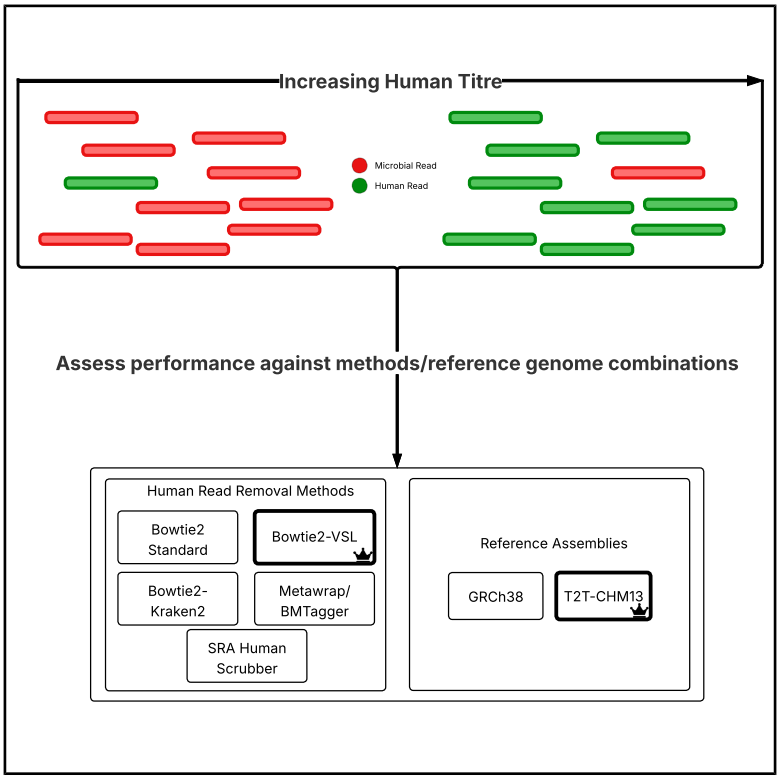
Cell Reports
Methods

Resource

Benchmarking of human read removal strategies for viral and microbial metagenomics

Graphical abstract



Highlights

- Systematic benchmarking of human read removal methods on synthetic and real data
- Bowtie2 “very-sensitive-local” mode with single-end reads gives best performance
- Trade-offs found between sensitivity, specificity, and computational efficiency
- Guidance for choosing and configuring methods to optimize accuracy and reduce risks

Authors

Matthew Forbes, Duncan Y.K. Ng, Róisín M. Boggan, ..., David K. Jackson, Kevin Howe, Ewan M. Harrison

Correspondence

ms65@sanger.ac.uk (M.F.), eh6@sanger.ac.uk (E.M.H.)

In brief

Forbes et al. benchmark human read removal methods for viral and microbial metagenomics. Using synthetic and microbiome datasets, along with a set of benchmark methods, they demonstrate that a custom Bowtie2 configuration (“very-sensitive-local” mode with single end) with the T2T-CHM13 reference best balances host read removal and microbial integrity.

Forbes et al., 2025, Cell Reports Methods *5*, 101218
November 17, 2025
https://doi.org/10.1016/j.crmeth.2025.101218

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Resource

Benchmarking of human read removal strategies for viral and microbial metagenomics

Matthew Forbes,[1,4,5,*] Duncan Y.K. Ng,[2] Róisín M. Boggan,[2,3] Andrea Frick-Kretschmer,[1] Jillian Durham,[4] Oliver Lorenz,[2] Bruhad Dave,[1] Florent Lassalle,[2] Carol Scott,[4] Josef Wagner,[2] Adrianne Lignes,[1] Fernanda Noaves,[1] David K. Jackson,[4] Kevin Howe,[1] and Ewan M. Harrison[2,3,*]

[1]Genomic Surveillance Unit, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[2]Parasites and Microbes Programme, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[3]Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 0AW, UK
[4]Informatics and Digital Solutions, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK
[5]Lead contact
*Correspondence: ms65@sanger.ac.uk (M.F.), eh6@sanger.ac.uk (E.M.H.)
https://doi.org/10.1016/j.crmeth.2025.101218

MOTIVATION Human read contamination in metagenomic data compromises analysis, increases computational burden, and poses privacy risks. While numerous removal methods exist, their accuracy and efficiency vary and the benefits of incorporating newer reference genomes such as T2T-CHM13 remain underexplored. Here, we benchmark leading approaches to establish best practices for reliable and secure human read removal, providing guidance on the choice and configuration of appropriate methodologies.

SUMMARY

Human reads are a key contaminant in microbial metagenomics and enrichment-based studies, requiring removal for computational efficiency, biological analysis, and privacy protection. Various *in silico* methods exist, but their effectiveness depends on the parameters and reference genomes used. Here, we assess different methods, including the impact of the updated telomere-to-telomere (T2T)-CHM13 human genome versus GRCh38. Using a synthetic dataset of viral and human reads, we evaluated performance metrics for multiple approaches. We found that the usage of high-sensitivity configuration of Bowtie2 with the T2T-CHM13 reference assembly significantly improves human read removal with minimal loss of specificity, albeit at higher computational cost compared to other methods investigated. Applying this approach to a publicly available microbiome dataset, we effectively removed sex-determining SNPs with little impact on microbial assembly. Our results suggest that our high-sensitivity Bowtie2 approach with the T2T-CHM13 is the best method tested to minimize identifiability risks from residual human reads.

INTRODUCTION

When performing next-generation sequencing directly on human-derived microbial or viral (referred to henceforth as pathogen) samples (e.g., clinical specimens), a significant proportion of reads will originate from the human host. This problem is particularly acute in shotgun metagenomic studies when there is no or only limited depletion of host DNA. The overall proportions expected will vary depending on the protocol employed. For instance, in microbiome studies, gut microbiome studies with samples derived from feces will typically have a low proportion of human reads (<10%), whereas saliva, nasal, skin, and vaginal samples can often be majority human in composition (>90% reads).[1]

This can have two important effects, relating to both downstream analyses and legal and ethical concerns. The human genome is around a 1,000-fold larger than the average bacterial genome and around 100,000-fold larger than a respiratory virus genome such as SARS-CoV-2 (~30 Kb for SARS-CoV-2 cf. ~3.2 Gb for the human genome). This means that a sample containing equal copies of human and pathogen genomes would result in approximately 99.9% of reads being of human origin,[2] which can have detrimental effects on the quality of variant calls.[3] Additionally, the increasing amount of human whole-genome sequences or genotyping data deposited in public archives raises a risk for potential identification of the individual from whom the sample was derived. This raises numerous concerns around privacy, lack of informed consent, and legal and regulatory compliance.[4] This risk can present a barrier to open and timely data sharing. Removal prior to analysis is important as it ensures reproducibility with data later submitted to public archives while optimizing data to mitigate erroneous calling,

with the additional benefit of reducing the size of data files used for complex downstream analysis, thus reducing the energy requirement for the bioinformatics analysis.[5]

Some methods seek to remove human reads *in vitro*. One method is to employ chemical depletion of host DNA. However, this can lead to issues such as incomplete removal of host DNA, loss of pathogen DNA, bias toward certain pathogen populations, and increased protocol complexity, time, and cost.[1] Other methods enrich the microbial component, for example, 16S ribosomal RNA sequencing and SARS-CoV-2 ARTIC amplicon sequencing[6]; these methods will reduce the numbers of human reads to negligible levels via specific amplification of target gene (s) or genomes; however, this comes at the expense of strain-specific information in other parts of microbial genomes and can only be employed to investigate a limited number of different species.

Another method is bait capture, where a synthetic oligonucleotide probe with complementarity to target sequences is used to selectively enrich the sequencing library with particular target genes or genomes. Baits can be designed for a range of targets, including respiratory and vector-borne viruses and bacteria. This can be very useful for enhancing the signal for a group of pathogens from the background and avoids issues associated with amplification, such as primer artifacts and dropouts; however, a background of human reads is still generated. As such, it is important to remove human reads *in silico*.

Computational human read removal (HRR) methods can often be broadly divided into the following classes: alignment and classifier-based. The alignment-based methodologies involve aligning reads to a human reference and retaining only the reads that do not align. Classification-based methods involve predicting the taxonomic identity of individual reads and removing reads that are classified as human.

Various HRR methods have previously been assessed by Bush et al.,[2] who investigated several alignment and classifier-based methodologies on a benchmark of multiple simulated mixtures of bacterial species. The authors found that alignment-based methodologies more accurately located human reads compared to classification-based methods because aligners directly compare the reads with the human genome sequence. They note a trade-off in terms of computational efficiency, with alignment-based methods working on substantially longer timescales as well as requiring greater computational resources. For short reads (<150 bp), the combination of two short-read aligners, Bowtie2 followed by SNAP, was the most effective HRR combination; however, they note the higher computational cost of combining the two alignment-based methods. Though the authors discuss a comparison of methodologies, they broadly implemented default parameters across the tested methodologies; however, aligners such as Bowtie2 can be further parameterized and run in alternative ways to enhance their efficacy for HRR, offering potential improvements for this use case.

Bush et al. also compared HRR performance with aligners and classifiers against the GRCh37[7] and GRCh38[8] references and found no difference between methods that use the different reference genomes.[2] However, since the publishing of this paper, a more complete telomere-to-telomere (T2T-CHM13) reference has been published, comprising the first full-length human genome sequence.[9] Compared to the GRCh37 assembly, GRCh38 was an improvement on the previous best reference, but T2T-CHM13 represents a paradigm shift. The use of long-read technologies meant that the T2T-CHM13 assembly was able to resolve previously "dark" regions in the previous assembly, including centromeres, telomeres, and ribosomal DNA (rDNA) arrays.[9] This offers potential for large improvements in the ability of alignment-based HRR tools to improve their ability to detect human reads that previously fell into these "dark zones" and ensure their removal from the dataset.

In this study, we investigated a number of HRR methods and their capacity to separate human reads from microbial reads (viral and bacterial). We created a synthetic titration benchmark dataset based on a gradient of bait-capture-derived reads mixed with reads derived from human samples derived from the 1000 Genomes Project[10] across multiple populations to ensure a range of demographic representation. A read-fate analysis was performed on multiple HRR methods using the synthetic titration dataset, comprising combinations of alignment and classification-based methods to identify the best-performing method and investigate performance differentials. We then applied our chosen method to a publicly available microbiome dataset and showed the effect that application has on the identifiability of individuals and confirm no adverse effects of implementing this methodology on either metagenomic-assembled genomes (MAGs) or consensus FASTA generation.

RESULTS

Comparison of methodology performance

We sought to identify the best practice for HRR from microbial metagenomics short read data (shotgun and bait capture). To do this, we investigated the performance of multiple HRR approaches (see details of selection in STAR Methods and Table 1) against a synthetic titration dataset constructed from viral reads mixed in with varying proportions of human reads (see Figure 1 and STAR Methods).

We defined a true positive (TP) as a read correctly classified as human, a true negative (TN) as a read correctly classified as non-human, a false positive (FP) as a non-human read incorrectly classified as human, and a false negative (FN) as a human read incorrectly classified as non-human. Using these definitions, we calculated a set of "all-round metrics"—accuracy, F1 score, and Matthews correlation coefficient (MCC)—to provide an overall view of method performance. Accuracy reflects the proportion of correct classifications overall but can be misleading when human and non-human reads are imbalanced. The F1 score combines information about correct and missed human read predictions but does not take the number of TNs into account. While this makes it less informative about the preservation of non-human reads, comparing F1 scores to accuracy and MCC across varying human titers can help indicate whether changes in performance are driven more by gains in HRR or losses in microbial retention. MCC, by contrast, incorporates all four components of the confusion matrix and provides a more balanced view, particularly under varying class distributions.

Table 1. Details of the HRR methods investigated in this study

Method name	Primary removal tool	Configurations	Wrapper	Human reference(s)
Bowtie2 Standard	Bowtie2 (v.2.5.4)	standard Bowtie2 parameters	HOSTILE (v.1.1.0)	T2T-CHM13 and GRCh38
Bowtie2 VSL	Bowtie2 (v.2.5.4)	"very-sensitive-local" parameter switched on and mapping performed on fastq pairs in "unpaired" mode	Kneaddata (v.0.1.2) (all functionality except human read removal switched off)	T2T-CHM13 and GRCh38
Bowtie2- Kraken2	Bowtie2 (v.2.5.4) for alignment, Kraken2 (v.2.1.3) for subsequent classification	standard parameters for Bowtie2 and Kraken2	Nextflow v.24.04.4	Bowtie2: T2T-CHM13 and GRCh38 Kraken2: Standard – 16 GB database from Ben Langmead (https://benlangmead.github.io/aws-indexes/k2)
Metawrap-QC	BMTagger (v.3.101)	standard BMTagger configurations	Nextflow (https://github.com/sanger-pathogens/generate_mags)	T2T-CHM13 and GRCh38
SRA Human Scrubber	SRA Human Scrubber (v.2.2.1)	standard SRA Human Scrubber configurations	SRA Human Scrubber (v.2.2.1)	20240718 human database filter (https://ftp.ncbi.nlm.nih.gov/sra/dbs/human_filter/human_filter.db.20240718v2)

We also calculated targeted metrics—precision, sensitivity (also known as recall), and specificity—to explore specific aspects of performance. Precision reflects how many of the reads removed were truly human, sensitivity captures how effectively human reads were identified and removed, and specificity indicates how well non-human reads were correctly retained.

All methodologies tested showed good performance in HRR across the synthetic titration benchmark, with the minimum MCC score at any human titer for any method of 0.94 (Figure 2A). As the human titer is increased, accuracy and MCC show a pattern of decay across all methods at similar rates, with no method converging or crossing across human titers. This decay is not shown for the F1 score (the harmonic mean of precision and recall), indicating that this performance decay was driven mainly by the increasing number of FNs (Figure S1), which is not a component of the F1 score.

Across the methodologies investigated, the best all-around performing methods were those utilizing Bowtie2 with the "very-sensitive-local" (VSL) configuration, as indicated by the clear difference in all-around performance showcased through accuracy, F1, and MCC scores. The Bowtie2-Kraken2 methods had the next best performance, with the poorest performance exhibited by the Bowtie2 Standard methodology with the GRCh38 human reference (Figure 2A).

Investigation of specific components of the overall performance indicated that the increase in sensitivity conferred by the Bowtie2 VSL methodology also resulted in a very small drop in specificity (TN rate, the proportion of actual negatives correctly identified by a test), indicating that this method was less specific than other methods and more prone to calling FPs than other methods. This difference is marginal, with a reduction in specificity of ~0.00003 from the highest specificity method (SRA Human Scrubber) and with a median false positive count of 5 reads with a maximum of 375 (the Bowtie-Kraken2 method has a higher maximum FP rate, see Figure S1; Table S1). This marginal reduction in specificity does not overcome the larger increase in sensitivity (TP rate, the proportion of actual positives identified by a test), contributing to its greater performance in the overall metrics compared to other methods.

Overall, the best-performing method, Bowtie2 VSL T2T-CHM13, exhibited an approximately 3.8-fold reduction in the number of FN reads (i.e., human reads incorrectly retained by the method) compared to the worst-performing method, Bowtie2 Standard GRCh38. It also exhibited an approximately 1.3-fold decrease compared to the next-best T2T-CHM13 method in Bowtie2-Kraken2.

The differences in human reference sequence appeared to have a different effect according to the different HRR methodologies (Figure 2B). All methodologies showed a significant improvement in HRR when the T2T-CHM13 reference sequence is applied according to the MCC scores (Wilcoxon signed-rank test p value <0.05). The extent of this was variable between methods, with Bowtie2 Standard and MetaWrap-QC showing the largest difference between the two methods, with a much smaller difference for the Bowtie2 VSL and Bowtie2-Kraken2 methods (Bowtie2 VSL also shows a higher test statistic, which suggests that the rank scores vary more within the dataset, and as such, there is less distinction between the two populations being compared; Figure 2B). This indicates that the sensitivity of the method employed is a more significant component of HRR performance than the reference sequence used and that the more sensitive methods (Bowtie2 VSL and Bowtie2-Kraken2) are less likely to be improved further by improvement in human genome assemblies reducing the maintenance burden of these methods.

We also examined resource usage across the tested methods (Figure 2C). The VSL parameterization had a significant impact on overall runtime (Wilcoxon rank-sum test, $p < 0.05$), resulting in a median increase of 6.42 min compared with the standard Bowtie2 implementation. However, memory consumption was

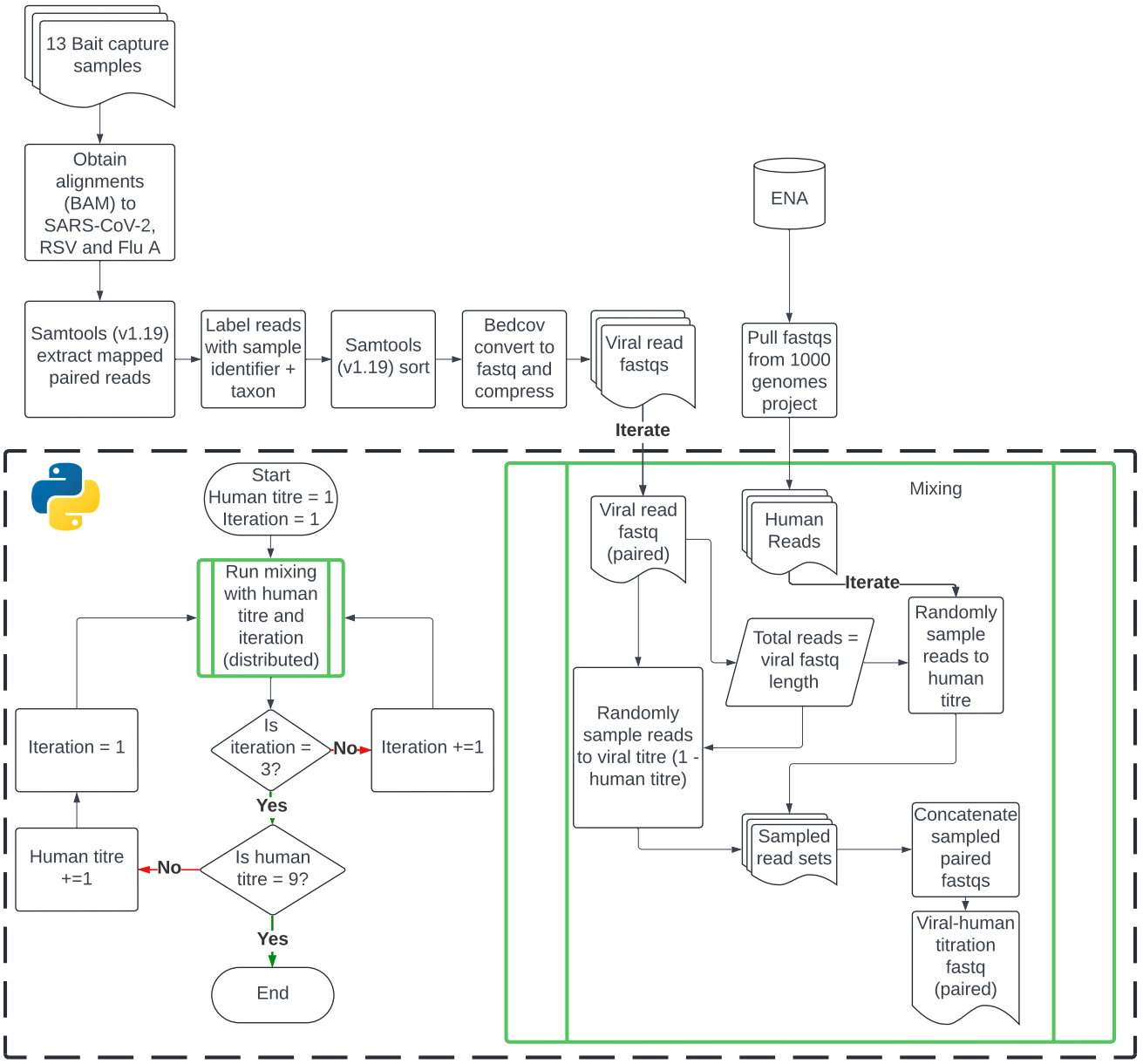


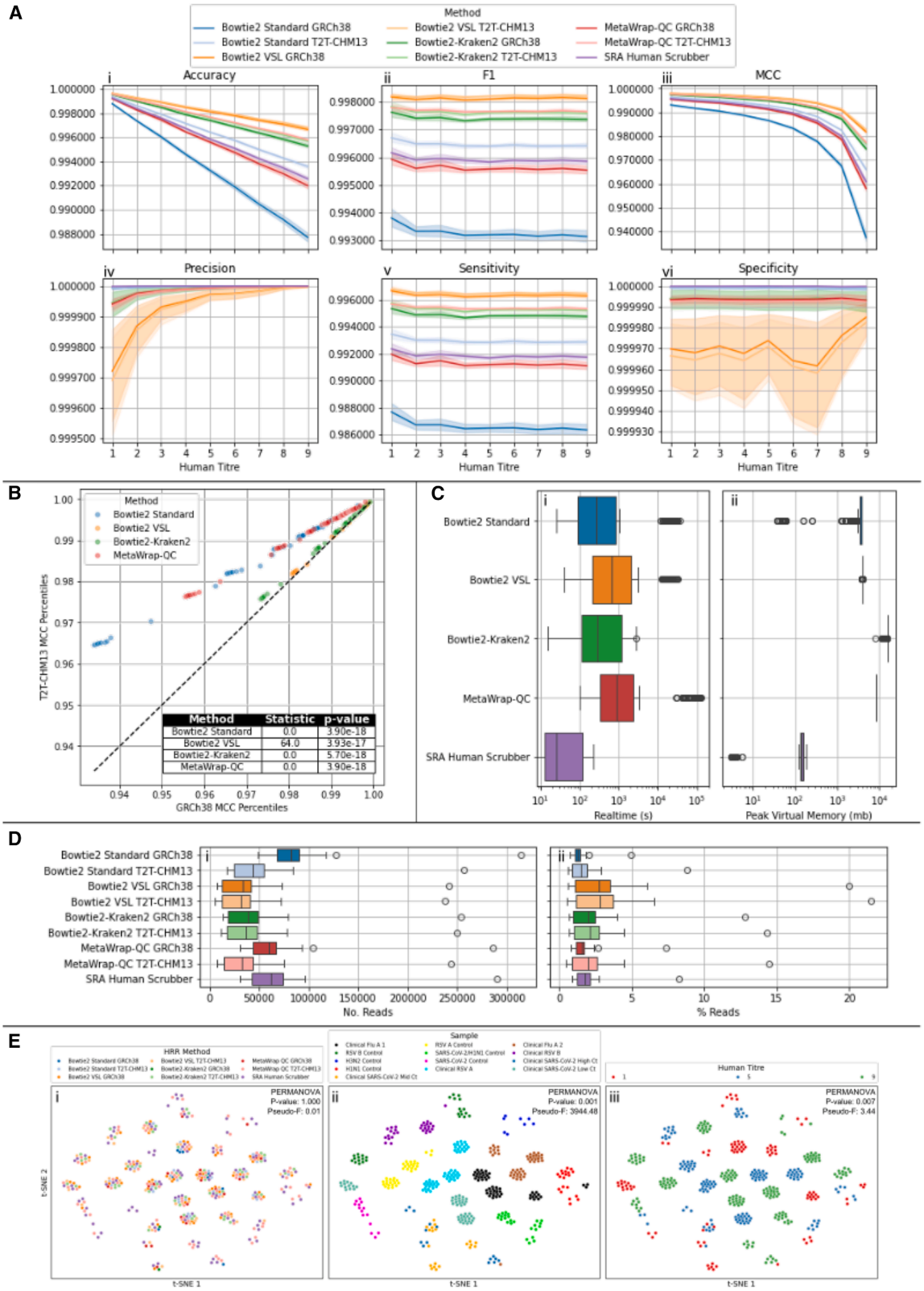
Figure 1. Flowchart depicting the process for developing the synthetic mixture titration dataset

similar between Bowtie2 VSL and the standard Bowtie2, while Bowtie2-Kraken2 and MetaWrap exhibited higher memory usage. In contrast, the SRA Human Scrubber was markedly more efficient than the other methods, with substantially lower runtime and peak memory consumption.

A key focus for HRR methodologies is the effective removal of host reads; therefore, we focused the analysis on the composition of reads that were recorded as FNs (reads arising from human samples that were not removed by a given methodology). As these FNs all arose from samples from the 1000 Genomes Project, we investigated the extent to which these retained reads were true human reads rather than probable microbial and/or other contaminants in the 1000 Genomes samples. To this effect, FN read sets were processed with Kraken2 against the "maxikraken" database (see STAR Methods), and their compositions were compared between methods.

As a percentage of FN reads, the Bowtie2 VSL methods exhibit significantly (Wilcoxon rank p value <0.05) increased microbial reads compared with other methods, indicating that human and unclassified reads make up for a much smaller component of these read sets resulting from these methods (Figure 2D, i). At the same time, we can see that the overall number of microbial reads retained by the VSL methods is decreased for these methods compared with other methods (Figure 2D, ii). Taken together, this provides a stronger indication that these methods have a greater propensity to deplete potential human

CellPress
OPEN ACCESS



(legend on next page)

reads from the dataset, however, with greater potential to additionally deplete some real microbial reads. FN reads that are either classified as human or remain unclassified after Kraken2 processing did not exhibit bias toward certain genetic ancestries (Figure S2). The HRR method employed appears to be a much stronger component of read retention than that of human genetic ancestry.

To assess the impact of different HRR methods on viral consensus genome FASTA generation, we investigated genome coverage patterns across different viral species with human titers 1, 5, and 9. Dimensionality reduction revealed that clustering patterns of consensus fasta coverage were primarily driven by the sample and differences in human titers (Figures 2E i and ii), with little to no contribution from the HRR methods (Figure 2E, iii). In line with this, PERMANOVA showed no significant grouping based on HRR methods (Figure 2E, iii). Collectively, these findings suggest that the HRR process had minimal influence on the overall viral consensus genome quality.

Evaluation of external metagenomic data

To evaluate the effect of HRR methods, we applied the approach to a publicly available shotgun metagenomic study. HRR had already been performed by the authors using Bowtie2 Standard GRCh38, prior to the dataset being uploaded into the public domain. We observed a significant difference in richness (number of observed bacterial species), Shannon diversity (diversity of the community taking into account the number of observed bacterial species and relative abundance), Inverse Simpson (similar to Shannon diversity but emphasizes on community evenness), and Pielou's evenness (how evenly distributed the species are in a community) between raw to SRA Human Scrubber, Bowtie2 VSL GRCh38, and Bowtie2 VSL T2T-CHM13 (Figure 3C). However, no significant differences were found between Bowtie2 VSL GRCh38 and T2T-CHM13. Bowtie2 VSL, regardless of reference, reduced detectable bacterial species from 31 in raw to 11 (Table S2). Shannon diversity and Inverse Simpson also decreased. Notably, Bowtie2 VSL increased evenness, suggesting preferential removal of less abundant species.

We evaluated the effect of HRR methods on beta diversity, which measures differences in microbial community composition between samples, using Bray-Curtis dissimilarity on relative abundance. Significant differences were observed between raw and SRA Human Scrubber ($R^2 = 0.0009$, $p = 0.003$), Bowtie2 VSL GRCh38 ($R^2 = 0.00357$, $p = 0.001$), and Bowtie2 VSL T2T-CHM13 ($R^2 = 0.00364$, $p = 0.001$). However, ordination did not show clear separation between clusters (Figure S3A). Variation partition analysis attributed 0.36% of variation to HRR methods and 97.26% to samples (Figure S4). After accounting for sample variation, a slight separation between methods was observed (Figure S3B), with Bowtie2 VSL clustering together regardless of the reference genome used. Bray-Curtis dissimilarity and Jaccard index scores were used to quantify differences (Figure 3C). SRA Human Scrubber had median scores of 0.05 (Bray-Curtis) and 0.1 (Jaccard). Bowtie2 VSL showed similar median scores across references: Bray-Curtis 0.067 (GRCh38) and 0.068 (T2T-CHM13); Jaccard 0.125 (GRCh38) and 0.1268 (T2T-CHM13).

Next, we compared the determined average nucleotide identity (ANI) of two common species found in the nasal microbiota, *Staphylococcus aureus* and *Corynebacterium accolens*, across the HRR methods (Figure 3D). We hypothesized that ANI increases when human reads misclassified as bacterial are removed, while ANI decreases when bacterial reads are misclassified and removed (Figure 3A). *S. aureus* ANI scores differed significantly between the raw dataset and Bowtie2 VSL (GRCh38 and T2T-CHM13) but not between raw and SRA Human Scrubber. No significant difference was observed between Bowtie2 VSL GRCh38 and T2T-CHM13, suggesting that Bowtie2 VSL itself, rather than the reference genome, drives the difference. SRA Human Scrubber improved *S. aureus* ANI scores, whereas Bowtie2 VSL reduced them (Figure 3E), suggesting that *S. aureus* reads are removed by Bowtie2 VSL. For *C. accolens*, ANI scores differed significantly between raw and both HRR methods but not between SRA Human Scrubber and Bowtie2 VSL, indicating similar performance. Notably, Bowtie2 VSL performed better than SRA Human Scrubber at maintaining or improving *C. accolens* ANI scores, suggesting that the extent of misclassified bacterial reads may vary by species.

To identify species removed by different HRR methods, we performed differential abundance analysis using MaAsLin2. Of the 3,115 species detected, 55 were depleted due to HRR (Figure 3F), primarily from the phyla *Pseudomonadota* and *Actinomycetota*. Bowtie2 VSL GRCh38 and T2T-CHM13 had

Figure 2. Comparison of HRR methods across the synthetic titration dataset

(A) Performance metrics across the dataset, top row (i–iii) shows overall classification performance: accuracy (i) reflects the proportion of correctly classified reads but can be misleading with class imbalance; F1 score (ii) balances precision and recall, making it useful when FPs and FNs have different consequences; Matthews correlation coefficient (MCC, iii) provides a more robust measure by incorporating all elements of the confusion matrix. The bottom row presents class-specific metrics: precision (iv) indicates the proportion of correctly identified human reads among those classified as human, sensitivity (v) shows how effectively human reads are removed, and specificity (vi) measures how well viral reads are retained. Lines indicate the mean across all samples and iterations at each human titer, with shading representing the 95% confidence interval. Different y axis scales between plots should be considered when interpreting values.

(B) Percentile-percentile plot showing the differences in MCC distributions between methods run with the GRCh38 and T2T-CHM13 human assemblies as reference sequences. Table shows the output of Wilcoxon signed-rank test for each method between the GRCh38 and T2T-CHM13 distributions.

(C) Boxplot showing differences in run time and resource usage between the different methodologies. (i) Real time for process completion in seconds. (ii) Peak Virtual Memory used in the completion of the process in megabytes.

(D) Boxplots showing distributions of FN (reads derived from human sources retained by HRR methods) reads, which are classified as microbial (i.e., neither human nor unclassified) by kraken2 against the maxikraken database across the synthetic titration dataset. (i) Number of FN reads classified as microbial, (ii) percentage thereof.

(E) t-SNE embeddings of consensus sequence coverage after HRR and comparing different human read titers. Points colored by the HRR method (i), sample (ii), and human titer (iii), with PERMANOVA p value and pseudo-F values for each grouping indicated in each plot.

CellPress
OPEN ACCESS

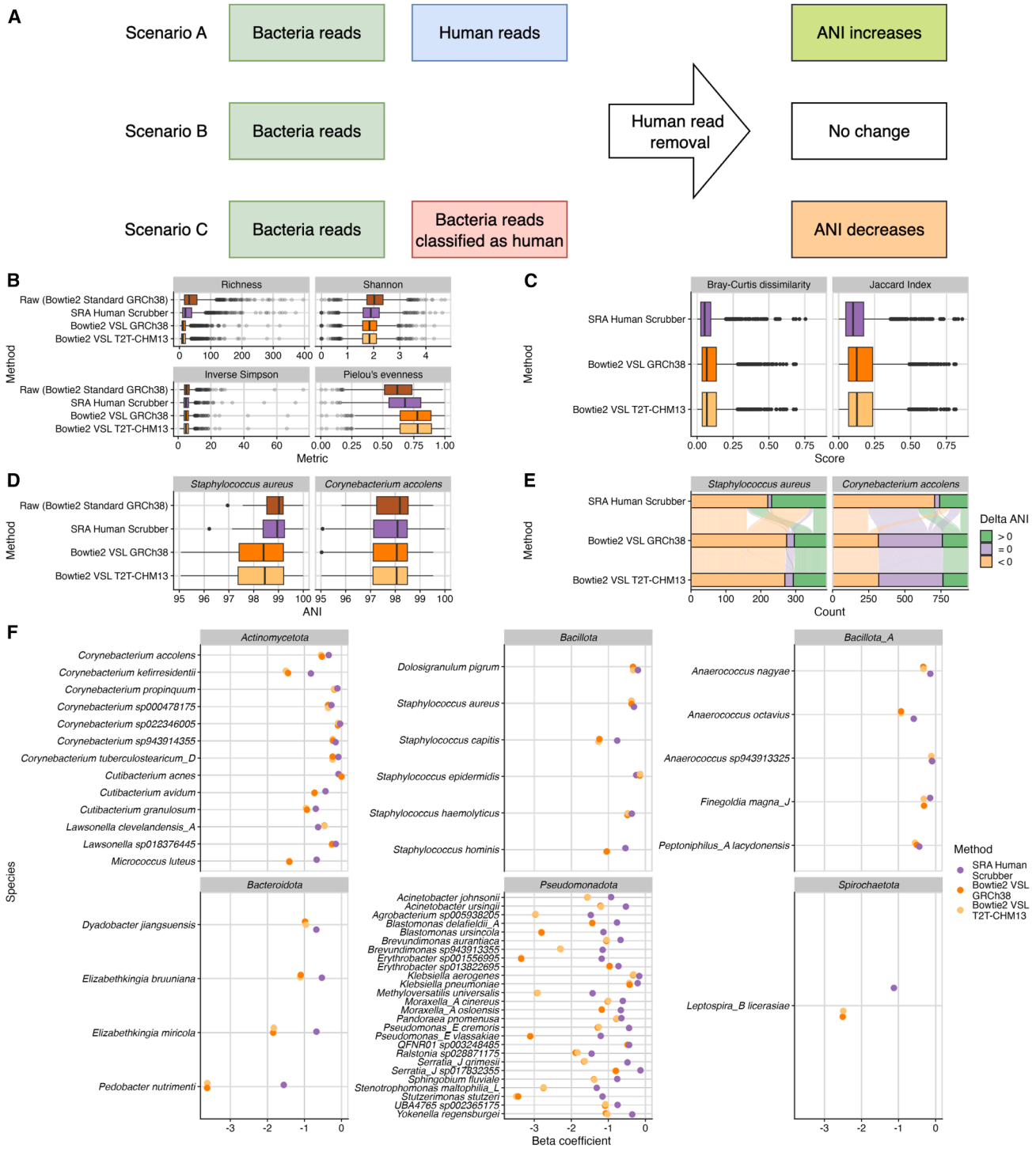


Figure 3. Comparison of metagenomic outputs arising from metagenomics analysis post implementation of HRR methods

(A) A flowchart illustrates how human read removal (HRR) affects average nucleotide identity (ANI) scores across different scenarios. ANI is calculated using Sylph, which utilizes k-mers and is sensitive to changes in read composition. In scenario A, ANI is calculated from both bacterial and human reads, where human reads resemble bacterial sequences, leading to a lower reported ANI score. Removing human reads increases the ANI score. In scenario B, no human reads interfere with ANI calculation, so applying HRR has no effect. In scenario C, some bacterial reads are misclassified as human reads and removed during HRR, reducing the total reads used for ANI calculation and lowering the ANI score.

(B) Alpha diversity comparison between three HRR methods and a standard HRR approach, Bowtie2 Standard GRCh38.

(legend continued on next page)

similar beta coefficients (mean delta: 0.0001). SRA Human Scrubber had a lower beta coefficient than Bowtie2 VSL, with a mean delta of 0.58 for both GRCh38 and T2T-CHM13 (Table S3).

In addition to quantifying the impact of HRR methodologies on microbiome diversity, we assessed the post-HRR samples for the presence of residual human reads. The aim of this analysis was to determine the level of human genetic information retained in post-HRR samples that could potentially be used to re-identify individuals. Reads were retained across autosomes and sex chromosomes in samples processed with the Bowtie2 Standard GRCh38 reference method and the SRA human scrubber method; Bowtie2 VSL with a GRCh38 reference genome retained notably fewer reads than both these methods (Figure 4A). Bowtie2 VSL with T2T-CHM13v2.0 reference genome did not detect retained reads, which was expected as the same reference genomes were used for initial HRR and subsequent re-alignment, but we have included this comparison for completeness.

Improvements made by the T2T-CHM13v2.0 reference genome include the resolution of highly repetitive regions, such as the human satellite array on chromosome 9, a region that has previously been of low mappability.[12] It is notable that a larger number of human reads were retained on chromosome 9 by the Bowtie2 Standard or SRA Human Scrubber methods (Figure 4B). This demonstrates the importance of using a high-quality reference genome for HRR.

We earlier highlighted the ethical concerns that surround the ability to re-identify an individual from human reads retained in microbiome data. This was previously tested in a cohort of individuals with gut microbiome samples and corresponding genomic samples.[4] Without access to human biobank data associated with the nasal microbiome samples, we are unable to fully test the ability to correctly match microbiome samples to human genetic samples.

As an approximation of re-identification, we performed a sex determination analysis of samples post HRR (Figure 4C). Sex determination was possible for all samples in the publicly uploaded nasal microbiome shotgun data (mean Y chromosome read depth in male samples, rd = 0.692) and remained possible after HRR with SRA Human Scrubber (rd = 0.420) or Bowtie2 VSL with a GRCh38 reference genome (rd = 0.00673), although with diminishing confidence. SRA scrubber and Bowtie2 VSL GRCh38 methodologies appear to be proficient in removing X chromosome reads but retain a larger proportion of Y chromosome reads, Bowtie2 VSL GRCh38 to a much lesser extent, reflecting the improved performance of the Bowtie2 VSL method. The impact of reference genome is highlighted by the differences observed between Bowtie2 VSL using the GRCh38 reference genome and the T2T-CHM13v2.0 reference genome, reflecting the improvements made by the T2T-CHM13v2.0 reference genome, including the more accurate representation of the Y chromosome included in the T2T-CHM13v2.0 reference sequence.[13]

DISCUSSION

This investigation demonstrates that the VSL mode in Bowtie2 significantly enhances HRR performance by increasing alignment sensitivity at the cost of longer runtime (Table 2). Unlike the default mode, VSL enables local alignment, allowing soft-clipping of non-matching bases at read ends, making it more tolerant to sequencing errors. The seed length is reduced from 22 to 20 bases, increasing sensitivity but also computational time, as shorter seeds match more frequently across the genome. Additionally, VSL increases re-seeding attempts from 2 to 3, improving alignment chances when the initial seed fails. The seed interval factor is lowered from 1.0 to 0.5, resulting in denser seed placement, which enhances alignment probability but requires more processing. Finally, the allowed backtracking is increased, allowing Bowtie2 to explore more alternative alignment paths. Collectively, these modifications increase the likelihood of mapping reads to the human reference but require additional computational resources.

Another key difference in the VSL methodology we have implemented is that reads are aligned to the reference by Bowtie2 in "unpaired" mode. This has the ability to confer several further sensitivity gains aside from the implementation of the VSL mode alone, including the removal of constraints around insert size restrictions, proper pairing of reads, greater sensitivity to partial matches (e.g., where one mate appears human and the other microbial), and potential handling of structural variants within human genomes, which may disrupt insert sizes. Furthermore, the removal of the requirement to compute pair restraints in unpaired mode likely reduces resource requirements, hence partly compensating for the increased resource usage of the VSL modality.

The combination of the parameter changes, allowance of soft-clipping (which Bowtie2 does not permit in its default end-to-end mode, unlike aligners such as BWA-MEM that perform local alignment by default), and removal of pairing constraints have a clear positive effect on the sensitivity of the Bowtie2 VSL methodology; however, this increase in sensitivity has obvious potential drawbacks, including additional computation time and a decrease in specificity. Indeed, we did observe a marginal decrease in specificity for the Bowtie2 VSL method over other methodologies. However, when comparing the difference between the top T2T-CHM13 methods (Bowtie2 VSL and MetaWrapQC), the difference in median sensitivity is approximately 191 times that of specificity. This is reflected in the fact that the overall metrics (accuracy, F1, and MCC) are in agreement that Bowtie2 VSL has the most optimal overall performance. Additionally, this decrease in specificity is reflected only in the removal of 5 viral reads in the median case, with 375 (out of a total of 8,005,608 reads) in the worst case (the worst case across the whole dataset is 429 reads from the Bowtie2-Kraken2 method out of a total of 8,563,378).

(C) Bray-Curtis dissimilarity and Jaccard index score of three HRR methods against the raw.
(D) Visualization of the ANI of two species commonly found in the nasal microbiota.
(E) Delta ANI score of three HRR methods against the raw.
(F) A differential analysis of HRR methods compared to the raw showing the depletion of bacterial species.

CellPress
OPEN ACCESS

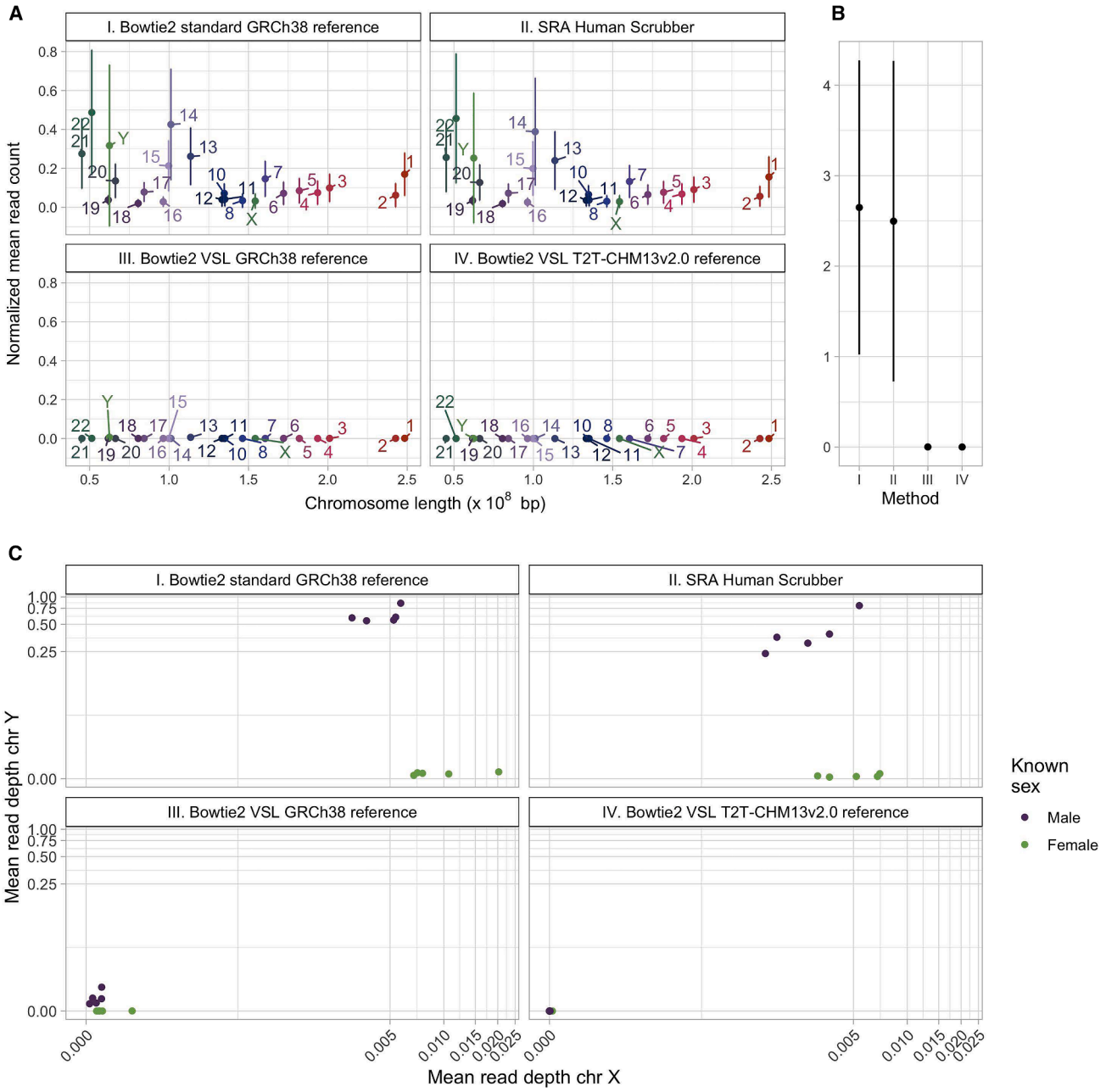


Figure 4. Post HRR analysis of public human nasal microbiome data

(A) Normalized mean read count when using (I) Bowtie2 Standard with a GRCh38 reference (data as released by Ju et al.[11]), (II) SRA human scrubber, (III) Bowtie2 VSL with a GRCh38 reference, and (IV) Bowtie2 VSL with a T2T-CHM13v2.0 reference.

(B) Normalized mean read count on chromosome 9.

(C) Plots compare the pseudo-log mean read depth of chromosome X and Y post HRR using (I) Bowtie2 Standard with a GRCh38 reference, (II) SRA human scrubber, (III) Bowtie2 VSL with a GRCh38 reference, and (IV) Bowtie2 VSL with a T2T-CHM13v2.0 reference.

Bowtie2 VSL has an effect on microbial composition, but the effect size is small. After HRR by Bowtie2 VSL, rarer species are excluded, leading to a more even community. HRR can introduce batch effects, and if multiple HRR methods are applied unevenly, they may bias the dataset. Authors should clearly report each step to ensure reproducibility. Bowtie2 VSL can also misclassify bacterial reads with GC content similar to humans (40.9%). This effect was more pronounced in *S. aureus* (GC 32.8%) than *C. accolens* (GC 59%), possibly due to Sylph's k-*mer*-based microbiota quantification, which is more sensitive to small changes in read composition.[14]

Table 2. Differences in Bowtie2 parameters between the default and the VSL modes

Mode	Alignment type	Seed length	Backtracking	Re-seeding	Seed interval factor
Default	end-to-end	22	15	2	1
Very-sensitive-local	local	20	20	3	0.5

Additionally, it is clear that implementation of Bowtie2 VSL in combination with the T2T-CHM13v2.0 reference genome is the most effective for removing host-contaminating reads, consequently providing the greatest security against identification, demonstrated by the reduced ability to identify the sex of the individual host. As such, although a small effect on microbial diversity is observed, this is outweighed by the reduction in the ability to identify human individuals.

More broadly, our investigation indicates that the selection of an HRR method and its configurations likely depends mostly on scalability and robustness. We observe that overall metrics such as accuracy and MCC show a reduction in overall performance as the proportion of human reads in the titration set is increased. This is likely driven by two main factors: the presence of more human reads translating into an increased chance of misidentification and the inclusion of more true microbial contaminants from the 1000 Genomes samples. Moreover, as the sensitivity remained relatively constant through the titration series for all methods, there is no indication that methods should be chosen on the basis of likely human titer. We also found no evidence of bias based on human genetic ancestry in the removal of human reads, with methodology seemingly the most important component. Therefore, without questions of scalability, the implementation of Bowtie2 VSL T2T-CHM13v2.0 would appear to be optimal.

When we consider scalability, however, we must take into consideration that Bowtie2 VSL shows a median time to completion of around 10.91 min, which is an increase of around 6.42 min per sample longer than the median time for the Bowtie2 Standard configuration. While this might be appropriate for users without time constraints or those with large computational resources allowing them to highly parallelize the process, these timescales can become burdensome. The memory profile, however, was fairly similar between the Bowtie2 Standard and VSL methods, with the addition of the Kraken2 classification step increasing the memory usage by around 4-fold.

By far, the most efficient tool by both runtime and memory usage was the SRA Human Scrubber, which uses the fast STAT methodology combined with a pre-indexed version of the human genome covering most regions of interest.[15] This allows the Scrubber to take a more lightweight approach to HRR, performing a fast scan over a large proportion of the identifiable human genome rather than a full-scale alignment. Our investigation indicates that this approach results in the highest specificity of any method in the benchmark, with a maximum of only 2 FP calls across the entire benchmark. However, we also observe that the scrubber's sensitivity is substantially lower than that of the best-performing methodology, Bowtie2 VSL, and is also lower than that of the Bowtie2 Standard implementation with the T2T-CHM13v2.0 reference genome. This indicates that a greater proportion of human reads will make up these samples, around 2,592 more reads than Bowtie2 VSL in the median case, which could have downstream effects on identifiability of human sequences, the composition of consensus sequences, and on downstream runtime. Our investigation indicates that coverage of consensus sequences does not appear to be affected by the use of either the SRA Scrubber or the Bowtie2 VSL methodology compared with other methods.

Overall, our investigation indicates that all the methods tested here are likely to be suitable for HRR in a variety of experiments; all the methods show high sensitivity and specificity. For the greatest reduction in potential risk for submission to public archives, the Bowtie2 VSL is most likely to remove the largest proportion of human reads and therefore reduce downstream risk to the greatest extent. There are two potential caveats to this: the potential removal of a larger number of microbial reads and the increased resource usage of the higher sensitivity bowtie2 methodology. In terms of incorrect removal, this analysis indicates that this effect, while present, is minimal and outweighed by the positive effect of increased removal of host reads and subsequent reduction in identification risk.

Users looking to minimize microbial read loss may wish to prioritize methods that offer greater specificity; in this study, this includes the Bowtie2 Standard method and the SRA Human Scrubber; however, this will likely entail a reduction in sensitivity, therefore increasing the risk of identifiability in your samples. Researchers may wish to consider the specifics of their study when selecting an HRR strategy, particularly if the accurate retention of certain microbial sequences is critical. For example, studies investigating human-associated viruses such as human papillomavirus (HPV) or Epstein-Barr virus (EBV), which can integrate into the host genome, may risk losing relevant reads due to sequence similarity with human DNA by implementation of the Bowtie2 VSL method.[16] In such cases, an alternative approach could involve first aligning raw reads to the microbial genome of interest using a sensitive mapping protocol to retain likely relevant reads, followed by HRR on the remaining reads using Bowtie2 VSL. This two-step approach can help preserve biologically meaningful signals while still ensuring rigorous HRR.

For resource usage, tools such as the SRA human scrubber are available; however, this investigation indicates that usage of these tools incurs a compromise on downstream data security. Instead, further developments could be made to implement methods that match the sensitivity of Bowtie2 VSL while reducing resource usage; a potential example may be the implementation of HISAT2, which may be able to offer similar sensitivity but reduce resource consumption through its usage of a hierarchical index, improved memory consumption, and cache optimization.[17]

Finally, when using a reference or reference k-mer database for alignment-based removal, the use of the latest T2T-CHM13v2.0 reference is recommended; however, using more sensitive methods such as Bowtie2 VSL can help to ensure

that your method remains robust to improvements in human genome assemblies without creating new vulnerabilities.

Limitations of the study

While this study provides a comprehensive benchmark of HRR approaches across synthetic and real-world datasets, several limitations remain. First, we focused on short-read sequencing data, and the performance of these methods may differ for long-read technologies. Second, although the synthetic titration mixtures included diverse human ancestries, they cannot fully capture the complexity of real clinical samples, which may contain additional contaminants or uneven coverage. Third, we concentrated on a subset of widely used tools and parameterizations; emerging or less commonly applied methods were not assessed. Finally, our evaluation emphasized alignment- and classification-based approaches against GRCh38 and T2T-CHM13 references, but future developments such as pangenome-based references may provide a richer representation of human genetic diversity and further influence HRR performance.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Matthew Forbes (ms65@sanger.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All sequencing datasets generated in this study have been deposited in the European Nucleotide Archive (ENA) under project accessions ERP169280 (viral read sets), ERP169281 (synthetic titration mixtures), and ERP169282 (post-HRR processed mixtures). Publicly available metagenomic data from the human nasal microbiome (CRA006819) were obtained from the Genome Sequence Archive (GSA) and re-analyzed as described.
- Benchmarking and HRR analyses were conducted using the pipeline available at https://github.com/wtsi-npg/hrr_tools/tree/benchmarking_paper-2025 (DOI: https://zenodo.org/records/17136543). Viral consensus sequence generation was performed using the viral-lens pipeline, available at https://github.com/genomic-surveillance/rvi-viral-lens.
- Any additional information required to re-analyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The authors would like to thank Antonio Marinho, Frank Schwach, Katherine Figueroa, Thomas Maddison, Simon Suddaby, Marissa Knoll, and Diego Teixeira for contributions to discussions of the work. Additionally, we would like to thank Katie Bellis, Steve Walton, Quan Lin, and Jayvant Desale for their support in submission of data to ENA and Jaime Tovar Corona, Salih Tuna, Waleed Osman, and Kevin Lewis for precursor evaluations leading to the work. This work has been supported by the Wellcome Trust [220540/Z/20/A].

AUTHOR CONTRIBUTIONS

Conceptualization, M.F., D.K.J., K.H., and E.M.H.; methodology, M.F., E.M.H., K.H., D.Y.K.N., R.M.B., and B.D.; software, J.D. and O.L.; validation, M.F., D.Y.K.N., R.M.B., A.F.-K., F.L., C.S., J.W., and D.K.J.; formal analysis, M.F., D.Y.K.N., R.M.B., and A.F.-K.; investigation, M.F., D.Y.K.N., R.M.B., A.F.-K., and J.D.; data curation, M.F. and D.Y.K.N.; writing – original draft, M.F., D.Y.K.N., R.M.B., and A.F.-K.; writing – review and editing, M.F., D.Y.K.N., R.M.B., A.F.-K., J.D., and E.M.H.; visualization, M.F., D.Y.K.N., R.M.B., and A.F.-K.; supervision, E.M.H., K.H., and D.K.J.; project administration, A.L. and F.N.; funding acquisition, E.M.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Human Read Removal Modules
  - Synthetic Mixture Titration Dataset Creation
  - Calculation of Statistics
  - Investigation of False Negative Read Sets
  - Evaluation of Consensus Fasta Generation
  - Evaluation of External Metagenomics Dataset
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2025.101218.

Received: April 11, 2025
Revised: August 4, 2025
Accepted: October 6, 2025
Published: November 5, 2025

REFERENCES

1. Marotz, C.A., Sanders, J.G., Zuniga, C., Zaramela, L.S., Knight, R., and Zengler, K. (2018). Improving saliva shotgun metagenomics by chemical host DNA depletion. Microbiome *6*, 42.
2. Bush, S.J., Connor, T.R., Peto, T.E.A., Crook, D.W., and Walker, A.S. (2020). Evaluation of methods for detecting human reads in microbial sequencing datasets. Microb. Genom. *6*, mgen000393. https://doi.org/10.1099/mgen.0.000393.
3. Goig, G.A., Blanco, S., Garcia-Basteiro, A.L., and Comas, I. (2020). Contaminant DNA in bacterial sequencing experiments is a major source of false genetic variability. BMC Biol. *18*, 24.
4. Tomofuji, Y., Sonehara, K., Kishikawa, T., Maeda, Y., Ogawa, K., Kawabata, S., Nii, T., Okuno, T., Oguro-Igashira, E., Kinoshita, M., et al. (2023). Reconstruction of the personal information from human genome reads in gut metagenome sequencing data. Nat. Microbiol. *8*, 1079–1094.
5. Lannelongue, L., Grealey, J., and Inouye, M. (2021). Green Algorithms: Quantifying the carbon footprint of computation. Adv. Sci. *8*, 2100707.
6. Ulhuq, F.R., Barge, M., Falconer, K., Wild, J., Fernandes, G., Gallagher, A., McGinley, S., Sugadol, A., Tariq, M., Maloney, D., et al. (2023). Analysis of the ARTIC V4 and V4.1 SARS-CoV-2 primers and their impact on the detection of Omicron BA.1 and BA.2 lineage-defining mutations. Microb. Genom. *9*, mgen000991.
7. Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., et al. (2011). Modernizing reference genome assemblies. PLoS Biol. *9*, e1001091.
8. Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.-C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thibaud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome

assemblies demonstrates the enduring quality of the reference assembly. Genome Res. *27*, 849–864.

9. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53.
10. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.
11. Ju, Y., Zhang, Z., Liu, M., Lin, S., Sun, Q., Song, Z., Liang, W., Tong, X., Jie, Z., Lu, H., et al. (2024). Integrated large-scale metagenome assembly and multi-kingdom network analyses identify sex differences in the human nasal microbiome. Genome Biol. *25*, 257.
12. Aganezov, S., Yan, S.M., Soto, D.C., Kirsche, M., Zarate, S., Avdeyev, P., Taylor, D.J., Shafin, K., Shumate, A., Xiao, C., et al. (2022). A complete reference genome improves analysis of human genetic variation. Science *376*, eabl3533.
13. Rhie, A., Nurk, S., Cechova, M., Hoyt, S.J., Taylor, D.J., Altemose, N., Hook, P.W., Koren, S., Rautiainen, M., Alexandrov, I.A., et al. (2023). The complete sequence of a human Y chromosome. Nature *621*, 344–354.
14. Shaw, J., and Yu, Y.W. (2024). Rapid species-level metagenome profiling and containment estimation with sylph. Nat. Biotechnol. *43*, 1348–1359.
15. Katz, K.S., Shutov, O., Lapoint, R., Kimelman, M., Brister, J.R., and O'Sullivan, C. (2021). STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. Genome Biol. *22*, 270.
16. Wylie, K.M., Weinstock, G.M., and Storch, G.A. (2012). Emerging view of the human virome. Transl. Res. *160*, 283–290.
17. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915.
18. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.
19. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol. *20*, 257.
20. Di Tommaso, P., Chatzou, M., Floden, E.W., Barja, P.P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nat. Biotechnol. *35*, 316–319.
21. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
22. Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.
23. Grubaugh, N.D., Gangavarapu, K., Quick, J., Matteson, N.L., De Jesus, J. G., Main, B.J., Tan, A.L., Paul, L.M., Brackney, D.E., Grewal, S., et al. (2019). An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol *20*, 8.
24. Constantinides, B., Hunt, M., and Crook, D.W. (2023). Hostile: accurate decontamination of microbial host sequences. Bioinformatics *39*, btad728. https://doi.org/10.1093/bioinformatics/btad728.
25. The Huttenhower Lab. KneadData – The Huttenhower Lab. https://huttenhower.sph.harvard.edu/kneaddata/.
26. Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome *6*, 158.
27. Nicholls, S.M., Quick, J.C., Tang, S., and Loman, N.J. (2019). Ultra-deep, long-read nanopore sequencing of mock microbial community standards. GigaScience *8*, giz043. https://doi.org/10.1093/gigascience/giz043.
28. Brister, J.R., Ako-Adjei, D., Bao, Y., and Blinkova, O. (2015). NCBI viral genomes resource. Nucleic Acids Res. *43*, D571–D577.
29. Jie, Z., Liang, S., Ding, Q., Li, F., Tang, S., Wang, D., Lin, Y., Chen, P., Cai, K., Qiu, X., et al. (2021). A transomic cohort as a reference point for promoting a healthy human gut microbiome. Med. Microecol. *8*, 100039.
30. Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. *50*, D785–D794.
31. R Core Team (2021). R: A Language and Environment for Statistical Computing (Preprint at R Foundation for Statistical Computing).
32. Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Solymos, P., Stevens, M.H.H., Wagner, H., et al. (2025). vegan: Community Ecology Package (R package version 2.7-0). https://vegandevs.github.io/vegan/.
33. Lahti, L., and Shetty, S. (2017). microbiome. Bioconductor. http://bioconductor.org/packages/microbiome/.
34. Arbizu, P.M. (2020). pairwiseAdonis: Pairwise multilevel comparison using adonis (Github). https://github.com/pmartinezarbizu/pairwiseAdonis.
35. Brunson, C. (2020). corybrunson/ggalluvial: faceting bug fix. Zenodo. https://doi.org/10.5281/ZENODO.4012484.
36. Mallick, H., Rahnavard, A., McIver, L.J., Ma, S., Zhang, Y., Nguyen, L.H., Tickle, T.L., Weingart, G., Ren, B., Schwager, E.H., et al. (2021). Multivariable association discovery in population-scale meta-omics studies. PLoS Comput. Biol. *17*, e1009442.
37. Paradis, E., and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics *35*, 526–528.

CellPress
OPEN ACCESS

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Synthetic RNA control – Influenza A (H3N2)	Twist Bioscience	Cat# 103002
Synthetic RNA control – Influenza A (H1N1)	Twist Bioscience	Cat# 103001
Synthetic RNA control – SARS-CoV-2	Twist Bioscience	Cat# 102024
Synthetic RNA control – SARS-CoV-2 + H1N1 mix	Twist Bioscience	Cat# 102024, 103001
Synthetic RNA control – RSV A	Amplirun	Reference: MBC041
Synthetic RNA control – RSV B	Amplirun	Reference: MBC083
Clinical respiratory virus samples	Respiratory Virus and Microbiome Initiative (RVI)	Collected under ethics approval: NRES Committee London – Surrey, REC ref. 22/PR/1222; IRAS ID 316530
Deposited data		
20240718 Human Database Filter	NCBI	https://ftp.ncbi.nlm.nih.gov/sra/dbs/human_filter/human_filter.db.20240718v2
T2T-CHM13v2.0	Nurk et al.[9]	https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/
GRCh38	Schneider et al.[8]	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/
Maxikraken	Loman Lab	https://lomanlab.github.io/mockcommunity/mc_databases.html
1000 Genomes Project human genomic samples	ENA	Accessions in supplemental information
Raw datasets	ENA	Accessions in supplemental information
Software and algorithms		
Bowtie2 v2.5.4	Langmead and Salzberg[18]	https://github.com/BenLangmead/bowtie2
Kraken2 v2.1.3	Wood et al.[19]	https://ccb.jhu.edu/software/kraken2/
BMTagger v3.101	NCBI	https://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/
MetaWrap-QC (as implemented in generate_mags pipeline)	Github – Sanger Pathogens (commit: ff4c635)	https://github.com/sanger-pathogens/generate_mags/tree/ff4c635fbc9801474cde5327bb4fa12239e17b8b
SRA Human Scrubber v2.2.1	NCBI	https://github.com/ncbi/sra-human-scrubber
Nextflow v24.04.4	Di Tommaso et al.[20]	https://www.nextflow.io
Kneaddata v0.1.2	The Huttenhower Lab	https://huttenhower.sph.harvard.edu/kneaddata/
Trimmomatic v0.39	Bolger et al.[21]	http://www.usadellab.org/cms/?page=trimmomatic
BWA v0.7.17	Li and Durbin[22]	http://bio-bwa.sourceforge.net/
iVar v1.4.3	Grubaugh et al.[23]	https://andersen-lab.github.io/ivar/html/
Sylph v0.6.1	Shaw and Yu[14]	https://github.com/bluenote-1577/sylph
Code Repository	This study (commit: f31a834)	https://github.com/wtsi-npg/hrr_tools
RVI viral lens v0.3.2	Github - Genomic Surveillance Unit	https://github.com/genomic-surveillance/rvi-viral-lens/releases/tag/v0.3.2

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

To construct the synthetic titration benchmark dataset, 13 samples were processed using a bait-capture protocol. Six were synthetic control samples obtained from Twist Bioscience, and seven were clinical respiratory samples collected under the Respiratory Virus and Microbiome Initiative (RVI).

Collection and use of the clinical samples were approved by the National Research Ethics Service Committee London – Surrey Research Ethics Committee (REC reference 22/PR/1222; IRAS project ID 316530; approval granted 12 September 2022).

No individual-level demographic information (age or sex) was available to the authors for the RVI samples.

Library preparation and sequencing were performed according to the published protocol RVI-seq – Multi-respiratory virus sequencing protocol V.1 (protocols.io, https://doi.org/10.17504/protocols.io.e6nvw1b62lmk/v1).

METHOD DETAILS

Human Read Removal Modules

To identify the best practice for human read removal from microbial metagenomics data (shotgun and bait capture), we investigated the performance of multiple Human Read Removal (HRR) approaches (Table 1). These methods were selected as a representative set covering popular alignment and classification-based approaches. Bowtie2 methods have previously been identified as highly effective for host read removal; as such we chose methods which facilitated comparison when sensitivity was adjusted and when combined with a classifier. We also wanted to investigate the performance of other methodologies which are widely used in the field, to assess their suitability against the popular alignment based methods. Three methods are based on the popular short-read aligner Bowtie2,[18] Bowtie2 standard is an implementation of the aligner with default parameters implemented within HOSTILE, which wraps Bowtie2 and performs the downstream removal of aligned reads.[24] Bowtie2 VSL implements the "very-sensitive-local" parameter within Bowtie2,[18] increasing the sensitivity of Bowtie2 to more localised matches, mediated through a Kneaddata[25] wrapper with all other functionality switched off and read pairs being mapped in "unpaired" mode, hence treating each independently. Bowtie2 was implemented with a further Kraken2[19] classification step for a 2-stage alignment followed by classification protocol. BMTagger, a k-mer based method, was also implemented within MetaWrap-QC,[26] and the SRA-Human-Scrubber which is based on the SRA Taxonomy Analysis Tool (STAT) which NCBI recommends is run on read-level FASTQ data prior to submission and provides functionality to run the tool upon bioproject submission.[15] For all methods, singleton reads were identified post-processing and removed, hence only paired reads which are not classified as human (either through alignment or classification) are retained for onward analysis.

Additionally, we sought to investigate the difference in performance when using different reference genomes, namely T2T-CHM13v2.0[9] or GRCh38.[7] To this end, the three Bowtie2-based methods (Bowtie2 Standard, Bowtie2 VSL and Bowtie2-Kraken2) were run over samples sets twice, once with a T2T-CHM13 and again with a GRCh38 reference for comparison. For MetaWrap-QC, indexes were created using srprism (version 2.4.24-alpha) for each of the T2T-CHM13 and GRCh38 reference assemblies. For the SRA-scrubber, this approach was not applied, and we tested only the latest human filter database made available through the FTP (see key resources table), date stamped 2024/07/18.

Methods were orchestrated through a nextflow[20] pipeline (see key resources table), and runtime and memory usage statistics were obtained from nextflow trace files.

Synthetic Mixture Titration Dataset Creation

We constructed a dataset to assess the performance of the different human read removal methods under investigation. This dataset contained real viral (RSV A and B, SARS-CoV-2 and Influenza A) and human reads mixed together in varying ratios, ensuring that reads from multiple human individuals with diverse genetic ancestry are present in the created read set mixtures. We selected 13 samples (7 from clinical sources, 6 synthetic control samples) sequenced through a bait capture procedure, and obtained all reads which aligned to one of three respiratory viruses: SARS-CoV-2, RSV A/B or Influenza A (see Data S1). Mapped paired-end reads were extracted, labeled for downstream identification and sorted with samtools v1.19. Human reads were obtained from a collection of 27 samples of different genetic ancestries from the 1000 genomes project,[10] which was previously used in the evaluation of the HOSTILE HRR tool[24] (Supplementary Spreadsheet). This dataset was used as it was created to ensure a publicly available set of human sequences with diverse origins.

Viral and human reads were mixed in such a way as to ensure that the total number of human reads was sampled with identical frequency from each of the 27 human samples, ensuring that samples investigated comprised human reads from populations of different genetic ancestry in equal weight, thus allowing downstream investigation of differential ability to dehumanise samples taken from different human populations.

For each sample investigated, titration sets were created with viral:human proportions ranging from 1:9 - 9:1 via random sampling of viral and human read sets. Random sampling was performed in triplicate for each sample/human titer combination. This resulted in the generation of 351 paired-end fastqs representing the 13 viral samples, with 9 human read titers and 3 iterations of each (Figure 1).

Read sets for the dataset are available through ENA, viral read sets for each sample can be obtained through the ENA: ERP169280 project, titration mixtures can be obtained through the ENA: ERP169281 project and titration mixtures processed through the various HRR methods can be obtained through ENA: ERP169282. A full mapping of all samples can be found in Data S1.

Calculation of Statistics

After titration sets were processed with each human read removal method, read names were extracted from both the removed and retained read sets to obtain the status of all reads in the investigation. Binary classifiers were calculated for each titration sample; we

defined a "positive" case as a read removed by a human read removal method and established the truth status depending on whether the read was incorporated into the dataset from viral or human origin (e.g., a read removed from the dataset by an HRR method which originated from a human sample will be labeled as a true positive).

These binary classifiers were subsequently used to classify performance metrics, this was split into "all-round" metrics such as accuracy, Matthew's Correlation Coefficient (MCC) and F1 scores, and more specific scores such as Precision, Sensitivity (also known as True Positive Rate or Recall) and Specificity (also known as True Negative Rate).

Investigation of False Negative Read Sets

False negative reads were processed using Kraken2 with default settings,[19] using the maxikraken database supplied by the Loman group for the Mock Communities project.[27] Percentages were extracted from Kraken2 reports and in-house code was used to parse reports and to merge counts across samples retaining hierarchies.

Evaluation of Consensus Fasta Generation

The human-read-removed titration datasets 1, 5, and 9 were trimmed and filtered using Trimmomatic (version 0.39), read sets were then taxonomically classified using Kraken2 (version 2.1.3) with a custom database optimised for reference selection consisting of the contents of Viral RefSeq,[28] with additional influenza and RSV genomes obtained from NCBI. To improve the species selection, a bespoke taxonomy was applied within the database which included 1) pruning sub-species leaf nodes for non-influenza/RSV genomes, 2) reorganising RSV genomes into a hierarchical sub-structure for types A and B, and 3) reorganising influenza genomes to effectively bin reads by influenza genome segments. For segments 4 and 6, an additional hierarchical layer was introduced to distinguish between HA and NA types. Viruses with more than 100 assigned reads were selected as references for mapping with BWA (version 0.7.17) and consensus sequences were generated with iVar (version 1.4.3), calling bases with coverage equal to or greater than 10× (implemented through RVI Viral Lens pipeline v0.3.2).

The percentage of sequence coverage against various selected reference sequences was assessed as the proportion of positions in the alignment covered to at least 10×, and was subsequently used to perform a Principal Component Analysis (PCA) followed by a t-distributed Stochastic Neighbor Embedding (t-SNE, perplexity 30, number of iteration 5000) for visualisation. A PERMANOVA (Permutational Multivariate Analysis of Variance) was performed to assess the statistical significance of differences between groups within the dataset.

Evaluation of External Metagenomics Dataset

We selected a publicly available shotgun metagenomics study of the nasal microbiota to benchmark our method, accession number CRA006819 deposited on Genome Sequence Archive (GSA). Samples for nasal microbiome data were collected as part of the 4D-SZ cohort, and represent 1593 individuals of East-Asian genetic ancestry, from the city of Shenzhen.[29] To date, this is the largest publicly available shotgun metagenomic dataset for the nasal microbiota. Before the data was uploaded, the authors removed human reads with Bowtie2 using the human genome GRCh38 as the reference[11]. Raw FASTQ files were downloaded from the GSA. The human read removal pipelines, as described in Table 1, were applied to the dataset. The microbiota taxonomic composition was determined using sylph 0.6.1[13] with GTDB release 220[30] sketched at c200.

Statistical analysis was performed in R version 4.4.2.[31] Alpha diversity was calculated with the vegan package version 2.6–10[32] with the functions specnumber and diversity. Pielou's evenness was calculated using the function diversity found within the microbiome package version 1.28.0.[33] Kruskal-Wallis test and pairwise Wilcox test were used to test for statistical significance. For beta diversity, Bray-Curtis dissimilarity and Jaccard Index were calculated using vegdist function found within vegan. A pairwise t-test was used for statistical testing. *p*-values were adjusted with the Benjamini-Hochberg procedure. Pairwise PERMANOVA was calculated with function pairwise.adonis2 from the package pairwiseAdonis.[34] Visualisation of the average nucleotide identity was done with ggalluvial.[35] MaAsLin 2 v1.20.0 was used to test for microbial association.[36] The fixed effect measured is the method used for human read removal while the random effect was the sample. The raw dataset (Bowtie2 Standard with GRCh38) was set as the reference. The default parameters were used. To visualise the difference between the methods, a PCoA was calculated using the pcoa package from ape v5.8-1[37] and the dbrda functions found within the vegan package. The function varpart was used to calculate the partition of variation in the data.

To assess the ability to determine human sample sex from uploaded metagenomic data, a subset of 10 samples (M = 5, F = 5) were used. We compared pre-human read removal samples (uploaded to GSA) with post-human read removal samples (GSA samples processed through SRA human scrubber, Bowtie2 VSL with a GRCh38 reference genome, and Bowtie2 VSL with a T2T-CHM13v2.0 reference genome). Post-human read removal samples were re-aligned to the T2T-CHM13v2.0 reference using Bowtie2 version 2.5.1 and the '–very-sensitive-local' option to identify retained human reads. Information about the proportion of reads mapping to the reference genome was gathered using samtools1.17. Mean read depth values were visualised using the 'pseudo_log_trans()' function from the 'scales' R package version 1.3.0.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical procedures are described in detail in the above sections entitled Human Read Removal Modules, Synthetic Mixture Titration Dataset Creation, Calculation of Statistics, Investigation of False Negative Read Sets, Evaluation of Consensus Fasta Generation and Evaluation of External Metagenomics Dataset.