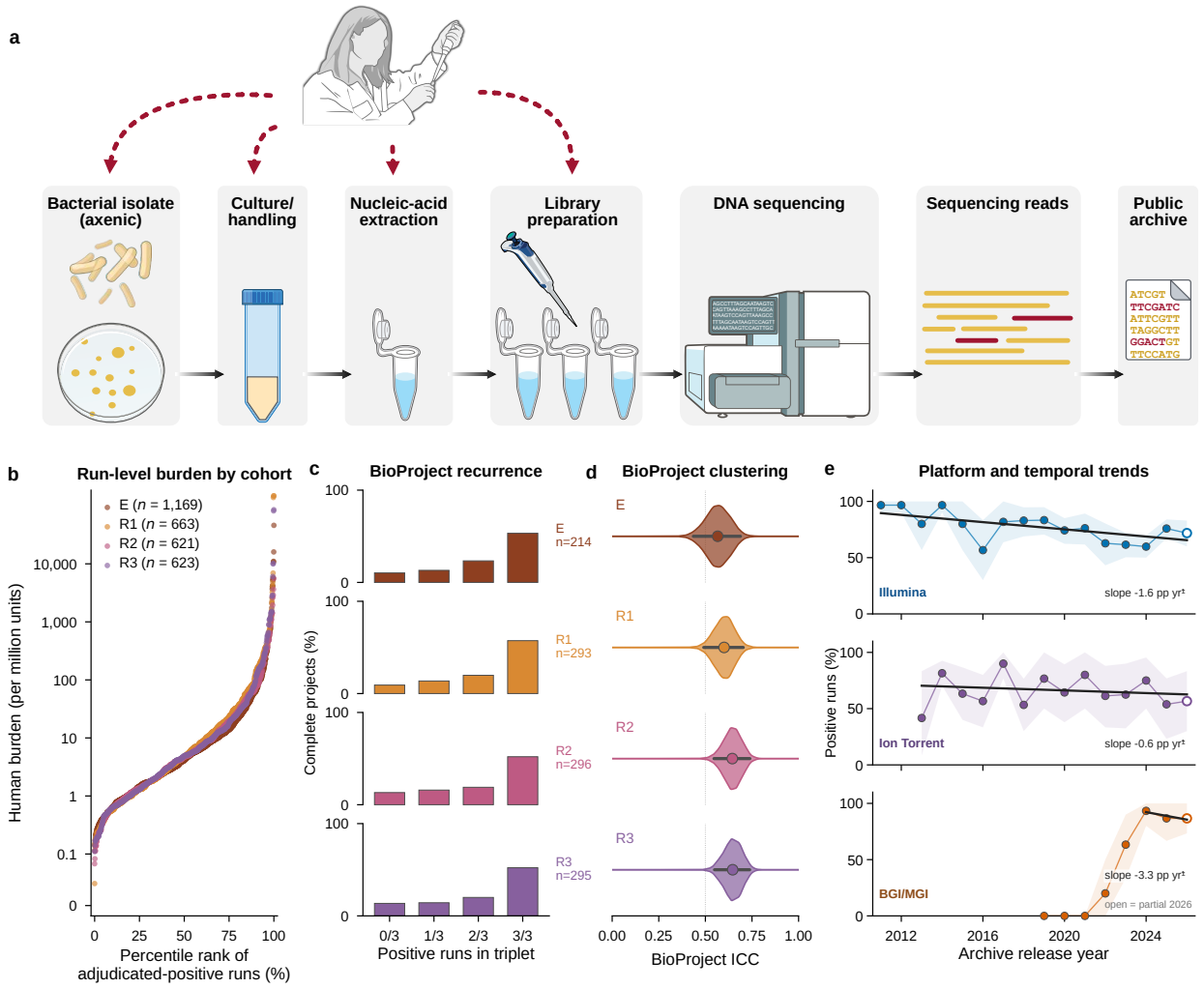


Fig. 1 | Human-sequence contamination is prevalent in bacterial isolate WGS and clusters within BioProjects. **a,** Conceptual bacterial isolate workflow showing possible points at which unintended human material could enter the sequencing workflow and be deposited in the read archive with microbial reads; the schematic does not identify a contamination mechanism. Public-domain scientific illustrations adapted from the NIAID NIH BioArt Source were used in the schematic. **b,** Distribution of human-sequence burden among adjudicated-positive bacterial isolate WGS runs in the exploratory cohort E and three systematic replication cohorts R1–R3 (n=1,169, 663, 621 and 623, respectively). Curves show human-sequence burden across percentile ranks of positive runs. Burden is expressed as human sequences per million quantification units (read pairs for paired-end libraries; reads for single-end libraries). **c,** Recurrence of adjudicated human-sequence positivity among BioProjects with three selected and evaluable sequencing runs in exploratory cohort E (214 BioProjects) and systematic replication cohorts R1 (293), R2 (296) and R3 (295). Bars indicate the percentage of BioProjects containing zero, one, two or three positive runs. **d,** BioProject-level clustering of human-sequence positivity in E and R1–R3, estimated using separate Bayesian random-intercept logistic models for each cohort. Violins show the posterior distributions of the BioProject intraclass correlation coefficient (ICC), points indicate posterior means, and horizontal segments denote 95% highest-density intervals. Posterior mean ICCs were 0.566, 0.600, 0.644 and 0.646 for E, R1, R2 and R3, respectively. **e,** Platform-specific temporal patterns in project-balanced human-sequence positivity by archive release year, analyzed in an expanded dataset of 2,109 runs from 558 BioProjects. Points indicate annual project-balanced positive fractions, with shaded regions representing 95% confidence intervals. Lines show temporal trends fitted to BioProject-level observations. Estimated slopes were -1.60 percentage points per year for Illumina (2011–2026), -0.60 for Ion Torrent (2013–2026) and -3.33 for BGI/MGI (2024–2026). Slopes were estimated with BioProject-cluster-bootstrap uncertainty and are distinct from the annual-summary fits reported in Extended Data Fig. 3. Open symbols denote partial-year observations for 2026.

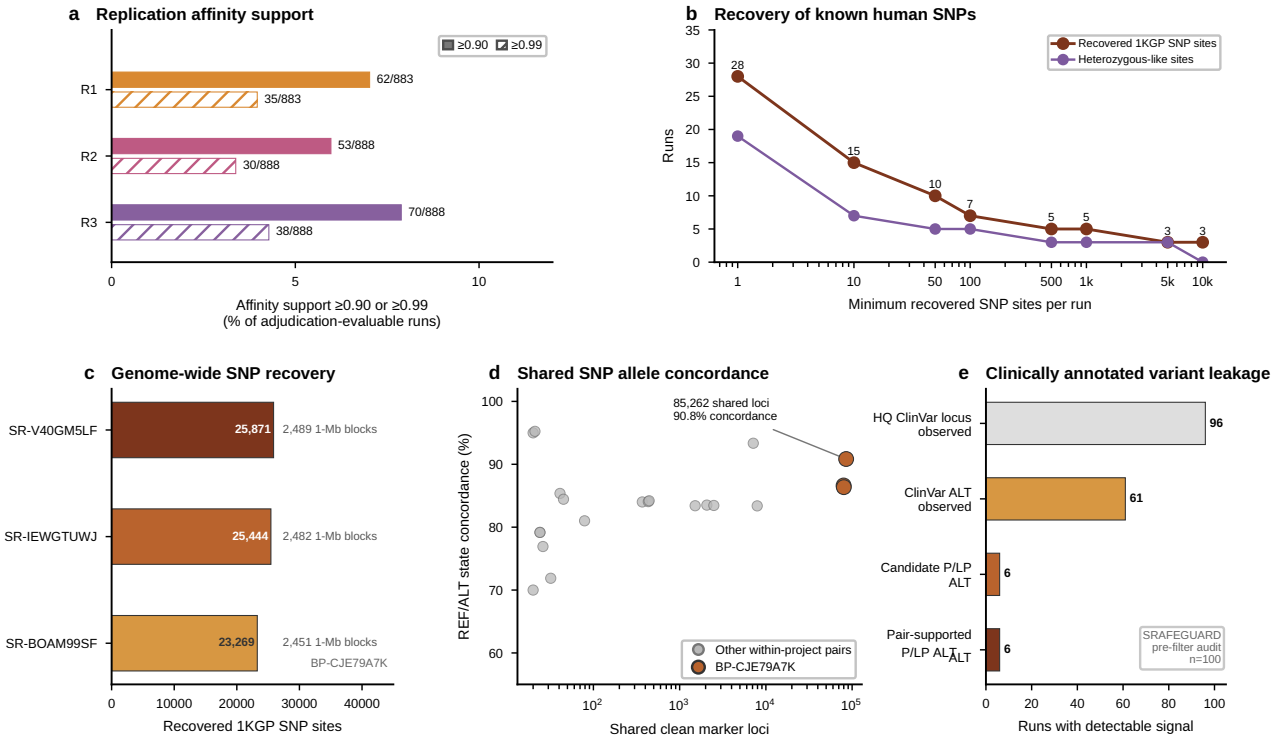


Fig. 2 | Contaminating human sequence retains population, genotype-like and clinically annotated information. a, Population-affinity information recovered from three systematic replication cohorts (R1–R3). Bars show the percentage of adjudication-evaluable runs with bootstrap support of at least 0.90 or 0.99 for their strongest affinity to one of five 1000 Genomes Project (1KGP) reference populations[10]. Affinity indicates resemblance to broad reference-population allele frequencies and does not establish individual ancestry, ethnicity or geographic origin. **b,** Conservative recovery of partial human genetic information in a separately selected historical analysis of 346 contamination events from 177 BioProjects. Curves show the number of runs meeting increasing thresholds of recovered 1KGP SNP sites and heterozygous-like observations. **c,** Genome-wide recovery of known human SNPs in three selected runs, showing recovered SNP-site counts and the number of 1-Mb genomic blocks represented. Labels use publication pseudonyms to reduce privacy exposure. Recovered SNP sites spanned all 22 autosomes; see Extended Data Fig. 4. **d,** REF/ALT allele-state concordance across clean SNP loci shared between pairs of contaminated bacterial sequencing runs from the same BioProject. Highlighted comparisons involve the BioProject illustrated in panel c; other within-project comparisons provide additional context. The highlighted comparison shared 85,262 loci, of which 77,451 had concordant and 7,811 discordant allele states (90.84% concordance). These results demonstrate that selected bacterial sequencing datasets can retain concordant partial human allele profiles. **e,** Clinically annotated human-variant information detected in a selected 100-run high-burden cohort E subset, analyzed as a privacy stress test. Bars show the number of runs with at least one high-quality observation at a ClinVar locus, an observed ClinVar alternate (ALT) allele, a candidate pathogenic/likely pathogenic (P/LP) ALT allele, or a read-pair-supported P/LP ALT allele. Read-pair support represents computational sequence evidence rather than clinical validation.

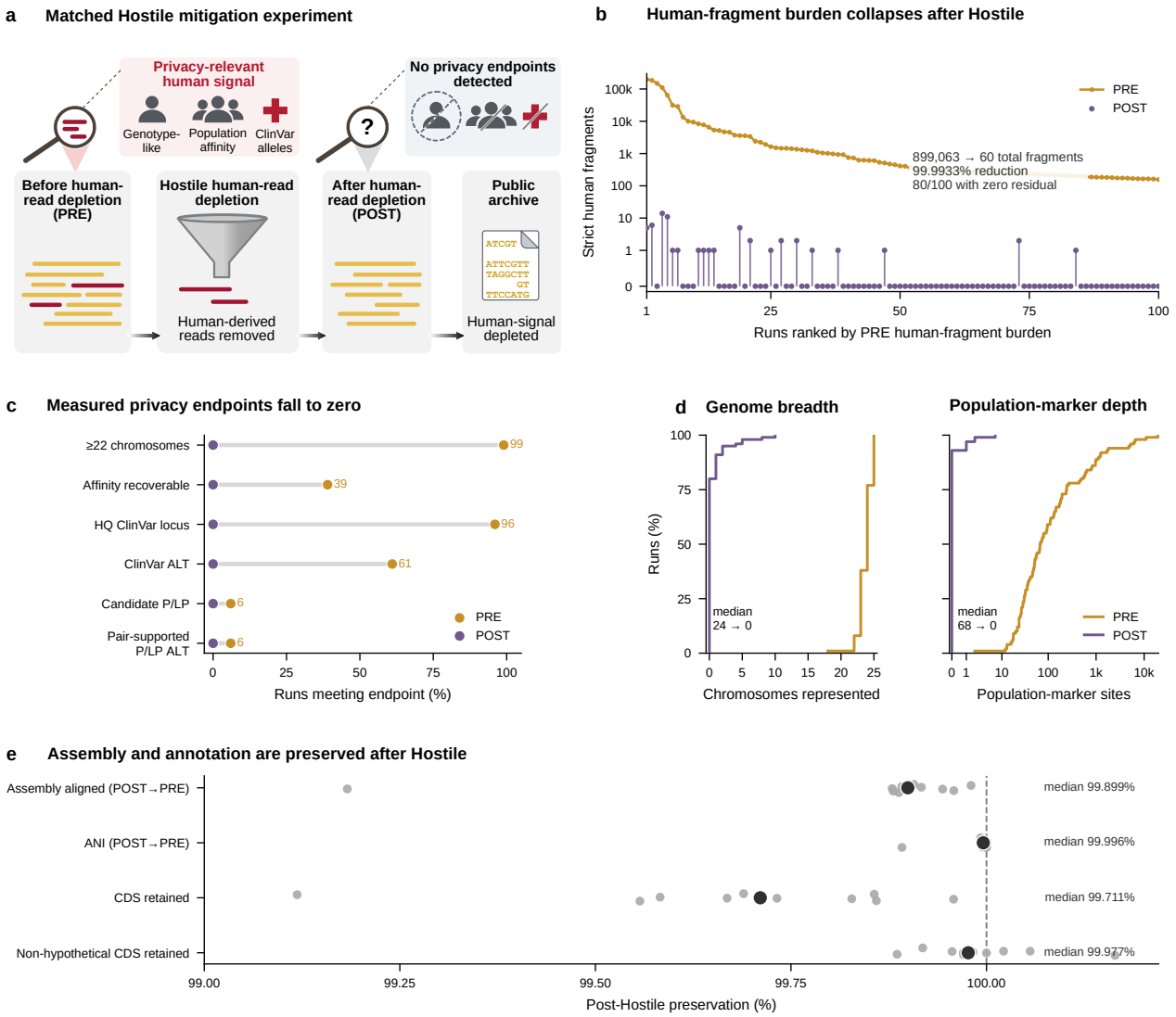


Fig. 3 | Human-read depletion sharply reduces recoverable human genetic information while preserving bacterial assembly and annotation metrics. **a,** Schematic of the human-read mitigation experiment using Hostile[11]. The high human-sequence burden 100 paired-end Illumina WGS runs from cohort E (see Fig. 2e) were evaluated before (PRE) and after (POST) Hostile human-read depletion using identical SRAFEGUARD contamination and privacy endpoints. **b,** Strict human-fragment counts before and after Hostile, with runs ranked by their PRE human-fragment burden. **c,** Recovery of six privacy-relevant endpoints before and after depletion: representation of at least 22 human chromosomes, recoverable population affinity, high-quality (HQ) ClinVar-locus observations, ClinVar alternate (ALT) alleles, candidate pathogenic/likely pathogenic (P/LP) ALT alleles and read-pair-supported P/LP ALT alleles. Points show the percentage of runs meeting each endpoint. All six endpoints fell to zero POST. Read-pair support represents computational sequence mapping evidence. **d,** Human genomic breadth and population-marker depth before and after depletion. Empirical cumulative distributions show the number of human chromosomes represented per run (left) and population-informative marker sites per run (right). Median chromosome breadth decreased from 24 to 0 and median marker depth from 68 to 0. Seven POST runs retained some population-marker observations, but none met the defined population-affinity endpoint. **e,** Preservation of bacterial assembly and annotation metrics in a selected 10-run high-burden stress-test panel. Gray points indicate individual runs, black points indicate medians and the dashed vertical line marks the PRE reference value (100%). Median POST-to-PRE assembly alignment was 99.899%, average nucleotide identity (ANI) was 99.996%, coding-sequence (CDS) retention was 99.711% and non-hypothetical CDS retention was 99.977%. These metrics demonstrate high preservation of the measured bacterial assembly/annotation features.

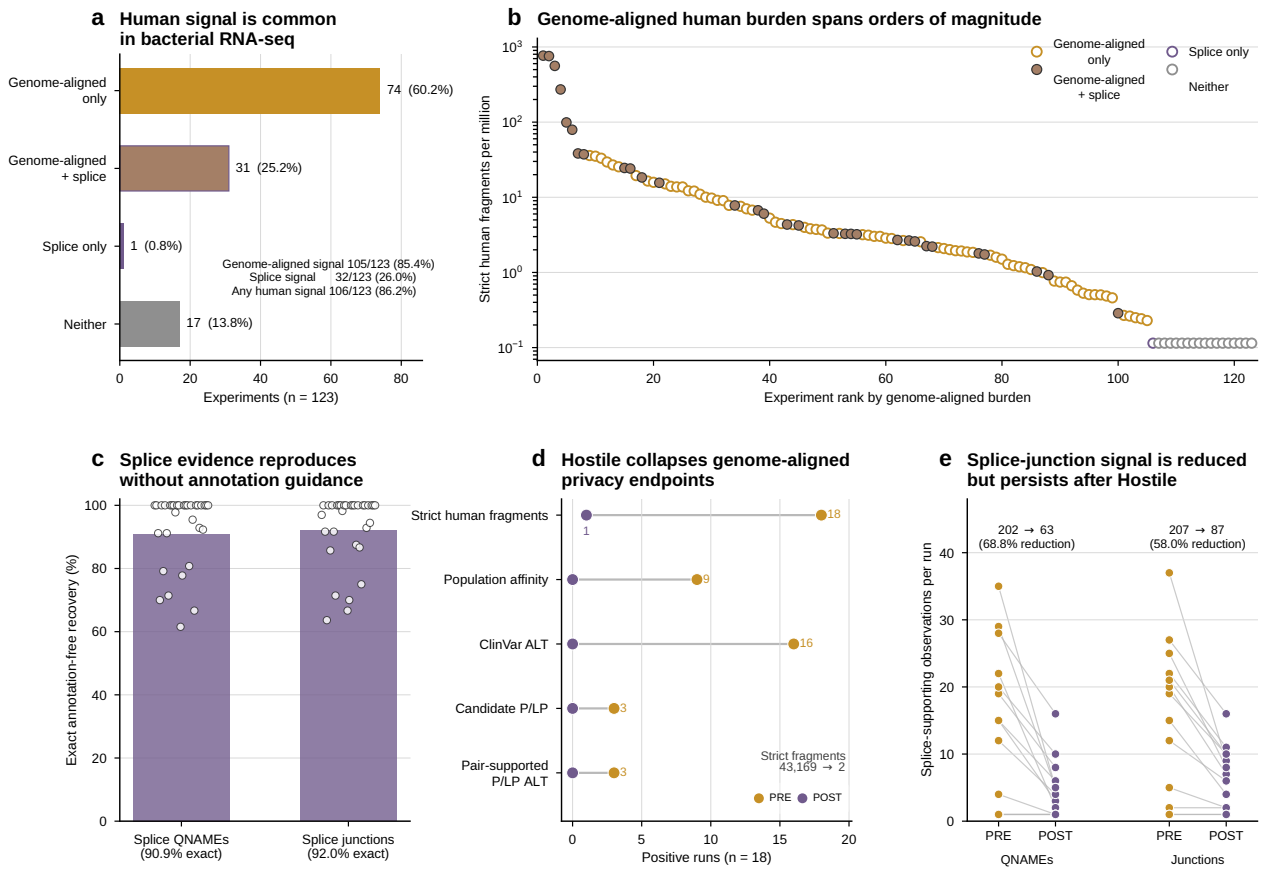
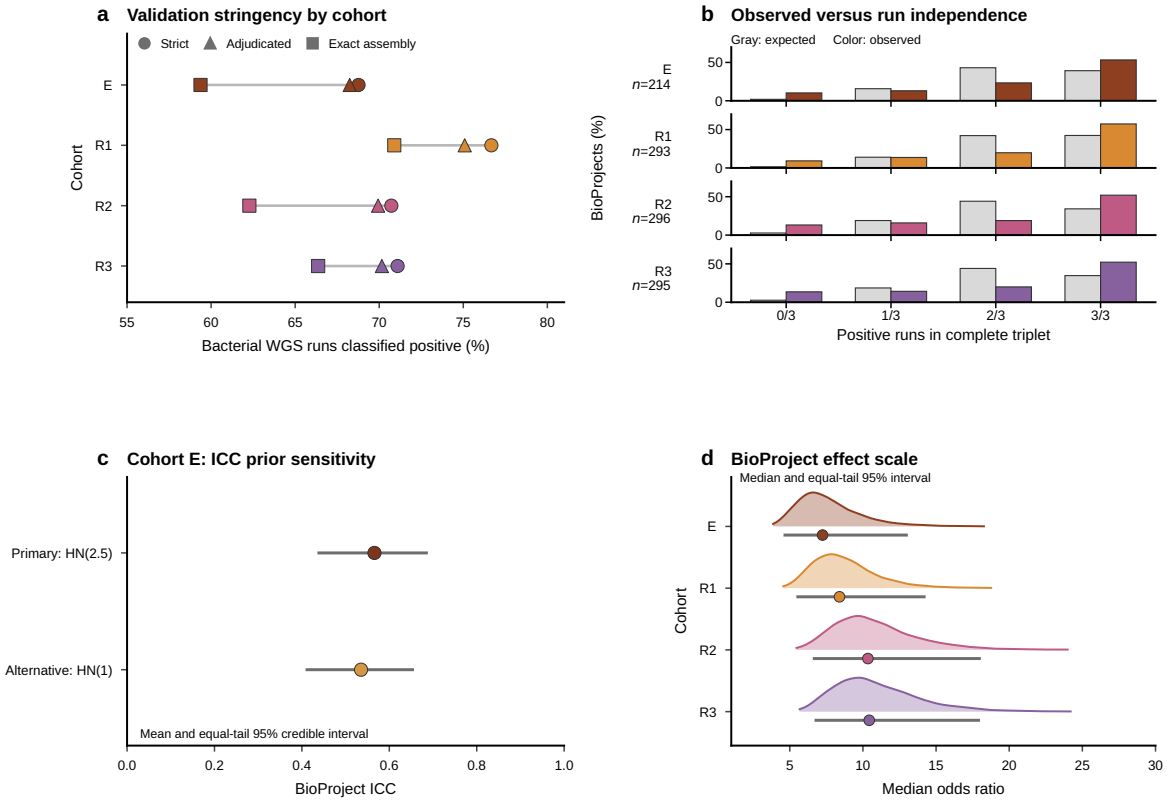
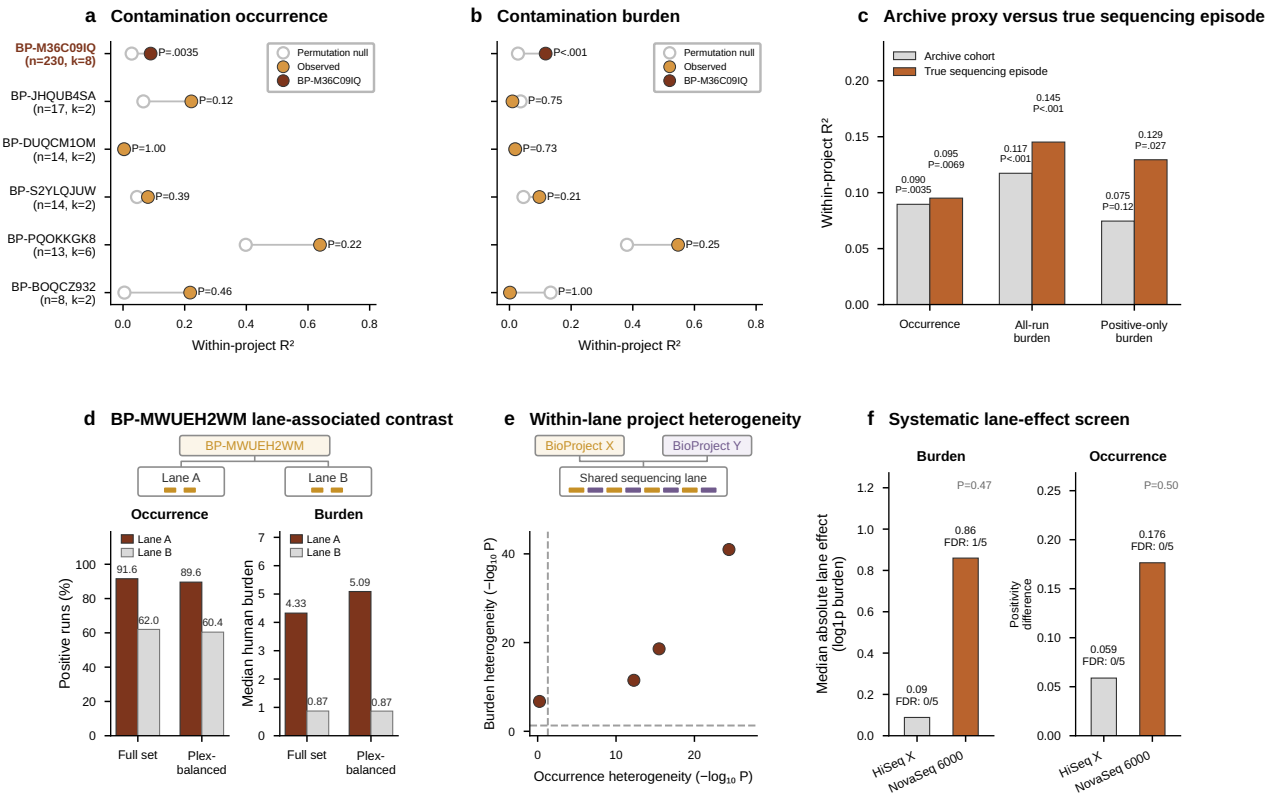


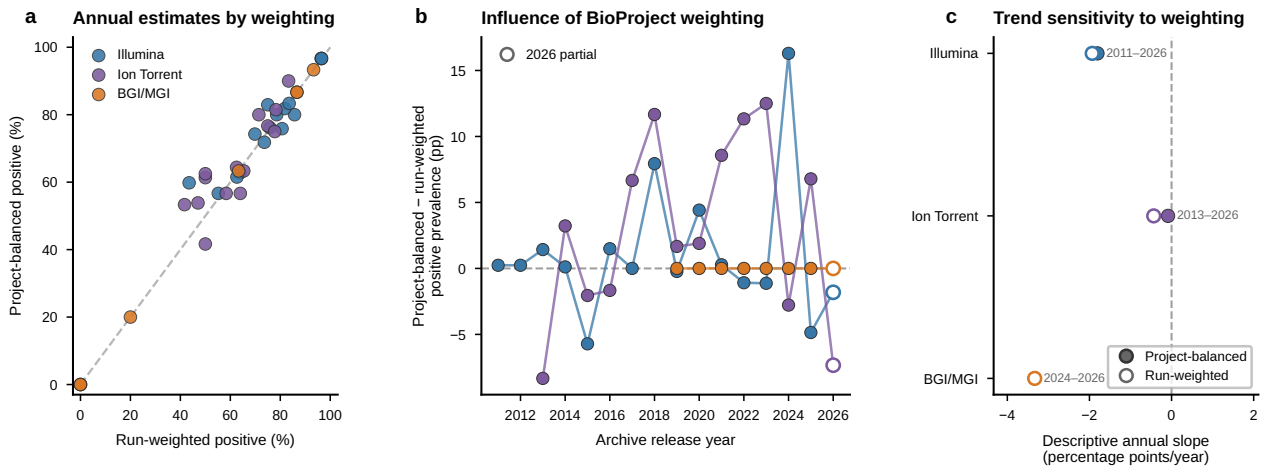
Fig. 4 | Human sequence is detectable in bacterial RNA-seq and is substantially reduced by human-read depletion. **a,** Distribution of human genome-aligned and splice-junction signals across 123 bacterial RNA-seq Experiments representing 153 sequencing runs. Splice-junction evidence is consistent with processed mature human RNA, whereas genome-aligned reads may originate from human DNA or genome-compatible RNA. **b,** Human genome-aligned sequence burden across the RNA-seq survey cohort, ranked by strict human fragments per million input read pairs. Point styles distinguish the four signal classes in panel a. Eighteen Experiments with zero genome-aligned burden are displayed at a plotting floor to accommodate the logarithmic scale. **c,** Annotation-free validation in 32 splice-positive runs: 330/363 testable QNAMEs had at least one exact target recovered; 329/363 recovered every target; 343/373 testable run-by-junction observations were exact. These results support the reproducibility of splice evidence without relying on transcript annotation. **d,** Human genome-aligned signal and privacy-relevant endpoints before (PRE) and after (POST) Hostile depletion in 18 RNA-seq runs selected for recoverable population affinity, ClinVar alternate (ALT) allele observations or computationally supported pathogenic/likely pathogenic (P/LP) ALT alleles. Points indicate the number of runs meeting each endpoint. Total strict human fragments decreased from 43,169 to two, with only one run retaining strict human-fragment evidence POST. Recoverable population affinity, ClinVar ALT alleles and P/LP endpoints fell to zero POST. **e,** Persistence of splice-junction evidence after Hostile in the 13 PRE splice-positive runs from the mitigation panel. Paired points show splice-supporting read-pair and junction observations per run. Splice-supporting QNAMEs decreased from 202 to 63 (68.8% reduction), and run-by-junction observations from 207 to 87 (58.0% reduction). All 13 runs remained splice-positive POST, and every retained POST observation had an exact PRE counterpart.



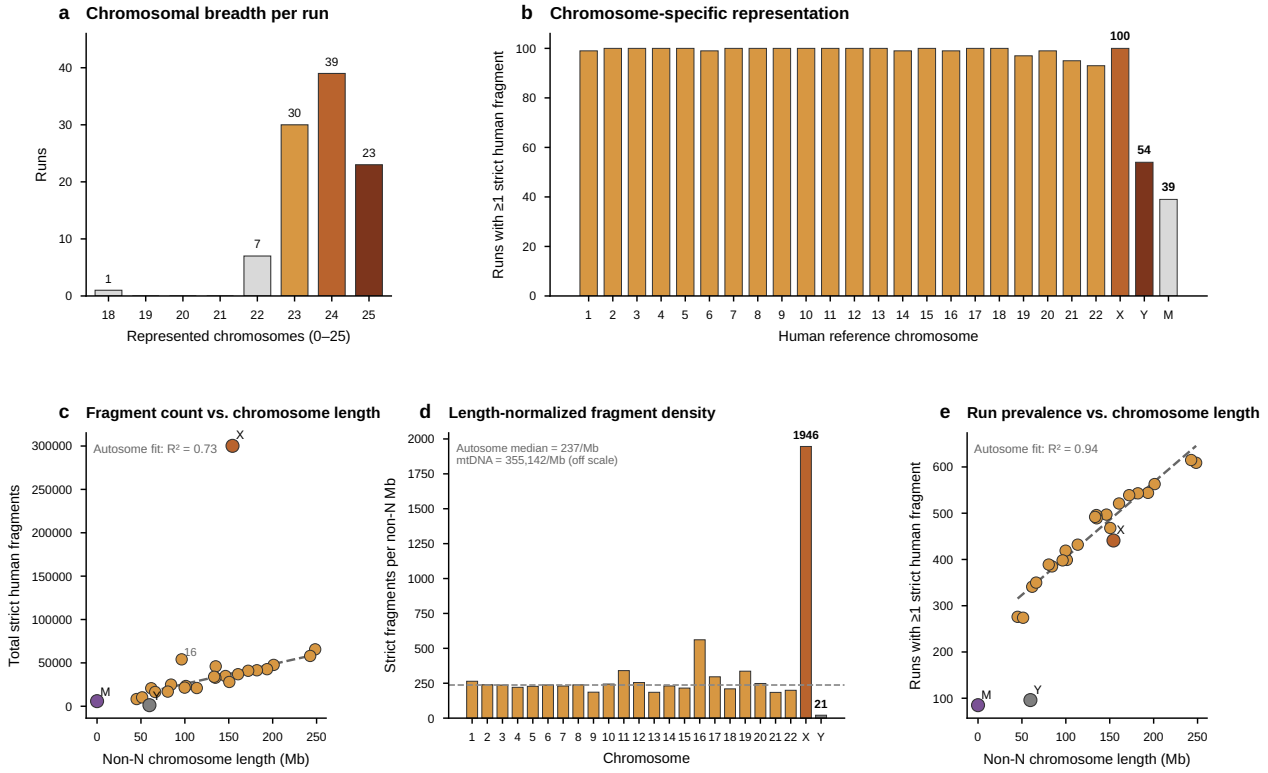
Extended Data Fig. 1 | Human-sequence positivity persists under stringent validation and exhibits strong BioProject-level clustering. **a,** Human-sequence positivity across increasingly conservative validation criteria in exploratory cohort E and systematic replication cohorts R1–R3. Strict mapping identifies human-reference-aligned fragments meeting stringent alignment and quality criteria. Bacterial-match adjudication excludes fragments that could also be explained by bacterial sequence similarity. Exact-assembly validation further requires a successful comparison against an exactly matched bacterial assembly. **b,** Observed recurrence of adjudicated human-sequence positivity among BioProjects represented by three selected, evaluable runs. The observed proportions of BioProjects with zero, one, two or three positive runs are compared with expectations under independent run outcomes, calculated separately for each cohort using its observed positive-run fraction. The analysis includes 214, 293, 296 and 295 BioProjects in E, R1, R2 and R3, respectively. **c,** Sensitivity of the BioProject intraclass correlation coefficient (ICC) to alternative prior specifications in Bayesian random-intercept logistic models fitted to the 642-run, 214-project panel from cohort E. Points show posterior estimates and horizontal intervals denote equal-tail 95% credible intervals. Similar estimates across prior specifications indicate that the inferred BioProject-level clustering is not strongly dependent on the selected priors. **d,** BioProject-level heterogeneity expressed as the median odds ratio (MOR), derived from separate Bayesian random-intercept models for E and R1–R3. Posterior distributions summarize the magnitude of between-project variation in contamination positivity on the odds-ratio scale. Points indicate posterior medians and horizontal segments denote equal-tail 95% credible intervals. Together, these analyses demonstrate that human-sequence positivity persists under increasingly stringent validation criteria and that substantial BioProject-level clustering is reproducible across independently sampled cohorts and robust to alternative model priors.



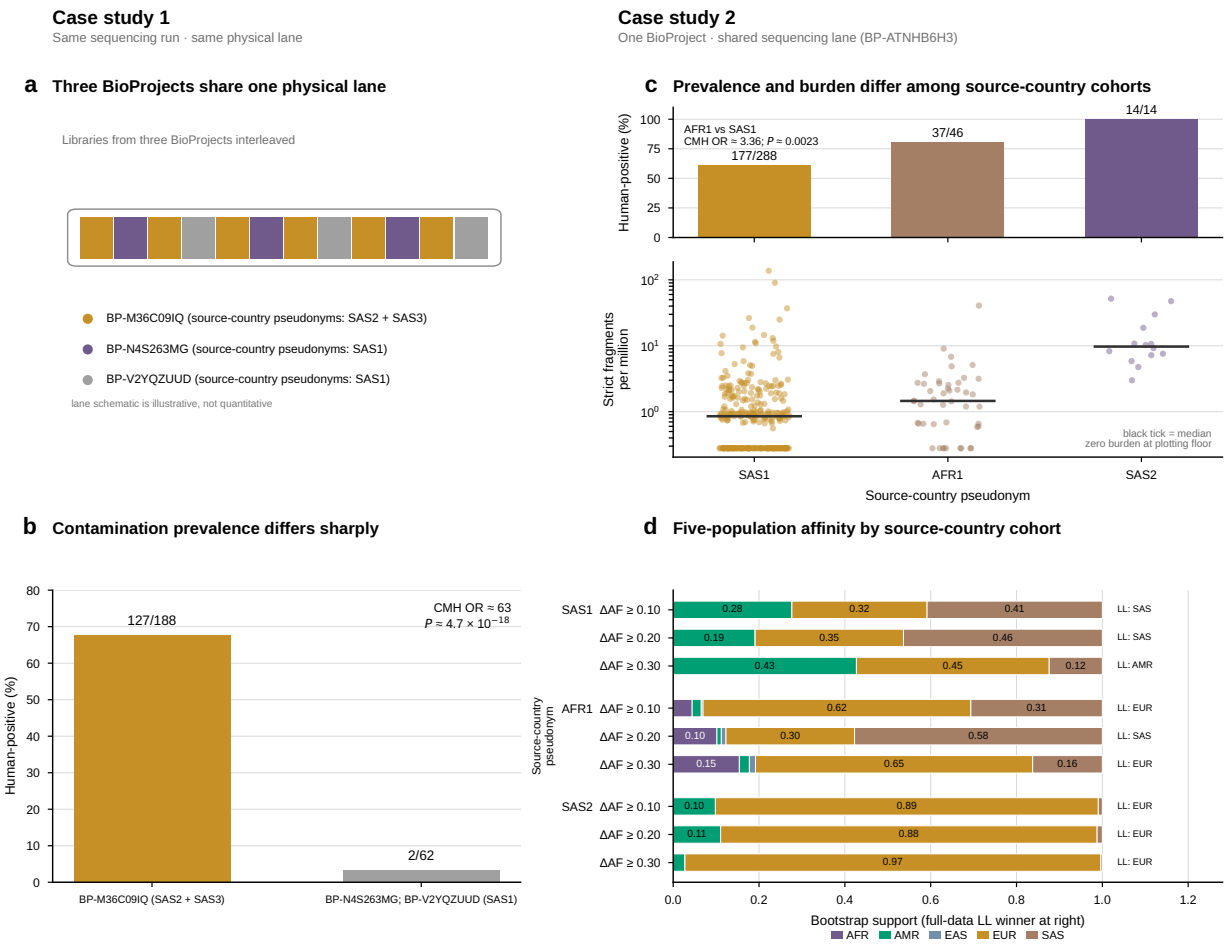
Extended Data Fig. 2 | Archive proxies and sequencing episodes reveal technical structure without a universal lane mechanism. a, Within-project variation in human-sequence positivity across archive-submission cohorts in 296 sequencing runs from six BioProjects in cohort E. Observed within-project R^2 values are compared with permutation-null expectations to evaluate whether contamination occurrence differs systematically among submission groups. Archive-submission cohorts serve as administrative proxies and might not necessarily correspond to physical sequencing or laboratory batches. **b,** Corresponding analysis of strict human-fragment burden across the same six BioProjects. Observed within-project R^2 values are compared with permutation-null expectations, testing whether submission-associated differences extend beyond contamination occurrence to the amount of human sequence recovered. **c,** Comparison of archive-submission cohorts with sequencing episodes reconstructed from read-group metadata in a focal 230-run BioProject. Within-project R^2 values are shown for contamination occurrence, all-run burden and positive-only burden. Reconstructed episodes explained 0.095, 0.145 and 0.129 of the corresponding variance, compared with 0.090, 0.117 and 0.075 for archive-submission cohorts. Reconstructed sequencing episodes capture additional technical structure but do not independently establish the stage at which contamination originated. **d,** Lane-associated differences in contamination occurrence and strict human-fragment burden within a single BioProject comprising 275 runs. Comparisons are shown for the complete dataset and a 96-run plex-balanced sensitivity analysis. In the complete dataset, 206/225 lane-2 runs (91.6%) and 31/50 lane-3 runs (62.0%) were positive. **e,** Between-project heterogeneity in contamination occurrence and burden across four targeted shared-sequencing-lane contexts. The analysis included 1,085 runs with available contamination results from 1,511 technically eligible runs, representing six pseudonymized BioProjects, with two or three projects sharing each lane context. Points show statistical evidence for project-level differences within each context; the dashed line indicates $P = 0.05$. These targeted comparisons were conditional on available contamination results and are not a systematic estimate of archive-wide lane effects. **f,** Comparison of lane-associated contamination effects between HiSeq X and NovaSeq 6000 using year-matched experiments comprising 340 runs from nine BioProjects. Bars show median absolute lane effects for log-transformed human-fragment burden and absolute differences in contamination positivity. Annotations indicate the number of experiments significant after false-discovery-rate correction. Platform-level permutation tests did not detect significant differences in lane-effect magnitude for burden ($P = 0.47$) or occurrence ($P = 0.50$). Together, these analyses reveal context-dependent contamination patterns, including differences between archive-submission groups and sequencing episodes (a–c), between sequencing lanes within a BioProject (d), and between BioProjects sharing the same lane (e), without identifying a universal lane- or platform-specific contamination mechanism (f).



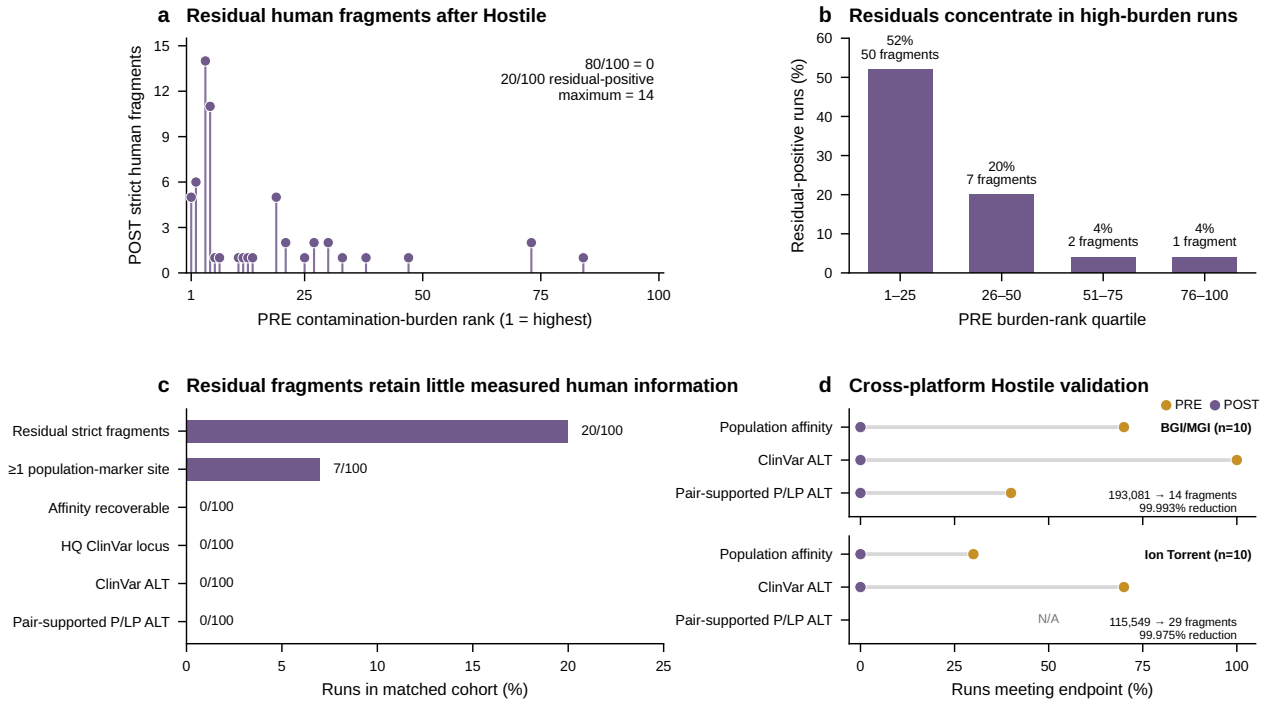
Extended Data Fig. 3 | Platform-specific temporal patterns are sensitive to BioProject weighting. a, Comparison of run-weighted and project-balanced annual strict human-sequence positivity across Illumina, Ion Torrent and BGI/MGI datasets, using archive release year. Each point represents one platform-year estimate. Run-weighted estimates give equal weight to individual sequencing runs, whereas project-balanced estimates give equal weight to contributing BioProjects within each year. The dashed diagonal indicates equality between the two estimates. **b,** Annual differences between project-balanced and run-weighted positivity estimates, expressed in percentage points. Positive values indicate higher prevalence after BioProject balancing, whereas negative values indicate the reverse. The largest absolute difference was 16.3 percentage points for Illumina in 2024. Open points denote partial-year observations for 2026. Archive release year should not be interpreted as sequencing year. **c,** Sensitivity of descriptive temporal slopes to the choice of weighting. Filled points represent slopes fitted to project-balanced annual estimates, and open points represent slopes fitted to run-weighted annual estimates. Project-balanced versus run-weighted slopes were -1.80 versus -1.93 percentage points per year for Illumina (2011–2026), -0.09 versus -0.43 for Ion Torrent (2013–2026) and -3.33 versus -3.33 for BGI/MGI (2024–2026). These descriptive annual-summary slopes are distinct from the BioProject-level fits with cluster-bootstrap uncertainty reported in Fig. 1e. Together, these analyses show that BioProject weighting can substantially influence individual annual prevalence estimates (a,b), while the broad platform-specific temporal patterns remain relatively similar across weighting methods (c).



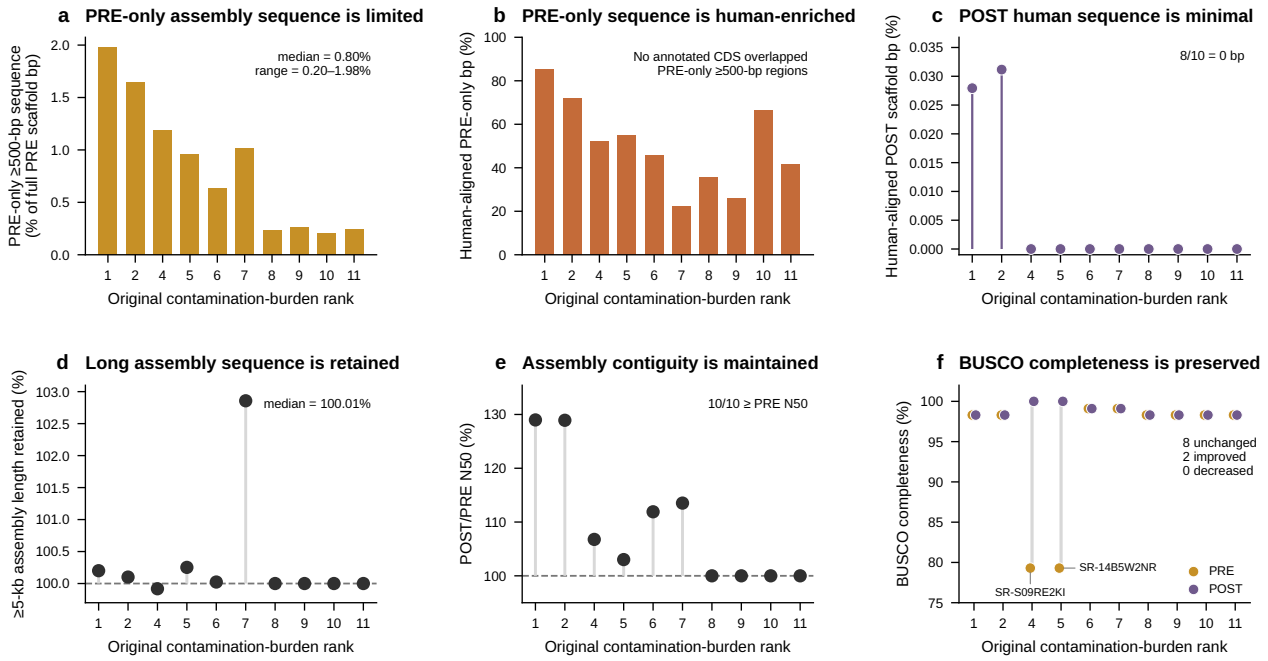
Extended Data Fig. 4 | Contaminating human sequence is broadly distributed across the genome and scales with chromosome size. Panels a,b examine the same 100 paired-end Illumina WGS runs from cohort E selected for the highest absolute strict human-fragment counts and subsequently analyzed for clinical-variant information and human-read depletion (Fig. 2e and Fig. 3). Panels c–e use the broader 1,713-run exploratory cohort E. **a,** Chromosomal breadth of contaminating human sequence in the selected 100-run panel. Bars show the number of runs with strict human fragments detected across 18–25 of the 25 human reference sequences (22 autosomes, chromosomes X and Y, and mitochondrial DNA). Twenty-three of the selected runs represented all 25 sequences; across the full E cohort, 30 runs represented all 25, including seven Ion Torrent runs. **b,** Chromosome-specific representation in the same 100 runs. Bars indicate the number of runs containing at least one strict human fragment from each reference sequence, illustrating the distribution of detectable human sequence across the genome. **c,** Total strict human-sequence counts across cohort E as a function of non-N chromosome length. The ordinary least-squares regression uses autosomes only ($R^2 = 0.73$), with chromosomes X, Y and mitochondrial DNA shown separately. Human-sequence counts generally increased with chromosome size. **d,** Chromosome-specific human-sequence density, expressed as strict units per non-N megabase. Length normalization reduces the chromosome-size dependence of absolute counts and highlights departures from the autosomal background. The mitochondrial value exceeds the plotted scale and is labeled separately. **e,** Number of runs containing at least one strict human-sequence unit from each chromosome, plotted against non-N chromosome length. The autosome-only linear fit ($R^2 = 0.94$) shows that larger autosomes were represented in more runs. For panels c–e, strict units denote fragments in paired-end libraries and reads in single-end libraries. Together, these analyses demonstrate widespread genomic representation rather than restriction to a few recurrent regions; chromosome representation indicates sequence detection, not complete chromosomal coverage or donor identification.



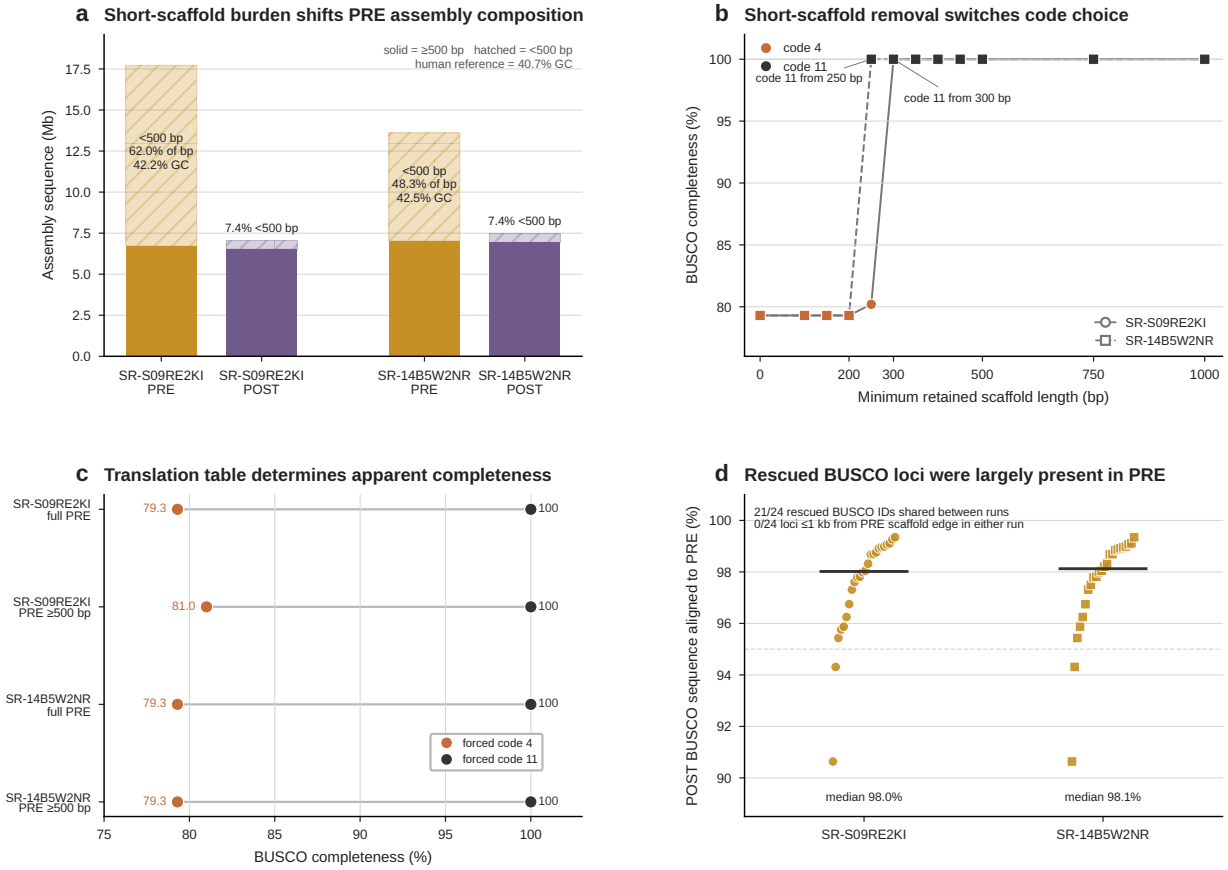
Extended Data Fig. 5 | Human-sequence contamination and population-affinity patterns differ among libraries sharing a final sequencing lane. a, Schematic of three bacterial BioProjects whose libraries were interleaved on the same physical HiSeq X Ten sequencing lane. The schematic illustrates their shared final sequencing context and is not quantitative. **b,** Human-sequence positivity among bacterial libraries sharing the same final sequencing lane. One BioProject associated with pseudonymized source-country groups SAS2 and SAS3 had 127/188 positive libraries, whereas two BioProjects associated with SAS1 together had 2/62 positive libraries. Source-country labels were pseudonymized to reduce privacy exposure. The plex-stratified Cochran–Mantel–Haenszel analysis yielded an odds ratio of approximately 63 ($P \approx 4.7 \times 10^{-18}$). The pronounced difference between groups sharing the same sequencing lane argues against a homogeneous lane-wide contamination source. **c,** Human-sequence positivity and strict-fragment burden across three pseudonymized source-country groups within a second shared sequencing context comprising 348 runs from one BioProject. The difference in positivity between AFR1 and SAS1 remained after controlling for plex position (Cochran–Mantel–Haenszel odds ratio, approximately 3.36; $P \approx 0.0023$). Black ticks denote burden medians; zero-burden observations are displayed at a plotting floor on the logarithmic scale. **d,** Human population-affinity patterns across the same pseudonymized source-country groups, evaluated using population-informative markers selected by maximum pairwise allele-frequency differences (ΔAF) across five 1KGP reference populations (AFR, AMR, EAS, EUR and SAS), at thresholds of at least 0.10, 0.20 and 0.30. Affinity was scored against all five populations. Colored segments show bootstrap support for the respective reference-population affinities, whereas the indicated maximum-likelihood population identifies the strongest fit to the full data; these need not agree. Affinity patterns differed among groups and across marker thresholds, with SAS2 showing relatively consistent EUR affinity and the other groups showing more mixed support. Reference-population affinity does not establish individual ancestry, ethnicity, nationality or geographic origin. Together, these case studies show that libraries sharing a final sequencing lane can retain substantial differences in contamination prevalence and burden (b,c), as well as differences in population-affinity patterns derived from contaminating human sequence (d), without establishing the precise stage at which contamination was introduced.



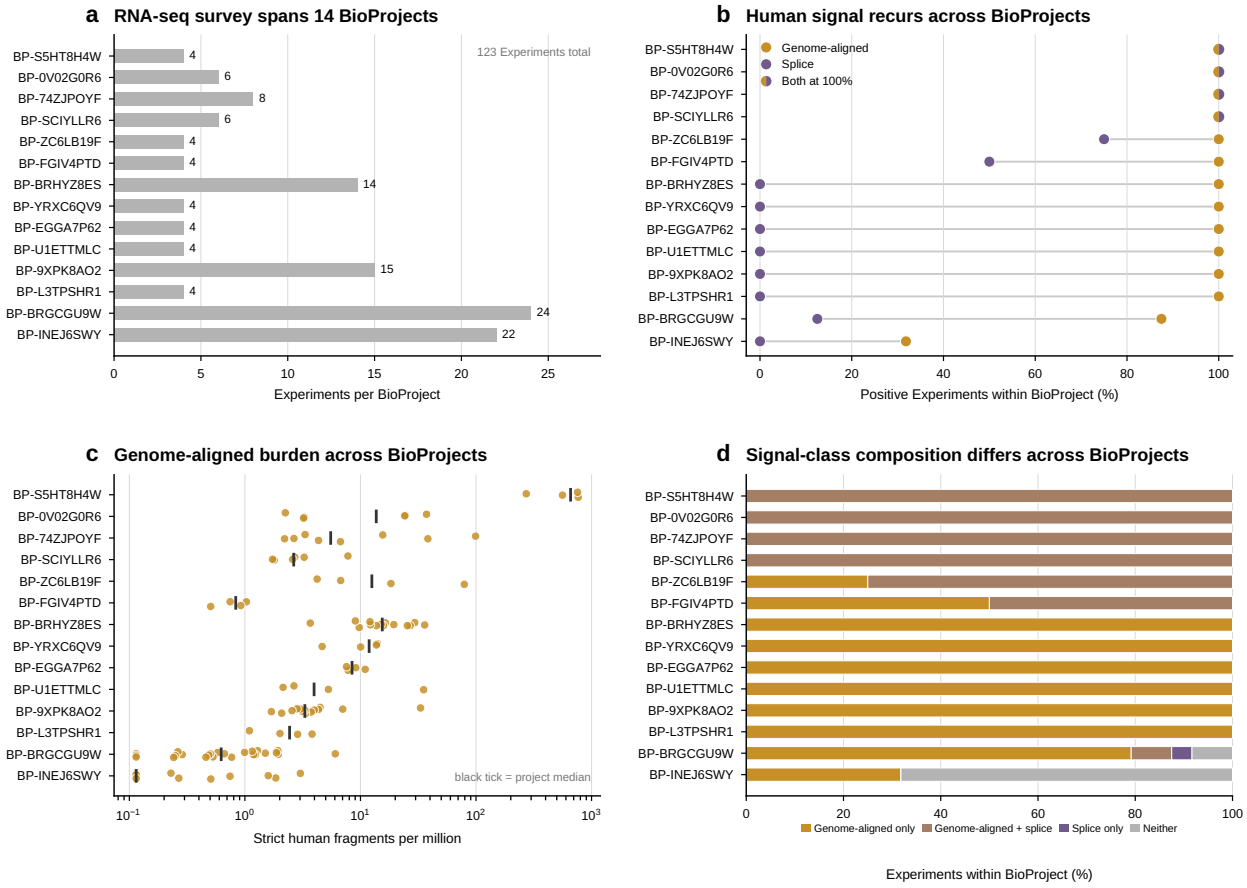
Extended Data Fig. 6 | Residual human signal after Hostile is sparse, information-poor and reduced across sequencing platforms. Panels a–c examine the same 100 paired-end Illumina WGS runs evaluated before (PRE) and after (POST) Hostile human-read depletion in Fig. 3. **a,** Distribution of residual strict human fragments after Hostile, with runs ranked by PRE human-fragment burden. Eighty of 100 runs had no residual strict human fragments, whereas 20 retained 60 fragments in total. The maximum residual count was 14 fragments in a single run. **b,** Frequency and magnitude of residual human-sequence detection across quartiles of PRE contamination burden. Bars show the percentage of runs retaining strict human fragments after depletion, and annotations indicate the number of residual fragments in each quartile. The highest PRE-burden quartile accounted for 50 of the 60 POST fragments, indicating that residual evidence was concentrated in the most heavily contaminated starting datasets. **c,** Privacy-relevant information retained after Hostile in the same 100-run cohort. Bars show the percentage of runs meeting independently evaluated POST endpoints, rather than sequential filtering stages. Although 20 runs retained strict human fragments and seven contained population-informative marker observations, none met the minimum marker-count requirement for population-affinity evaluation or retained high-quality ClinVar-locus observations, ClinVar alternate (ALT) alleles or computationally supported pathogenic/likely pathogenic (P/LP) ALT alleles. **d,** Cross-platform evaluation of Hostile in separately selected BGI/MGI paired-end and Ion Torrent single-end WGS runs (10 runs per platform). Strict human-sequence counts decreased from 193,081 PRE to 14 POST for BGI/MGI (99.993% reduction) and from 115,549 to 29 for Ion Torrent (99.975% reduction). Points show the percentage of runs meeting population-affinity, ClinVar ALT and P/LP ALT evidence endpoints before and after depletion. Recoverable population affinity decreased from 7/10 to 0/10 for BGI/MGI and from 3/10 to 0/10 for Ion Torrent; ClinVar ALT exposure decreased from 10/10 and 7/10 PRE, respectively, to zero POST in both platforms. For BGI/MGI, pair-supported P/LP ALT evidence decreased from 4/10 PRE to 0/10 POST. Physical read-pair support is inapplicable to single-end Ion Torrent data; its separate single-read P/LP candidate screen was negative in all ten runs both PRE and POST. Pair-supported P/LP ALT evidence represents computational sequence-placement support, not clinical validation. Together, these analyses show that substantial depletion of human sequence reduces the measured privacy-relevant information across sequencing platforms (a–d), although non-detection under the evaluated endpoints does not establish complete elimination of human-derived information.



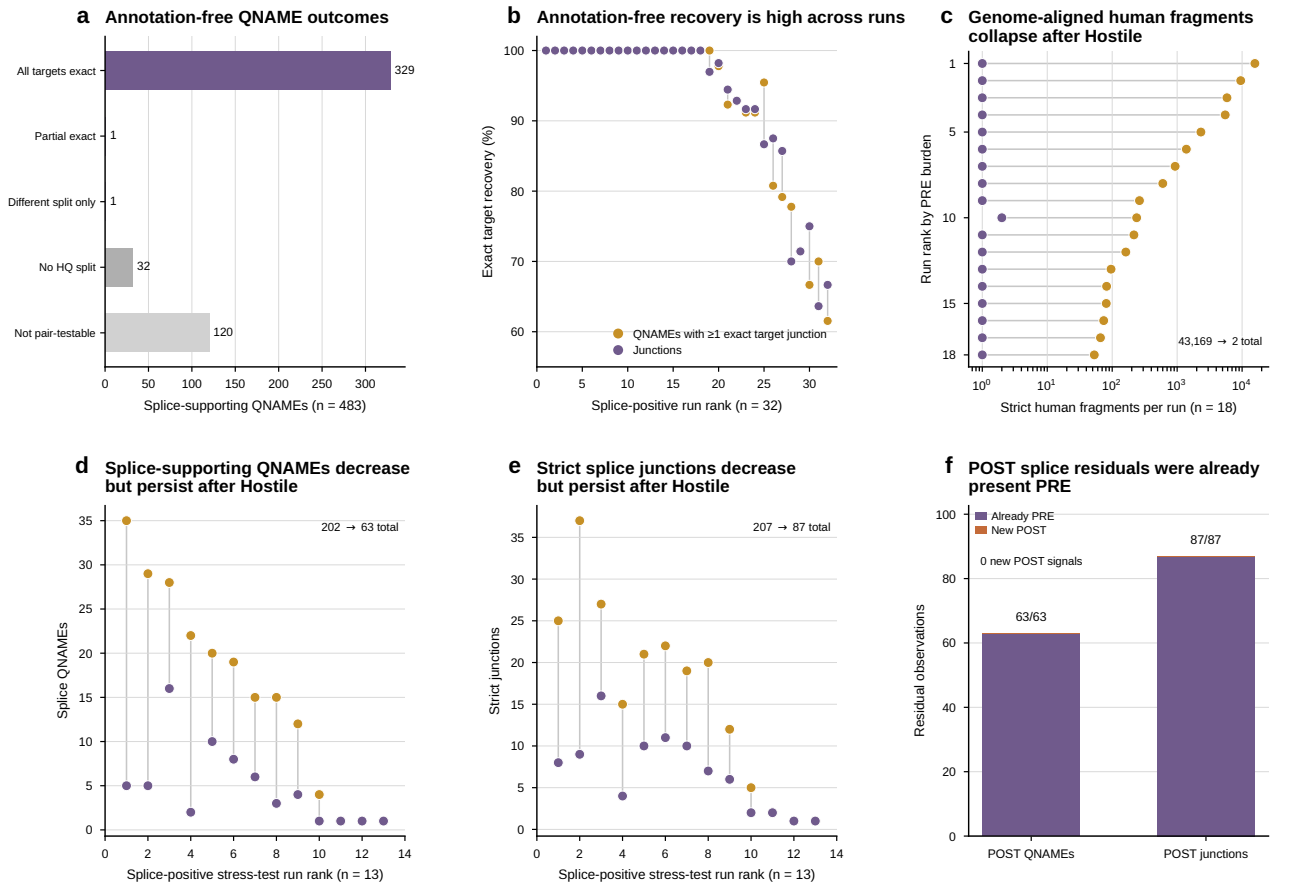
Extended Data Fig. 7 | Host-read depletion preferentially removes human-enriched assembly sequence while preserving bacterial genome quality. Ten high-human-sequence-burden paired-end Illumina WGS runs were evaluated before (PRE) and after (POST) Hostile depletion and reassembly, corresponding to the assembly-preservation stress test in Fig. 3e. Runs are identified by their original contamination-burden ranks (1, 2 and 4–11, rank 3 was excluded because its failed PRE assembly). **a,** Fraction of PRE scaffold-assembly sequence represented by PRE-only regions, defined as regions not covered by alignment of the corresponding POST assembly and restricted to lengths of at least 500 bp. PRE-only sequence constituted a median of 0.80% of the PRE assembly. **b,** Percentage of PRE-only assembly sequence in regions at least 500 bp long that aligned to the human reference. No annotated bacterial CDS overlapped these PRE-only regions. **c,** Human-reference-aligned sequence remaining in the full POST scaffold assemblies after Hostile depletion and reassembly. Eight of ten POST assemblies contained no human-aligned sequence. The two remaining assemblies contained only 0.0279% and 0.0312% human-aligned scaffold sequence, respectively. **d,** Total length of scaffolds at least 5 kb in POST relative to PRE assemblies. The median POST/PRE length ratio was approximately 100%, indicating little change in assembled sequence length. The dashed line indicates the PRE reference value (100%). **e,** Assembly contiguity after depletion, expressed as POST/PRE N50. N50 was equal to or greater than its PRE value in all ten runs, with a median increase of 4.9%. The dashed line indicates 100%. **f,** BUSCO completeness before and after depletion. Paired points represent individual runs. Completeness was unchanged in eight runs and increased by approximately 20.7 percentage points in two, with no decreases observed. These apparent improvements reflected changes in automated genome-completeness assessment rather than recovery of previously absent bacterial genes, as examined in Extended Data Fig. 8. Together, these results show that human-read depletion removes a small, human-enriched component of the assembled sequence (a–c) while maintaining the measured bacterial assembly-quality metrics (d–f). PRE-only assembly sequence should not be interpreted as sequence physically deleted from bacterial chromosomes.



Extended Data Fig. 8 | Short-scaffold burden can alter automated translation-table selection and apparent bacterial genome completeness. Two high-burden WGS datasets showing an approximately 20.7-percentage-point increase in BUSCO completeness after Hostile depletion and reassembly (Extended Data Fig. 7f) were examined to distinguish recovery of bacterial sequence from changes in automated genome-quality assessment. **a,** Scaffold composition before (PRE) and after (POST) Hostile depletion and reassembly. Bars partition the assemblies into scaffolds at least 500 bp long and shorter scaffolds. Short scaffolds accounted for 62.0% and 48.3% of PRE assembly sequence in the two datasets, respectively, compared with approximately 7.4% POST. Their GC contents (42.2% and 42.5%) differed substantially from the longer-scaffold component (approximately 65%) and were closer to that of the human reference (40.7%). GC similarity is consistent with human origin but is not sufficient to establish sequence provenance. **b,** BUSCO completeness after progressively increasing the minimum retained scaffold length in the original PRE assemblies, without Hostile depletion or reassembly. The abrupt increase in apparent completeness coincided with a switch in the automatically selected Prodigal translation table from code 4 to code 11. The switch occurred at minimum scaffold lengths of 300 bp and 250 bp in the two datasets, respectively, after which completeness reached 100%. **c,** BUSCO completeness of the original complete PRE assemblies and their at-least-500-bp subsets evaluated with forced translation tables 4 or 11. Code 11 yielded 100% completeness even for the unfiltered PRE assemblies, whereas code 4 produced completeness estimates of approximately 79–81% regardless of scaffold filtering. This establishes that the apparent completeness increase was primarily associated with translation-table selection rather than recovery of previously absent bacterial sequence. **d,** Recovery of BUSCO sequence information in the original PRE assemblies for loci classified as Missing or Fragmented PRE but Complete POST. Each dataset contained 24 such BUSCO loci. POST BUSCO sequences aligned to PRE at median query coverages of 98.0% and 98.1%, respectively, with horizontal lines indicating medians. Twenty-one of the 24 BUSCO identities were shared between the two datasets, and none of the recovered loci lay within 1 kb of a PRE scaffold boundary. Together, these analyses demonstrate that the apparent BUSCO improvements in these two datasets were predominantly attributable to changes in automated gene-prediction settings (b,c), with the corresponding bacterial loci already largely represented in the original assemblies (d), rather than reconstruction of missing genes.



Extended Data Fig. 9 | Human RNA-seq signals recur across independent BioProjects and vary within projects. The RNA-seq survey comprises 153 runs representing 123 SRA Experiments, 117 BioSamples and 14 BioProjects (Fig. 4). Genome-aligned denotes strict human-reference-aligned sequence and does not independently establish genomic DNA origin, as human RNA can also contribute to this signal. **a,** Distribution of the 123 surveyed Experiments across 14 pseudonymized BioProjects. Bars indicate the number of Experiments contributed by each BioProject. BioProjects are displayed in a consistent order across panels a–d. **b,** Within-project prevalence of genome-aligned and stringent human splice-junction signals. Points indicate the percentage of Experiments positive for each signal class; coincident 100% values are represented by split symbols. At least one Experiment with genome-aligned human signal was detected in all 14 BioProjects, whereas splice-junction evidence occurred in seven. These results demonstrate broad project-level recurrence of genome-aligned signal but more restricted distribution of splice evidence. **c,** Genome-aligned human-sequence burden within and between BioProjects, expressed as strict human fragments per million input read pairs. Each point represents an Experiment, and black ticks denote BioProject medians. Human-sequence burden varies substantially both among Experiments within individual BioProjects and between projects. Experiments with zero measured burden are displayed at a common plotting floor on the logarithmic axis. **d,** Distribution of mutually exclusive human-sequence signal classes within each BioProject. Stacked bars show the proportions of Experiments exhibiting genome-aligned signal only, genome-aligned plus splice-junction signal, splice-junction signal only or neither measured signal. Across the 123 Experiments, the respective counts were 74, 31, one and 17, matching the categories in Fig. 4a. Together, these analyses demonstrate that human-sequence signals recur across independently represented BioProjects (a,b), while the magnitude and composition of these signals vary both within and between projects (c,d).



Extended Data Fig. 10 | Annotation-free validation confirms human splice evidence and reveals residual processed-RNA signal after Hostile depletion. Panels a,b examine independent annotation-free validation of human splice-junction evidence, whereas panels c–f examine human-sequence persistence before (PRE) and after (POST) Hostile depletion in bacterial RNA-seq datasets (Fig. 4). **a,** Annotation-free validation outcomes for 483 splice-supporting read-pair identifier (QNAME) records. Of these, 120 were not pair-testable and were excluded from the exact-recovery denominator. Among 363 pair-testable QNAMEs, 329 had all target junctions recovered exactly, one had partial exact recovery, one supported a different high-quality split and 32 had no qualifying high-quality split recovered. Thus, 330/363 (90.9%) had at least one exact target recovery. Non-testable observations were not counted as validation failures. **b,** Run-level annotation-free recovery of splice-supporting QNAMEs and junctions across 32 splice-positive runs. Points show the percentage of exactly recovered observations for individual runs after remapping without supplied transcript annotation or known splice sites. Across runs, 330/363 pair-testable QNAMEs (90.9%) and 343/373 testable run-by-junction observations (92.0%) were recovered exactly. Median run-level recovery was 100% for both measures. **c,** Genome-aligned strict human fragments before and after Hostile depletion in 18 matched RNA-seq stress-test runs, ranked by PRE burden. Total strict human fragments decreased from 43,169 PRE to two POST, with 17/18 runs retaining no detectable strict human fragments. Zero counts are displayed at a common plotting floor on the logarithmic axis. **d,** Splice-supporting QNAME observations before and after depletion in the 13 PRE splice-positive runs from the same stress-test cohort. Total observations decreased from 202 PRE to 63 POST (68.8% reduction), but all 13 runs retained detectable splice-supporting QNAMEs. **e,** Stringent splice-junction observations in the same 13 runs before and after depletion. Total run-by-junction observations decreased from 207 PRE to 87 POST (58.0% reduction), with all 13 runs remaining splice-positive. **f,** Exact PRE–POST overlap of residual splice evidence. All 63 POST splice-supporting QNAME observations and all 87 POST junction observations had exact counterparts in the corresponding PRE datasets, with no newly detected POST observations. Together, these analyses demonstrate independent recovery of human splice-junction evidence (a,b) and show that genome-reference-based Hostile depletion nearly eliminates genome-aligned human fragments while leaving a reduced but detectable component of pre-existing processed-RNA evidence (c–f).

References

[1] Alexander Astashyn, Eric S. Tvedte, Deacon Sweeney, Victor Sapojnikov, Nathan Bouk, Victor Joukov, Eyal Mozes, Pooja K. Strope, Pape M. Sylla, Lukas Wagner, Shelby L. Bidwell, Larissa C. Brown, Karen Clark, Emily W. Davis, Brian Smith-White, Wratko Hlavina, Kim D. Pruitt, Valerie A. Schneider, and Terence D. Murphy. Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biology*, 25:60, 2024. doi: 10.1186/s13059-024-03198-7.

[2] David Yuan, Alisha Ahamed, Awais Athar, Rajkumar Devaraj, Dipayan Gupta, Muhammad Haseeb, Maira Ihsan, Eugene Ivanov, Vishnukumar Kadhirvelu, Amnon Khen, Manish Kumar, Ankur Lathi, Isuru Liyanage, Lili Meszaros, Colman O'Cathail, Joana Pauperio, Ruben Paz, Stephane Pesant, Nadim Rahman, Jeena Rajan, Iva Tutis, Marianna Ventouratou, Senthilnathan Vijayaraja, Zahra Waheed, Peter Woollard, Tony Burdett, Guy Cochrane, and Ugis Sarkans. The European Nucleotide Archive in 2025. *Nucleic Acids Research*, 54(D1):D120–D127, 2026. doi: 10.1093/nar/gkaf1295.

[3] Maria Lourdes Sampietro, M. Thomas P. Gilbert, Oscar Lao, David Caramelli, Martina Lari, Jaume Bertranpetit, and Carles Lalueza-Fox. Tracking down human contamination in ancient human teeth. *Molecular Biology and Evolution*, 23(9):1801–1807, 2006. doi: 10.1093/molbev/msl047.

[4] Mark S. Longo, Michael J. O'Neill, and Rachel J. O'Neill. Abundant human DNA contamination identified in non-primate genome databases. *PLoS ONE*, 6(2):e16410, 2011. doi: 10.1371/journal.pone.0016410.

[5] Kirill Kryukov and Tadashi Imanishi. Human contamination in public genome assemblies. *PLoS ONE*, 11 (9):e0162424, 2016. doi: 10.1371/journal.pone.0162424.

[6] Florian P. Breitwieser, Mihaela Pertea, Aleksey V. Zimin, and Steven L. Salzberg. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Research*, 29(6):954–960, 2019. doi: 10.1101/gr.245373.118.

[7] Galo A. Goig, Silvia Blanco, Alberto L. Garcia-Basteiro, and Iñaki Comas. Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. *BMC Biology*, 18(1):24, 2020. doi: 10.1186/s12915-020-0748-z.

[8] Stephen J. Bush, Thomas R. Connor, Tim E. A. Peto, Derrick W. Crook, and A. Sarah Walker. Evaluation of methods for detecting human reads in microbial sequencing datasets. *Microbial Genomics*, 6(7):e000393, 2020. doi: 10.1099/mgen.0.000393.

[9] Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, et al. The complete sequence of a human genome. *Science*, 376(6588):44–53, 2022. doi: 10.1126/science.abj6987.

[10] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015. doi: 10.1038/nature15393.

[11] Bede Constantinides, Martin Hunt, and Derrick W. Crook. Hostile: accurate decontamination of microbial host sequences. *Bioinformatics*, 39(12):btad728, 2023. doi: 10.1093/bioinformatics/btad728.

[12] Maura Costello, Mark Fleharty, Justin Abreu, Yossi Farjoun, Steven Ferriera, Laurie Holmes, Brian Granger, Lisa Green, Tom Howd, Tamara Mason, Gina Vicente, Michael Dasilva, Wendy Brodeur, Timothy DeSmet, Sheila Dodge, Niall J. Lennon, and Stacey Gabriel. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, 19(1):332, 2018. doi: 10.1186/s12864-018-4703-0.

[13] Yoshihiko Tomofuji, Kyuto Sonehara, Toshihiro Kishikawa, Yuichi Maeda, Kotaro Ogawa, Shuhei Kawabata, Takuro Nii, Tatsusada Okuno, Eri Oguro-Igashira, Makoto Kinoshita, Masatoshi Takagaki, Kenichi Yamamoto, Takashi Kurakawa, Mayu Yagita-Sakamaki, Akiko Hosokawa, Daisuke Motooka, Yuki Matsumoto, Hidetoshi Matsuoka, Maiko Yoshimura, Shiro Ohshima, Shota Nakamura, Hidenori Inohara, Haruhiko Kishima, Hideki Mochizuki, Kiyoshi Takeda, Atsushi Kumanogoh, and Yukinori Okada. Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. *Nature Microbiology*, 8(6):1079–1094, 2023. doi: 10.1038/s41564-023-01381-3.

[14] Caitlin Guccione, Lucas Patel, Yoshihiko Tomofuji, Daniel McDonald, Antonio Gonzalez, Gregory D. Sepich-Poore, Kyuto Sonehara, Mohsen Zakeri, Yang Chen, Amanda Hazel Dilmore, Neil Damle, Sergio E. Baranzini, George Hightower, Teruaki Nakatsuji, Richard L. Gallo, Ben Langmead, Yukinori Okada,

Kit Curtius, and Rob Knight. Incomplete human reference genomes can drive false sex biases and expose patient-identifying information in metagenomic data. *Nature Communications*, 16:825, 2025. doi: 10.1038/s41467-025-56077-5.

[15] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018. doi: 10.1093/bioinformatics/bty191.

[16] Derrick E. Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20:257, 2019. doi: 10.1186/s13059-019-1891-0.

[17] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9 (4):357–359, 2012. doi: 10.1038/nmeth.1923.

[18] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, Sergey I. Nikolenko, Son Pham, Andrey D. Prjibelski, Alexey V. Pyshkin, Alexander V. Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max A. Alekseyev, and Pavel A. Pevzner. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19 (5):455–477, 2012. doi: 10.1089/cmb.2012.0021.

[19] Guillaume Marçais, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, 14(1): e1005944, 2018. doi: 10.1371/journal.pcbi.1005944.

[20] Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. High throughput ANI analysis of 90k prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9:5114, 2018. doi: 10.1038/s41467-018-07641-9.

[21] Oliver Schwengers, Lukas Jelonek, Marius Alfred Dieckmann, Sebastian Beyvers, Jochen Blom, and Alexander Goesmann. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, 7(11):000685, 2021. doi: 10.1099/mgen.0.000685.

[22] Mosè Manni, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, 38(10):4647–4654, 2021. doi: 10.1093/molbev/msab199.

[23] Daehwan Kim, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37 (8):907–915, 2019. doi: 10.1038/s41587-019-0201-4.

[24] Adam Frankish, Sílvia Carbonell-Sala, Mark Diekhans, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research*, 51(D1):D942–D949, 2023. doi: 10.1093/nar/gkac1071.

Supplementary Files

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

- SupplementaryData1.xlsx
- HumansequencecontaminationinmicrobialsequencingarchivesvSubSI.pdf
- SupplementaryTable1.xlsx
- SupplementaryData2.xlsx