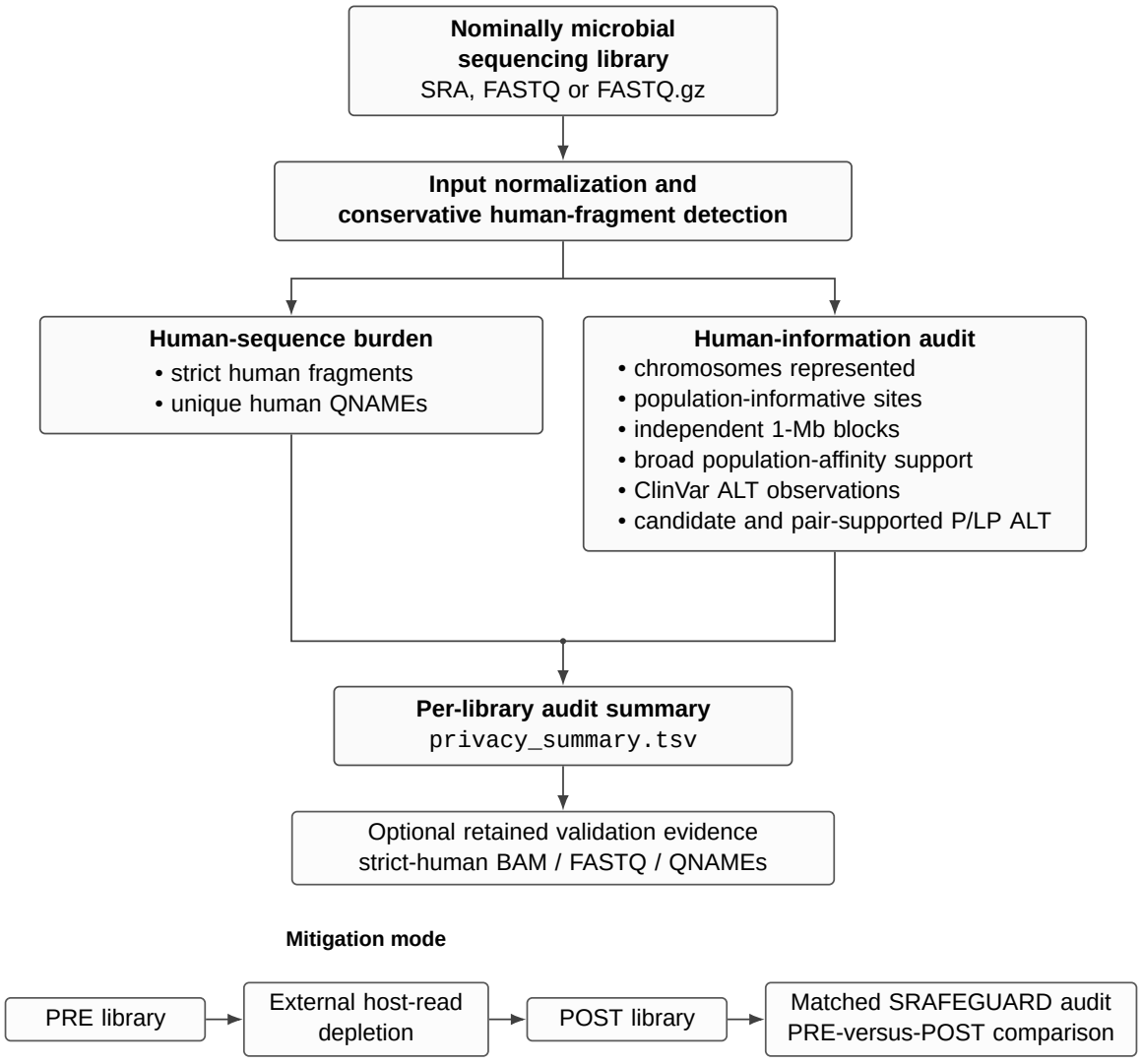
Supplementary Information

Supplementary Note 1 — SRAFEGUARD: an audit framework for recoverable human information in microbial sequencing data

SRAFEGUARD (v0.1.0) was developed to convert the study's human-contamination analyses into a reproducible audit workflow for nominally microbial sequencing libraries. The tool accepts a local SRA file or FASTQ/FASTQ.gz input and applies the same conservative human-evidence framework used in this study. Its purpose is measurement rather than depletion: SRAFEGUARD does not remove human reads, but reports how much human sequence information remains recoverable before or after an external filtering step.

Workflow



For each library, SRAFEGUARD identifies strict human-sequence evidence using the predefined mapping and fragment-level criteria (Methods), and summarizes human-fragment burden and chromosome representation. Population-informative autosomal SNPs are ascertained using the maximum allele-frequency difference across five 1000 Genomes[1] superpopulations (AFR, AMR, EAS, EUR and SAS). For the analyses reported here, SRAFEGUARD used a marker threshold of $\Delta AF \geq 0.30$ and scored clean REF/ALT observations against the same five reference populations. Affinity support was evaluated using 1,000 within-library bootstrap resamples of observed chromosome-qualified 1-Mb genomic blocks. The affinity-recoverability endpoint required at least 20 clean marker loci and bootstrap support ≥ 0.90 for the population with the highest full-data likelihood. Marker positivity, marker eligibility and affinity recoverability were treated as distinct endpoints. Reference-population affinity does not establish individual ancestry, ethnicity, nationality, identity or the causal source of contamination.

SRAFEGUARD also audits clinically annotated human variation. High-quality ClinVar-locus observations, ClinVar ALT observations, candidate pathogenic/likely pathogenic (P/LP) ALT observations and read-pair-supported P/LP ALT observations are treated as distinct endpoints. Read-pair support requires the ALT-carrying read to retain its allele at the expected human locus during qualifying paired-end remapping, subject to the prescribed mapping-quality, paired-placement and competing-alignment criteria. This confirmation endpoint applies to paired-end Illumina and BGI/MGI libraries. Ion Torrent libraries were single-end, so read-pair confirmation was not applicable (N/A), rather than a measured zero. The candidate P/LP ALT screen identified no observations in the ten Ion Torrent libraries either before or after depletion (0/10 PRE and 0/10 POST). These are computational allele-placement observations, not clinically validated variants, confirmed donor genotypes or evidence of phenotype or identity.

The principal machine-readable output is a per-library `privacy_summary.tsv` containing human-fragment burden, chromosome breadth, population-affinity metrics and ClinVar/P/LP exposure metrics. Read-level evidence can optionally be retained as strict-human BAM/FASTQ/QNAME files for validation; otherwise these intermediate evidence files can be removed after analysis. Batch and re-analysis modes support repeated, configuration-controlled auditing of larger cohorts.

For mitigation experiments, SRAFEGUARD was applied to matched original and externally depleted sequencing libraries. Hostile provided the human-read depletion step, whereas SRAFEGUARD independently quantified residual human-sequence information. In the 100-library bacterial WGS stress test, 91 libraries met the population-marker eligibility criterion before depletion, of which 39 met the affinity-recoverability criterion. Following depletion, seven libraries retained at least one clean population-marker locus, but none met the marker-eligibility or affinity-recoverability criteria. In the 18-library bacterial RNA-seq stress test, affinity-recoverable human information decreased from nine libraries to none, while splice-junction evidence was evaluated separately. These findings demonstrate substantial reductions in the measured genetic-information endpoints without establishing complete removal of all human-derived sequence.

SRAFEGUARD is designed as a conservative screening and verification framework rather than an identity-inference tool. Its current genomic audit also does not replace transcriptome- or splice-aware screening where processed human RNA is a concern.

Supplementary Note 2 — Read-pair-supported P/LP ALT observations in bacterial sequencing contamination

To illustrate the type of medically interpretable information carried by contaminating human reads, we examined read-pair-supported pathogenic or likely pathogenic (P/LP) ClinVar ALT observations in a deliberately selected set of 100 high-human-contamination bacterial WGS libraries drawn from exploratory cohort E. The examples below were selected for clinical interpretability and read-level support rather than to estimate prevalence. Accession, BioProject, sequencing-provider and geographic identifiers are deliberately omitted to avoid linking clinically annotated alleles to individual public microbial datasets.

Gene / variant	ClinVar classification / review status	Selected associated condition or context	ClinVar Variation ID	HQ depth	REF	ALT	Pair-supported ALT reads
CFTR, chr7:118903127 G>T	Pathogenic; practice guideline	Cystic fibrosis; CFTR-related disorders	7115	1	0	1	1
HBB, chr11:5285096 T>C	Pathogenic; criteria provided, multiple submitters, no conflicts	Hb SS disease; beta thalassemia; beta0 thalassemia	21191	1	0	1	1
TAP2, chr6:32653707 C>T	Likely pathogenic; criteria provided, single submitter	Bare lymphocyte syndrome type 1	534707	1	0	1	1
HBB, chr11:5286133 A>G	Pathogenic; no assertion criteria provided	Hemolytic anemia	15251	30	28	2	2
KMT2C, chr7:153354669 G>A	Pathogenic; criteria provided, single submitter	Cancer-associated context; see below	1216958	1	0	1	1

Each value under “Pair-supported ALT reads” counts an ALT-containing read whose complete paired-end fragment passed the computational locus-placement procedure: the ALT-supporting read remapped to the expected human locus with the ALT preserved, sufficient mapping quality and proper paired-end placement, and without an equal-or-better competing genomic placement. The mate read itself was not required to overlap or carry the ALT allele. Thus, the counts above are read-pair-supported ALT observations: their placement was corroborated by the paired-end fragment, and they are not independent genotype calls. Human contamination depth at these loci was usually sparse; accordingly, no diploid genotype was inferred.

The KMT2C observation provides an additional example of why such leakage can be clinically interpretable even when the database record itself is not tied to a specific disease label. The recovered allele corresponds to NM_170606.3:c.6082C>T (p.Arg2028Ter; R2028*), a pathogenic truncating variant (ClinVar Variation ID 1216958). The identical R2028* alteration has been reported in the 639-V urothelial carcinoma cell line[2] and in a triple-negative breast cancer-derived model, 885512-296-R-J2[3]. These reports provide an independently published cancer-model context for the same molecular alteration; they do not establish that the recovered allele caused cancer, or even that any contaminating contributor had cancer.

Overall, these examples demonstrate that contaminating human sequence in nominally bacterial datasets can contain medically interpretable alleles, including variants associated with serious inherited disease. Detection of a P/LP ALT allele does not establish contributor identity, genotype certainty, penetrance, phenotype or disease status. The examples therefore illustrate the type of information recoverable from contamination, rather than a diagnosis or characterization of any individual.

Supplementary Table 1 — Exploratory and systematic replication cohort characteristics

Aggregate cohort characteristics, prevalence, human-fragment burden distributions and run accounting are reported separately for exploratory cohort E and systematic replication cohorts R1–R3. E comprises 1,713 retained runs from 426 BioProjects. Each replication cohort contains 900 selected run memberships; core-evaluable counts are 883, 888 and 889 for R1, R2 and R3, while adjudication-evaluable counts are 883, 888 and 888. R3 therefore retains separate strict and adjudicated denominators because one adjudication result is unresolved. Cohorts are not pooled because their selection designs differ. Exact run/BioProject rosters and original sampling seeds are not included.

Supplementary Data 1 — Supporting analytical summaries

This workbook provides privacy-revised aggregate and figure-reconstruction summaries supporting the main figures, Extended Data figures and Supplementary Notes. It contains cohort-level results and panel-specific numerical observations required to reconstruct plotted measurements. The five-population SNP-recovery cohort underlying Figure 2b–d comprises 346 runs from 177 BioProjects, selected from 853 historical candidates with retained sequencing evidence. Its SNP-configuration recovery and within-project allele-state concordance analyses are distinct from the replication-cohort population-affinity analyses in Figure 2a. These measurements represent read-derived allele observations rather than validated donor genotypes or individual identification. The workbook excludes exact selected run/project rosters, original sampling seeds, accession-to-pseudonym mappings, accession-linked human sequence evidence and full run-level burden vectors. The RNA-seq survey comprises 153 runs, 123 Experiments, 117 BioSamples and 14 BioProjects; detailed identity-linked records remain restricted.

Supplementary Data 2 — Systematic replication analyses

This workbook reports BioProject recurrence, intraclass correlation coefficients (ICC), median odds ratios (MOR), cross-cohort overlap and statistical methods for exploratory cohort E and systematic replication cohorts R1–R3, retaining their distinct sampling designs and endpoint denominators. Replication ICC and MOR estimates are summarized using posterior means and 95% highest-density intervals. The historical exploratory-cohort ICC estimate is likewise reported as a posterior mean with a 95% highest-density interval, whereas its MOR is reported as a posterior median with the retained historical interval. Extended Data Figure 1d presents MOR posterior medians and equal-tailed 95% intervals from the same frozen adjudicated posterior draws. These reporting conventions summarize the underlying posterior distributions differently and do not represent independently fitted models. Strict-endpoint model summaries remain separate from the adjudicated MOR estimates displayed in Extended Data Figure 1d. In R3, 889 runs were strict-evaluable and 888 were adjudication-evaluable because one adjudication result remained unresolved.

References

[1] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015. doi: 10.1038/nature15393.

[2] Alexander Lang, Merve Yilmaz, Christiane Hader, Sammy Murday, Xenia Kunz, Nicholas Wagner, Constanze Wiek, Patrick Petzsch, Karl Köhrer, Julian Koch, Michéle J. Hoffmann, Annemarie Greife, and Wolfgang A. Schulz. Contingencies of UTX/KDM6A action in urothelial carcinoma. *Cancers*, 11(4):481, 2019. doi: 10.3390/cancers11040481.

[3] Thomas S. Dexheimer, Nathan P. Coussens, Thomas Silvers, Eric M. Jones, Li Chen, Jianwen Fang, Joel Morris, Jeffrey A. Moscow, James H. Doroshow, and Beverly A. Teicher. Combination screen in multi-cell type tumor spheroids reveals interaction between aryl hydrocarbon receptor antagonists and E1 ubiquitin-activating enzyme inhibitor. *SLAS Discovery*, 29:100186, 2024. doi: 10.1016/j.slasd.2024.100186.