



OPEN

## CleanSeqU algorithm for decontamination of catheterized urine 16S rRNA sequencing data

Sung Min Yoon<sup>1,3</sup>✉, Chang-Seok Ki<sup>1</sup> & Ju Sun Song<sup>1,2,3</sup>✉

Contamination in low-biomass samples, such as urine, presents a major challenge for 16S rRNA gene sequencing, as extraneous DNA from reagents and the environment often obscures microbial signals. Existing *in silico* decontamination algorithms face limitations in accurately identifying and removing these contaminants. To address this issue, we developed CleanSeqU, a novel decontamination algorithm designed to enhance the accuracy of 16S rRNA gene sequencing data for catheterized urine samples. This approach is grounded in the principle that the compositional pattern of potential contaminant taxa remains similar between biological samples and blank controls. Also, the algorithm identifies potential contaminants based on ecological plausibility and custom blacklist. We evaluated CleanSeqU's performance using vaginal microbiome dilution experiments as a proxy for low-biomass urine samples and compared it to the Decontam, Microdecon, and SCRUB algorithm. CleanSeqU consistently outperformed Decontam, Microdecon, and SCRUB across various contamination levels, with superior accuracy, F1-scores, and reduced beta-dissimilarity. CleanSeqU improved specificity and positive predictive value by correctly identifying and removing a higher number of contaminant amplicon sequence variants (ASVs). Furthermore, the reduced alpha diversity in the decontaminated datasets suggests more precise contaminant elimination. With its practical use of a single blank extraction control per batch and adjustable decontamination rules, CleanSeqU provides an efficient and scalable solution that delivers accurate microbial profiles. Our findings highlight its potential to significantly advance urine microbiome research by delivering more accurate microbial profiles.

**Keywords** 16S rRNA gene sequencing, Low biomass samples, Blank extraction control, Microbial contamination, Decontamination algorithms, Urine microbiome

### Abbreviations

PPV	Positive predictive value
ASV	Amplicon sequence variant
PCR	Polymerase chain reaction
MAD	Median absolute deviation
AAD	Average absolute deviation
TP	True positive
TN	True negative
FP	False positive
FN	False negative

Next-generation sequencing has advanced microbiome research because of the ability to perform more sensitive surveys of microbial communities, genomes, and functions. Microbiome studies initially focused on the gut microbiota, followed by other high-biomass organs, such as the vagina, skin, and mouth, which were the major body sites in the Human Microbiome Project launched in 2007<sup>1</sup> and have been further extended to low-biomass samples, such as urine<sup>2</sup>, placenta<sup>3</sup>, and the lower airway<sup>4</sup>. Among them, the urinary tract has a unique microbiota, even in the absence of urinary tract infection<sup>5</sup> and its microbial burden can be  $10^{3-5}$  bacteria per 1 ml of urine. This is at least  $10^6$  times smaller than the  $10^{11}$  bacteria per 1 ml of gut content<sup>6</sup>.

<sup>1</sup>Department of Laboratory Medicine, GC Genome, Seoul 16924, Korea. <sup>2</sup>Department of Laboratory Medicine, Green Cross Laboratories, Seoul 16924, Korea. <sup>3</sup>Sung Min Yoon and Ju Sun Song contributed equally to this work. <sup>4</sup>This work is part of Ju Sun Song's Ph.D. thesis at Sungkyunkwan University, Suwon, Korea. <sup>✉</sup>email: ysm0503v@gccorp.com; sjusun277@gmail.com

Urological disorders were previously thought to have no microbiological etiology; however, the discovery that the urinary tract is not a sterile environment and has a diverse and distinct urobiome has changed our understanding of these conditions. The function of the urobiome in a variety of urological diseases is gaining attention, and its alterations have been reported in a variety of urological diseases, such as chronic recurrent cystitis, neurogenic bladder dysfunction, interstitial cystitis, urgency urinary incontinence, urolithiasis, overactive bladder, and bladder cancer<sup>7-18</sup>.

As with other microbiome studies, the most commonly used method for urobiome research is marker gene (amplicon) sequencing because of its low cost and speed. To investigate bacterial communities, a partial hypervariable region of the 16S rRNA gene is specifically targeted. This process consists of extracting bacterial DNA from a sample, amplifying it using polymerase chain reaction (PCR), and sequencing. Amplicon sequencing is an extremely sensitive method, even for low-biomass specimens and has increased our ability to detect microbes in such samples. However, accurate characterization of microbial communities using marker gene sequencing is challenging in low-biomass specimens containing very little endogenous DNA. Contamination with bacterial DNA from exogenous sources introduced during sample collection and processing. This contamination can skew results<sup>19-22</sup>.

Extracellular microbial DNA can last for thousands of years and is found in nearly all ecosystems, including soils, sediments, freshwater, and oceans<sup>23</sup>. In addition, its effects are widespread in laboratory environments. Contaminant bacterial DNA can be isolated from many sources, including plastic consumables<sup>24</sup>, molecular biology grade water<sup>25,26</sup>, nucleic acid extraction kits<sup>19,22,27</sup>, and PCR master mixes<sup>20,26,28,29</sup>. Contaminated laboratory reagents in 16S rRNA gene-based experiments have long been recognized in the scientific literature<sup>30</sup> and these contaminating sequences have been previously reported to match water- and soil-associated bacterial genera such as *Acinetobacter*, *Alcaligenes*, *Bacillus*, *Bradyrhizobium*, *Herbaspirillum*, *Legionella*, *Leifsonia*, *Mesorhizobium*, *Methylbacterium*, *Microbacterium*, *Novosphingobium*, *Pseudomonas*, *Ralstonia*, *Sphingomonas*, *Stenotrophomonas*, and *Xanthomonas*<sup>22</sup>.

Several methods are available for detecting and eliminating contamination from microbial sequencing data, including (i) removal of sequences that appear in controls, (ii) removal of sequences below an ad hoc relative abundance threshold, (iii) removal of sequences previously identified as contaminants, and (iv) bioinformatics methods<sup>42</sup>. The most popular method for controlling and mitigating the impact of contaminant bacterial DNA in low-biomass samples is to sequence blank extraction controls along with the samples, relying on the assumption that sequencing of appropriate blank extraction controls will reveal background contaminants that could possibly occur in the associated clinical samples. The majority of contaminants are present in low abundance and are randomly included during pipetting of the PCR template<sup>31</sup>. They are subject to the rule of small numbers, which states that a random sample is unlikely to accurately represent the population from which it is obtained<sup>32</sup>. Therefore, these contaminants will not occur in the blank extraction control, particularly when the number of controls is limited.

Catheterized urine samples are also susceptible to contamination, but many urobiome studies have been published without appropriate decontamination procedures. Consequently, it is difficult to reach a consensus on the connection between urological illnesses and the urobiome. In this study, we developed a novel decontamination algorithm, CleanSeqU, which integrates multiple decontamination rules to overcome the limitations of existing methods. CleanSeqU classifies taxa identified in samples into three groups based on contamination levels measured in blank extraction control and applies tailored filtering rules to each. The algorithm utilizes Euclidean distance similarity analysis to identify and retain genuine taxa among highly abundant contaminants, Z-score-based filtering to distinguish true signals from low-level contamination, and ecological plausibility assessment using external databases to eliminate non-biological contaminants. Additionally, an in-house blacklist, curated from large-scale laboratory experimental data, is employed to remove recurrent contaminants specific to laboratory environments. We rigorously validated the algorithm using datasets generated by a multiple dilution series of human vaginal microbial samples and demonstrated that CleanSeqU outperformed an algorithm reported to remove contamination, with the highest accuracy among decontamination tools recently reported to date.

## Materials and methods

### CleanSeqU: decontamination model description

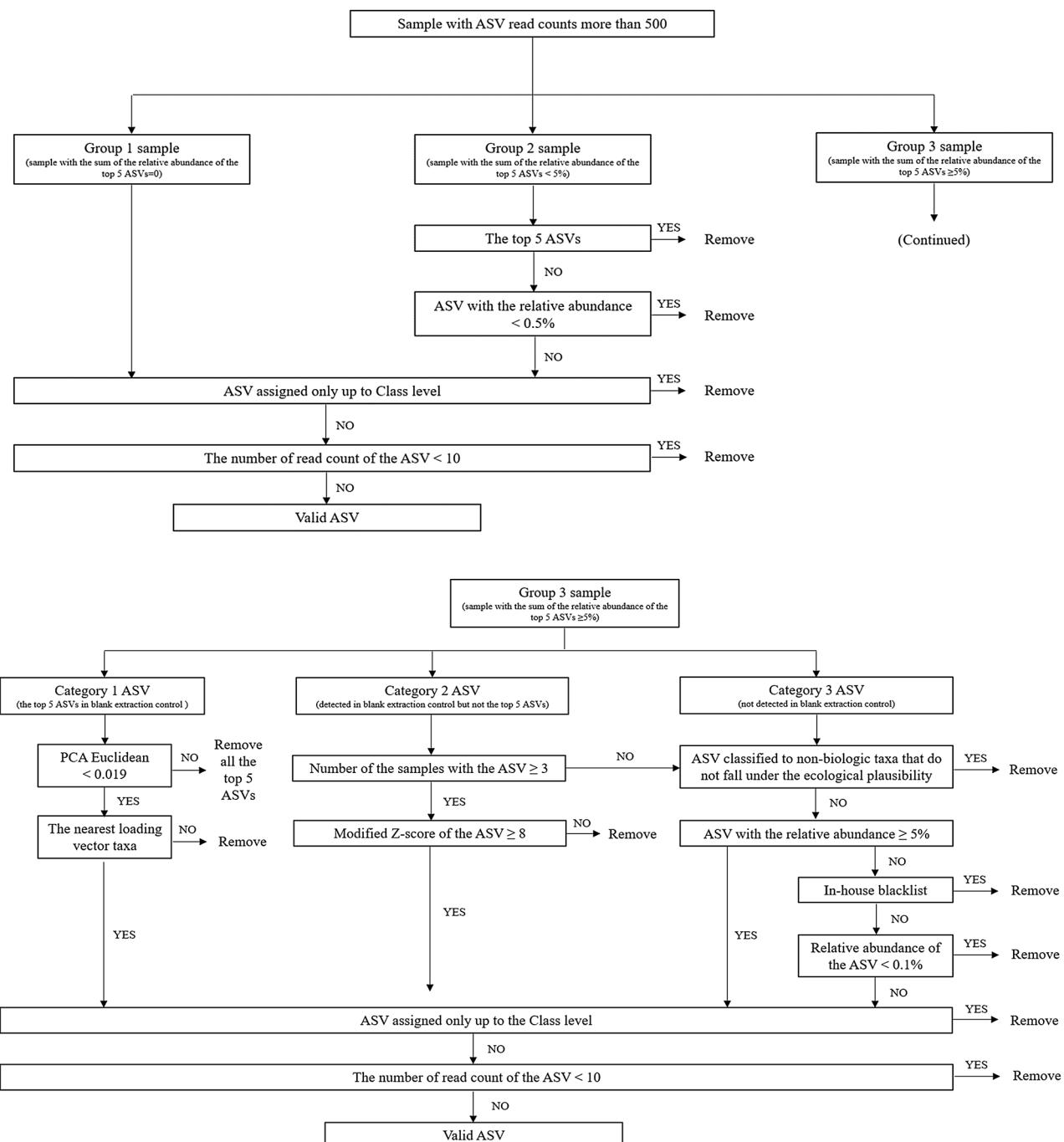
In this study, we developed a novel decontamination algorithm called CleanSeqU, in which various decontamination rules were adapted and integrated to complement their inherent limitations. CleanSeqU is composed of less than ten rules and each rule discriminates between contaminated ASV sequences from true ASV sequences. CleanSeqU is applied to each experimental batch and operated based on the 16S rRNA gene sequencing results of one blank extraction control processed together in each batch. Also, CleanSeqU only accounts for samples with ASV read counts of more than 500 because inadequately sequenced samples may not effectively reflect the overall bacterial community truly present in the sample.

The samples in the processed batch were first classified into three groups according to the level of contamination, which was determined based on the sum of the relative abundances of the five ASVs identified at the highest abundance in the blank extraction control of each batch in the sample. The five ASVs identified with the highest abundance in the blank extraction control are henceforth referred to as the top 5 ASVs in this paper. Samples in Group 1 are uncontaminated and were defined as samples in which the sum of the relative abundances of the top 5 ASVs was 0. The samples in Group 2 have a low level of contamination, as indicated by the sum of the relative abundances of the top 5 ASVs in the sample of less than 5%. The sum of the relative abundances of the top 5 ASVs in Group 3 samples is 5% or above, suggesting a moderate to high level of contamination. In the CleanSeqU algorithm, different decontamination rules are applied to the three groups, depending on the degree of contamination. CleanSeqU assesses sample contamination by quantifying dominant blank control

ASVs, classifies samples into three contamination groups and applies distinct decontamination procedures to each group (Fig. 1).

For Group 1 samples, all ASVs detected in the 16S rRNA gene sequencing results were considered valid sequences, and none of them were removed. Because the Group 2 samples have a low level of contamination, and contaminants rather than the top 5 ASVs in the sample were thought to be at a very low abundance, we removed the top 5 ASVs as well as the ASVs with a relative abundance of less than 0.5%.

ASVs found in Group 3 samples were further classified into three categories depending on how abundant they were or whether they existed in the blank extraction control of the experimental batch, as follows: the top 5 ASVs were classified as category (1) ASVs that were not among the top 5 ASVs but were detected in the blank extraction control of the experimental batch were classified as category (2) ASV that were not present in the blank extraction control of the experimental batch were classified into category (3) We applied different decontamination rules according to the ASV category.



**Fig. 1.** Flowchart of the CleanSeqU decontamination processes. Top 5 ASVs refers to the five ASVs identified at the highest abundance in the blank extraction control.

### Category 1 ASV—the top 5 ASVs

Abundant contamination, such as category 1 ASVs, was robustly detected across all moderately to highly contaminated samples as well as the blank extraction control in a sequencing run. The relative proportions of these abundant contaminants were similar in all contaminated samples as well as in the blank extraction control because multiple taxa present in the contamination source were introduced together in the samples. However, it is possible that this feature is both a contaminant and genuine in the studied ecosystem. In this case, the genuine feature was present in the sample of interest at a much higher prevalence compared to the other abundant contaminants, breaking the similar proportions observed across the blank extraction control and contaminated samples. To distinguish between contaminants and genuine features, we measured the Euclidean distance similarity between the compositional data of each sample and a blank extraction control using biplot analysis, in which the relative abundances of the top 5 ASVs of the samples and blank extraction control were normalized to 100. The larger the Euclidean distance similarity, the more similar the proportion of the top 5 ASVs in the sample is to that of the blank extraction control, indicating that the top 5 ASV in the sample are contaminants. The smaller the Euclidean distance similarity, the greater the proportion of the top 5 ASVs in the sample that deviated from that of the blank extraction control; some of these may be genuine features. The cutoff of the Euclidean distance similarity was set at 0.019 pragmatically, based on the observation from our 16S rRNA gene sequencing data that the composition of the top 5 ASVs in the sample with Euclidean distance similarity below this cutoff seemed to be biased considerably from the composition shown in the blank extraction control. Once the samples with Euclidean distance similarity below this cutoff were identified, the feature with the highest loading vector among the top 5 ASVs was considered a genuine feature, and the remaining features were removed.

### Category 2 ASV—ASV detected in the blank extraction control but not top 5 ASVs

Category 2 ASVs indicated a relatively low abundance of contaminant microbial DNA. The majority of contaminants might fall into this category. These low-abundance contaminant ASVs will be detected at a low prevalence in the sample, similar to the blank extraction control. However, some genuine features in a sample may be classified as category 2 because of the well-to-well leakage phenomenon, in which an abundantly present genuine feature in a sample is cross-contaminated into a blank extraction control. Well-to-well leakage commonly occurs within batches during experimental procedures. While contaminant ASVs belonging to category 2 were detected at low levels in most samples, as well as in the blank extraction control, these genuine features might be detected at a much higher ratio. To distinguish between the contaminant and truly present taxa among the ASVs in this category, we used the Z-score method, which deals with shared information across samples within the batch. Statistically, the Z-score quantifies the distance (in standard deviations) of a data point from the mean of the dataset. A high absolute z-score indicates that a data point is far from the mean, suggesting that it may be an outlier. Because ASVs that exist as contaminants in an experimental batch will be identified at a proportion similar to that of the blank extraction control in samples where the ASV is found, the Z-score, which is calculated using the proportion of ASVs, will show a low value close to zero for these ASVs. Meanwhile, ASV, which is truly present as a genuine feature in a sample, can be expected to be identified at much higher levels than the other samples as well as the blank extraction control. In this case, the Z-score of the ASVs presented as genuine taxa will show a higher value than the other samples. Therefore, we evaluated the Z-score for each ASV belonging to Category 2 to differentiate the truly present features.

Because the relative abundance of low-abundance contaminants in biological samples may vary according to their contamination levels, the Z-score calculated using the relative abundance of ASVs may have limited accuracy in distinguishing truly present features from contaminants. Therefore, in our framework, the adjusted Z-score, using the value of the relative abundance of the ASVs divided by the sum of the top 5 ASVs, was used rather than a simple Z-score using the relative abundance of the ASV itself.

Additionally, we utilized a modified Z-score that employs the median rather than the mean to calculate the Z-score because this is a more robust way to detect outliers. The cutoff of the adjusted modified Z-score was pragmatically set to 8, and an ASV with an adjusted modified Z-score of 8 or more was considered an actual feature in the sample. Only ASVs found in three or more biological samples were subjected to the adjusted modified Z-score analysis, whereas ASVs found in two or fewer biological samples were subjected to the decontamination rule for category 3 ASV, which are described in the next section.

### Category 3 ASV—ASV not detected in the blank extraction control

Most ASVs belonging to category 3 are likely to be contaminants present in low abundance; consequently, many contaminants may only exist in samples without being detected in the blank extraction control. Therefore, it is necessary to distinguish ASVs that can be represented as contaminants in samples, even if they are not detected in a blank extraction control. To determine whether category 3 ASVs are contaminants, we applied ecological plausibility and created an in-house database in our framework.

To determine ecological plausibility, we used the BacDive database, which is the largest worldwide database for standardized bacterial information and include isolation sources of the strains<sup>33</sup>. If none of the strains belonging to a genus had been isolated from human-related sources, we classified the genus as a non-human source. As result, among the 3239 genera for which isolation sources are registered in the Bacdive database, 2257 genera were classified as non-human sources. In addition, *Proteobacteria*, particularly *Alpha-* and *Beta-proteobacteria*, include bacteria that dominate aquatic and terrestrial systems<sup>34,35</sup>. Overall, taxa belonging to *Alpha-Proteobacteria* or *Beta-Proteobacteria* as well as taxa belonging to the genus classified as non-human source from the Bacdive database were defined as non-biological contaminants that do not fall under ecological plausibility, and if an ASV classified as category 3 in the sample falls under this list, it was removed as a contaminant.

Specific contaminants originating from the laboratory environment, consumables, and reagents may also be present. We created an in-house blacklist that referred to features that were specifically and recurrently detected in our 16S rRNA gene sequencing data produced from 2912 clinical urine samples and 148 blank extraction controls. To create an in-house blacklist, we assumed two concepts: (1) contaminants should be present in the blank extraction controls at a higher relative abundance compared to the biological samples, and (2) category 3 contaminants might not be discovered in a high ratio. Then, we classified the features into an in-house blacklist that met the following criteria: (1) maximum relative abundance of the feature across all biologic urine samples was less than 1%, (2) maximum relative abundance of the feature across all biologic urine samples was less than 5% when the mean relative abundance of all blank extraction controls was higher than the mean relative abundance of the feature of all biologic urine samples, and (3) maximum relative abundance of the feature across all biologic urine samples was less than 5% when the genus assigned to the feature was listed on the contaminants list in the GRIMER repository more than three times. GRIMER<sup>36</sup> is a tool for analyzing, visualizing, and exploring microbiome studies with a focus on contamination detection and compiles an extensive list of common contaminants containing 210 genera and 627 species reported in 22 published articles. There were 85 genera listed more than three in the contaminant list in the GRIMER repository (Supplementary Table S1). By applying these criteria for in-house blacklist, 54,721 out of 56,010 ASVs identified from 3060 16S rRNA gene sequencing data were classified as blacklists. Of the 54,721 blacklisted ASVs, 491 were detected in more than 100 of the 2912 urine samples (Supplementary Table S2).

The following is the order in which contaminants are removed from ASVs that are classified as category 3. First, ASVs corresponding to non-biological contaminants were removed, and then features corresponding to the in-house blacklist with a relative abundance of less than 5% were removed. Additionally, rare features with a relative abundance of less than 0.1% were removed.

Among the ASV features remaining valid after decontamination processes in all categories, ASV features whose sequences were only assigned up to Class level were additionally removed because it is reasonable for genuine taxa to be appropriately assigned to low-level taxonomies using well-curated and complete reference taxonomy databases and well-performing taxonomy assignment algorithms. Finally, ASVs with read counts below 10 were eliminated to exclude the possibility of low-frequency artifacts (e.g., sequencing artifacts or low-lying PCR contamination).

### Evaluation of decontamination methods

#### *Human vaginal microbial Dilution series data*

We chose to evaluate our algorithm using a vaginal microbiome dataset instead of catheterized urine samples, because catheterized urine samples from UTI or asymptomatic bacteriuria cases have a high microbial burden, but they typically exhibit limited diversity with a few dominant species. This makes them less suitable for testing the algorithm on a diverse microbiome composition. Vaginal and urinary microbiomes are similar<sup>37</sup>, therefore, we are confident that the algorithm's performance can be generalized to urine microbiome data as well. We prepared a human vaginal microbial dilution series using ten leftover human vaginal microbiome samples. Vaginal microbiome samples were collected using a sterile swab kit containing preservatives (Noble Biosciences, Republic of Korea). Preservative solutions of each vaginal sample were first diluted to 1/1000 and had further undergone six rounds of serial two-fold dilutions with nuclease-free water (NFW) (Invitrogen, USA). Nucleic acid concentrations of the undiluted vaginal samples ranged from 4 to 40 ng/μl. Experiments of 16S rRNA gene sequencing for a total of 10 sets of the vaginal sample dilution series were conducted in the same manner with other catheterized urine samples requested to the laboratory and processed divided into 6 experimental batches along with urine samples. A blank extraction control was included in each batch.

This study was approved by the Ethics Committee of GC Laboratories (GCL-2023-1075-02) and was carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects at the time of initial sample collection for diagnostic purposes and this study was conducted using residual human microbiome samples collected after diagnostic testing, with all samples fully anonymized prior to analysis. As this study utilized de-identified residual samples, the requirement for additional informed consent specific to this research was waived by the IRB, in accordance with institutional and national guidelines.

DNA extraction was performed using the MagMAX™ Microbiome Ultra Nucleic Acid Isolation Kit (ThermoFisher Scientific, Waltham, MA, USA) according to the manufacturer's instructions. The prepared DNA was used for 16S library construction using NEXTflex 16S V4 Amplicon-Seq (Bioo Scientific, Austin, TX, USA). The amplification cycle was 8 cycles for PCR I amplification and 22 cycles for PCR II amplification. The final library products were diluted, pooled, and sequenced using the MiSeq system (Illumina) with a paired-end 500-cycle kit. The vaginal microbial dilution series and blank extraction controls were subjected to the same procedure.

#### *Bioinformatic analysis*

QIIME 2 (2021.11.0) was used to analyze the 16S rRNA gene sequence data<sup>38</sup>. Demultiplexed and primer-trimmed data were quality-filtered and denoised using DADA2 (Divisive Amplicon Denoising Algorithm 2, 1.18.0), which uses a parametric model to infer exact biological sequences from quality-filtered reads, known as ASVs<sup>39,40</sup>. In DADA2, independently denoised forward and reverse reads were merged at the end of the workflow, and the chimeric ASVs were removed. For the taxonomic classification of ASVs, a multinomial naive Bayes machine-learning classifier in the q2-feature-classifier was used against the refseq database<sup>41</sup>. Finally, ASVs that were not assigned to bacteria at the domain level were removed.

### Benchmarking of decontamination methods

To evaluate the performance of CleanSeqU, we compared the outcomes of the decontamination process using CleanSeqU with the previously published decontamination method decontam (1.13.0)<sup>42</sup>, microdecon (1.0.2)<sup>43</sup>, and SCRuB (0.0.1)<sup>44</sup> using 16S rRNA gene sequencing data produced from the vaginal microbial dilution series, as described above.

To assess the overall performance of the decontamination methods, we calculated the alpha-diversity metrics and Bray–Curtis dissimilarity between the decontaminated data and 16S rRNA gene sequencing data from the undiluted vaginal samples. To evaluate the overall accuracy of each decontamination method, we categorized ASVs as either correctly or incorrectly identified as ground truth or contaminants, following an approach similar to that used by Karstens et al. (2019)<sup>45</sup>. We defined the ground truth for ASV classification based on 16S rRNA gene sequencing data generated from undiluted vaginal samples. Specifically, an ASV was considered a contaminant if it was not expected to be present in undiluted vaginal samples, whereas a ground-truth ASV was one that occurred in these samples. ASVs correctly classified as ground truth were referred to as true positives, while ASVs correctly classified as contaminants were considered true negatives. Conversely, ASVs incorrectly classified as ground truth were false positives, and those incorrectly classified as contaminants were false negatives.

Additionally, we compared CleanSeqU with Decontam, Microdecon, and SCRuB using the publicly available mock community dilution series from Karstens et al. (2019)<sup>45</sup>.

### Statistical analysis

All statistical analyses were performed using R version 4.0.5. Category 1 ASVs were decontaminated using the prcomp and biplot functions, and Euclidean distance similarity was calculated using the proxy package. When calculating the adjusted modified Z-score, if the median absolute deviation (MAD) value was nonzero, it was multiplied by a weighting factor of 1.4826 and used as the denominator. However, if the MAD value was zero, the average absolute deviation (AAD) was multiplied by 1.2533 and used as the denominator. A statistical hypothesis test comparing the two groups was performed using the Wilcoxon signed-rank sum test, which is a nonparametric test. The smooth curve of the numerical changes according to the proportion of total contaminants in the figures was analyzed using LOESS.

## Results

### 16S rRNA gene sequencing of vaginal microbial Dilution series and blank extraction control

The total microbial community of the ten undiluted vaginal microbiome samples consisted of 107 ASV features. Among them, 49 ASV features were found at a relative abundance of 1% or above in at least one undiluted vaginal sample, and they mapped to 34 distinct taxa (Supplementary Table S3).

Among the 6 blank extraction controls, 570 ASV features were detected. The most abundant genus was *Pseudomonas*, followed by *Janthinobacterium*, *Stenotrophomonas*, *Cutibacterium*, and *Undibacterium*. The assigned taxa of the 13 ASV features that were found in the blank extraction controls and had an average prevalence > 1% (Supplementary Table S4). The average proportions of phyla for ASV features detected in the blank extraction controls were 70, 14, 9, and 5% for *Proteobacteria*, *Actinobacteria*, *Bacteroidetes*, and *Firmicutes*, respectively.

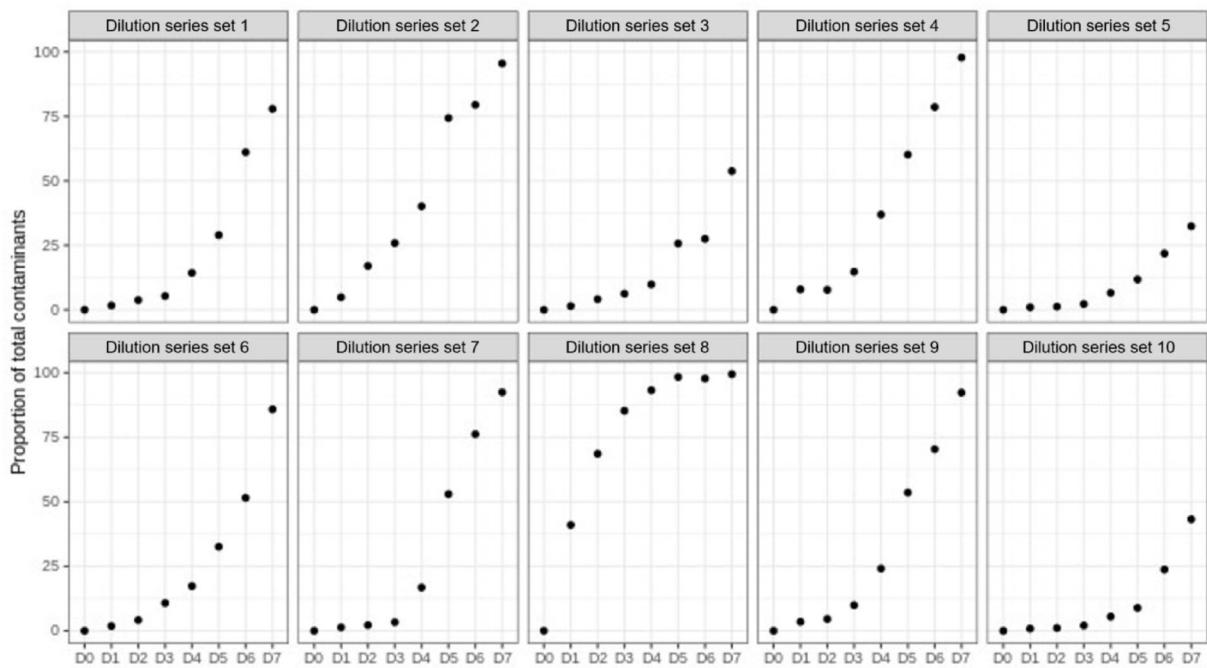
In the dilution series of vaginal microbial samples, samples with more diluted material were characterized by higher proportions of contaminants, as defined by sequences that did not match the expected undiluted vaginal microbial characteristics, although not as linearly as expected (Figs. 2 and 3; Supplementary Table S5). The proportion of total contaminants ranged from 1 to 99%.

Comparison of decontamination performance between Decontam, Microdecon, SCRuB and CleanSeqU using vaginal microbial dilution series.

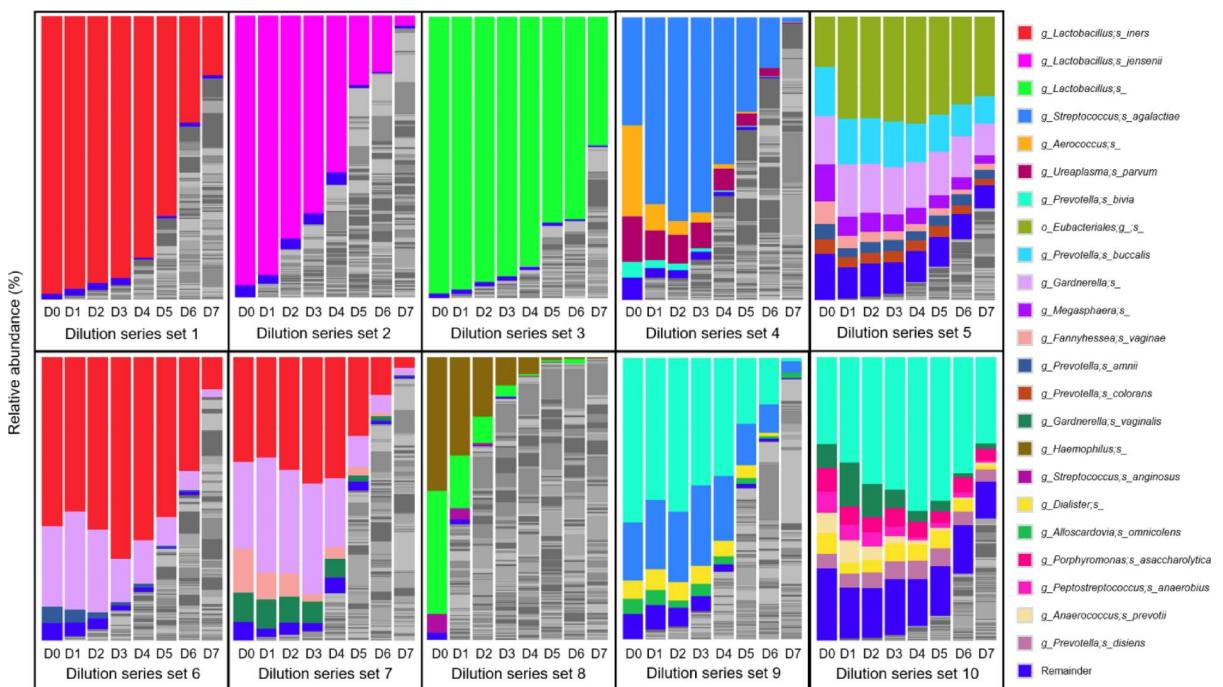
We tested the performance of Decontam, Microdecon, SCRuB and CleanSeqU algorithms in identifying and removing contaminant ASVs from a vaginal microbial dilution series. Compared with decontam, microdecon, and SCRuB, the relative abundance of the removed ASVs as contaminants was higher in the CleanSeqU at all dilution stages, and this difference tended to become more pronounced as the contamination proportion increased in the more diluted stages (Fig. 4). Alpha diversity calculated by Chao1 estimating species richness showed that more types of ASVs were removed by CleanSeqU than by decontam, microdecon, and SCRuB across all dilution stages, indicating CleanSeqU usually recognize more types of ASV as contaminants than the other algorithms (Fig. 5).

To evaluate the ability of each decontamination method to recover the expected vaginal microbial community profiles from the contaminated dilution series samples, we compared the accuracy, F1-score, and output similarity to the ground truth using the Bray–Curtis dissimilarity between the algorithms in each dilution series set. ASVs classified as correctly or incorrectly identified as undiluted vaginal microbial communities or contaminants for the 10 sets of the vaginal sample dilution series are presented in Fig. 6 and Supplementary Table S6. Among them, decontam exhibited the highest false positive rate, followed by microdecon and SCRuB, while CleanSeqU had the lowest. Decontam and Microdecon showed slightly higher false negative rates compared to the SCRuB and CleanSeqU, although the false negative rates were generally low across all four algorithms. CleanSeqU showed higher accuracy, and F1-score, and lower Bray–Curtis dissimilarity compared to the other algorithms in most diluted samples (Fig. 7).

Both the accuracy and F1-score gradually decreased as the contaminant proportion increased, and the beta dissimilarity gradually increased as the contaminant proportion increased (Fig. 8). In particular, the values of the F1-score and beta-dissimilarity tended to change sharply in the highly contaminated samples. Because highly contaminated samples produce imbalanced data, it can be said that the F1-score, interpreted as the harmonic



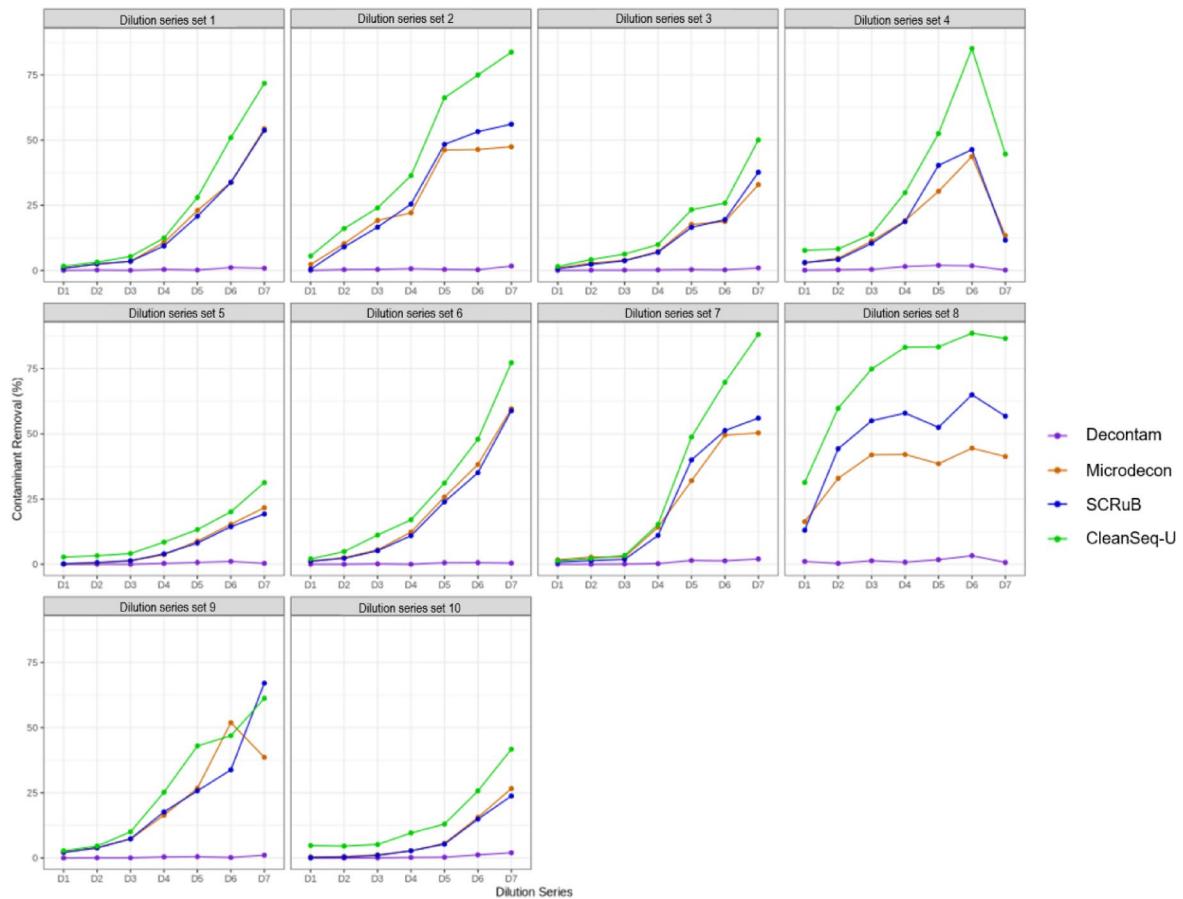
**Fig. 2.** The proportion of total contaminants increases with decreasing amount of bacterial input material.



**Fig. 3.** Stacked bar plot representing the bacteria identified in dilution series of each set. The expected bacteria from the undiluted vaginal microbial community are displayed in color, while contaminant bacteria are in grayscale. Bacteria that existed at a prevalence of less than 5% in the undiluted vaginal microbial community were designated as the other genuine bacteria.

mean of precision and recall, and beta-dissimilarity, quantifying differences in overall taxonomic composition, reflect more accurate performance rather than accuracy.

Furthermore, we divided all diluted samples into two groups based on a 90% cut-off for the contaminant proportion and compared the differences in F1-score and beta-dissimilarity between the algorithms to evaluate whether their performance varied depending on the contaminant proportion. The CleanSeqU showed a



**Fig. 4.** The proportion of removed ASV reads in the dilution series samples in each batch determined using Decontam, Microdecon, SCRuB and CleanSeq-U.

significantly better F1-score and beta-dissimilarity than the other algorithms in the group with a contaminant proportion of less than 90%; in contrast, there was no significant difference in those parameters in the group with a contaminant proportion of more than 90% (Fig. 9).

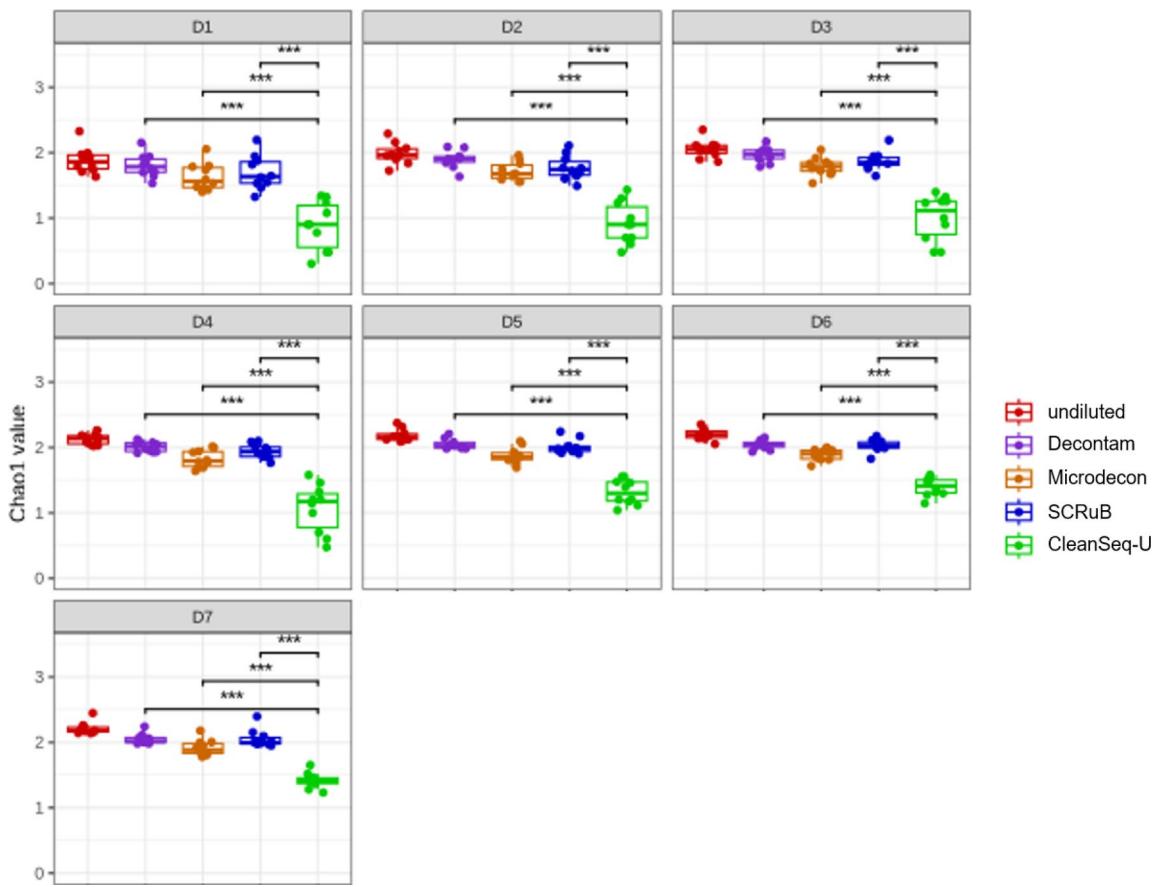
Comparison of decontamination performance between Decontam, Microdecon, SCRuB and CleanSeqU using publicly available mock community dilution series.

To further validate the performance of CleanSeqU in different low-biomass environments, we additionally applied our algorithm to the mock community dataset from Karstens et al. alongside Decontam, Microdecon, and SCRuB for comparison (Supplementary Fig. S4). The results demonstrated that CleanSeqU effectively reduces contaminant sequences while maintaining the overall microbial community structure, showing improved performance relative to the other decontamination methods in dilution samples with low contamination rates. However, in dilution samples with a higher contamination rate, the performance of CleanSeqU was comparable to or lower than that of Decontam and Microdecon. The SCRuB exhibited a markedly reduced performance at all dilution levels compared to the other algorithms.

The stacked bar plot of ASV classification demonstrated that in samples with a low contamination rate, CleanSeqU had the lowest combined proportion of false-positive and false-negative ASV features. However, as the contamination rate increased, the proportion of incorrectly classified ASVs gradually increased across all algorithms (Supplementary Fig. S5). Microdecon, which showed the best performance in high-contamination samples, had the lowest proportion of false-positive ASV features among the four algorithms. The next best-performing algorithm, Decontam, exhibited the lowest proportion of false-negative ASV features. In contrast, SCRuB, which had the poorest performance among the four algorithms, consistently showed a high proportion of false-negative ASV features across all dilution levels and a higher proportion of false-positive ASV features than the other algorithms, especially at high dilution levels.

## Discussion

Several software tools have been developed to identify and control bacterial DNA contamination in 16S rRNA gene sequencing data. Decontam<sup>42</sup> operates in a set of rules in which contaminant taxa are recognized and removed, which are more prevalent in controls than in the samples of interest and/or are more frequent in samples with lower DNA concentrations. However, there is a limit to identifying and removing a taxon if it is both a contaminant in certain samples and genuinely present in others, as Decontam cannot distinguish



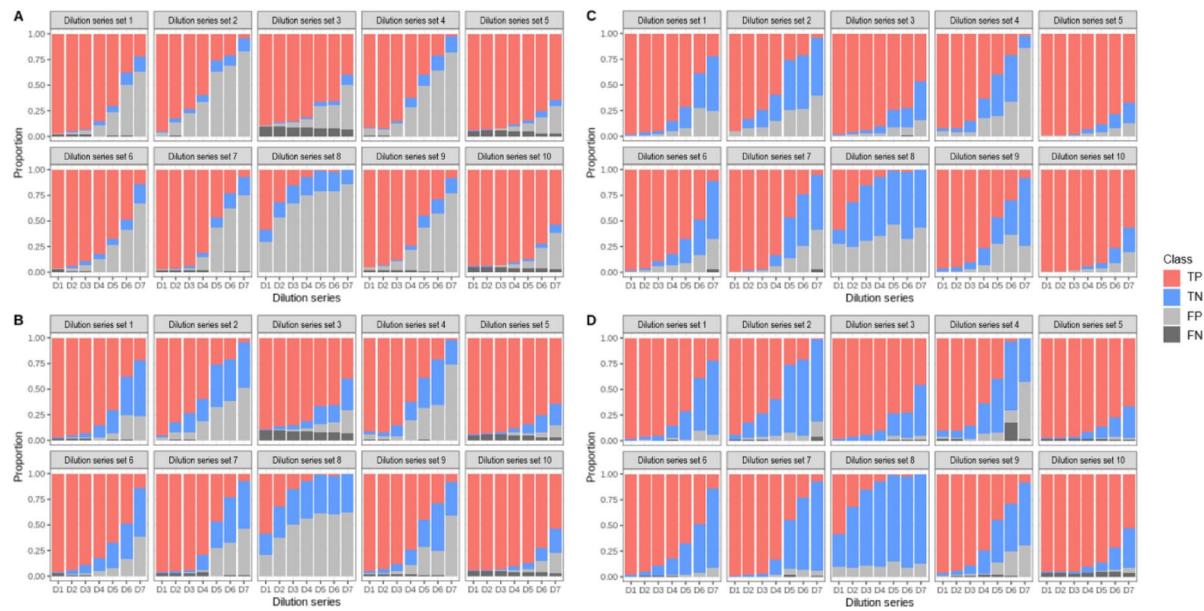
**Fig. 5.** Alpha diversity calculated by Chao1 estimating species richness showed that more types of ASVs usually were removed by CleanSeqU than Decontam, Microdecon, and SCRUB across all dilution stages. An asterisk (\*\*\*) indicates that Wilcoxon rank  $P < 0.001$ .

between these contexts and will remove the taxon as a contaminant across all samples. To address this issue, Microdecon<sup>43</sup> partially removed possible contaminants by calculating the ratio of taxa found in the controls to anchor contaminants. However, Microdecon processes only one sample at a time, disregarding the data shared among the samples. Over the last decade, several computational techniques have been proposed for tracking and identifying potentially complex microbial community origins, a process known as “microbial source tracking.” These methods have shown great promise, particularly for quantifying contaminants<sup>44,46,47</sup>. In particular, SCRUB<sup>44</sup> performs highly and precisely identifies and removes latent contamination in a sample of interest, but also enables the partial removal of taxa that are both contaminants and present in the ecosystem of interest. Notably, it handles well-to-well leakage, in which material from biological samples leaks into controls during experimental procedures, especially during DNA extraction. In Decontam and Microdecon, truly present taxa accompanied by well-to-well leakage were misclassified as contaminants and removed.

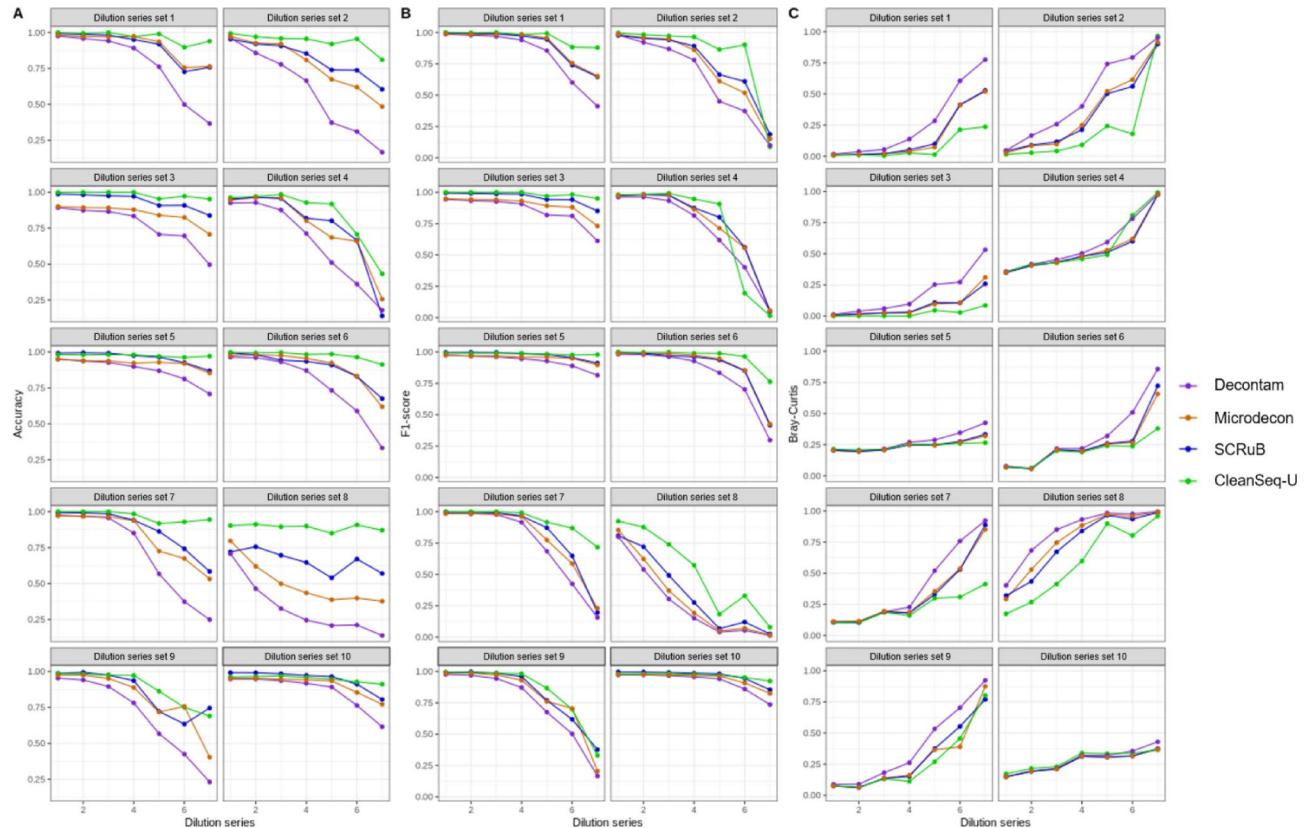
Despite high performance of SCRUB, it functions well when controls represent multiple distinct contamination sources that potentially affect the samples of interest. However, it is difficult to obtain multiple controls that reveal as many distinct contamination sources as possible during the actual experimental process. Additionally, the contaminant taxonomic profile changes over time according to the researcher, external environments, and seasons; therefore, blank extraction controls should be included and sequenced for every batch of extraction<sup>48</sup>.

CleanSeqU makes it simple to apply because it uses a single blank extraction control in the processed batch to eliminate contamination from the samples of the relevant batch. Furthermore, CleanSeqU consists of conceptual and intuitive rules for distinguishing contaminants from true features and can be applied to any data regardless of experimental method with some modification and adjustments.

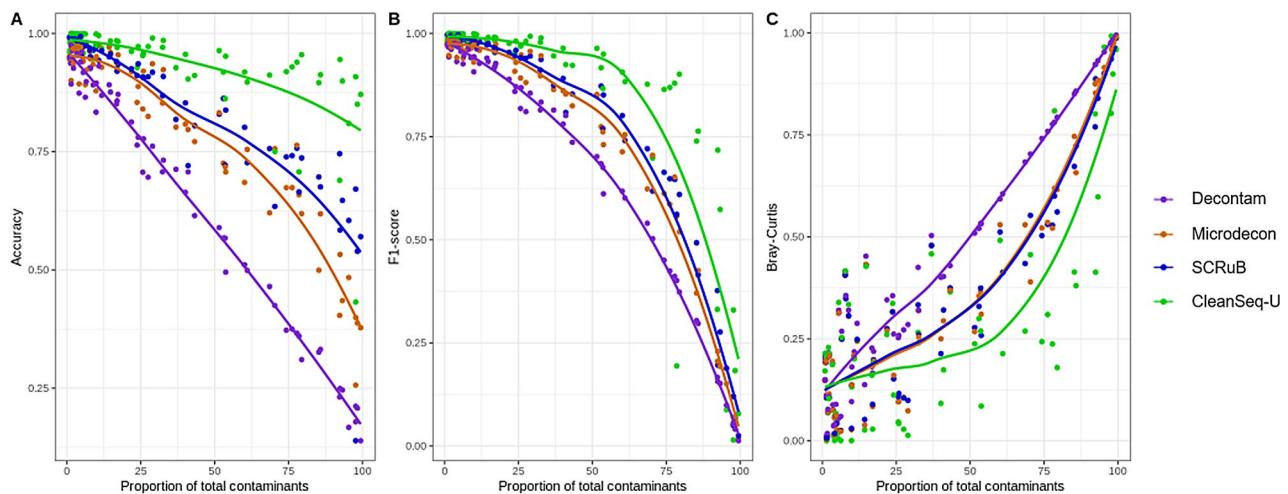
CleanSeqU outperformed SCRUB as well as Decontam and Microdecon, as shown by the accuracy, F1-score and beta-dissimilarity results in the evaluation study using a vaginal microbiome dilution series. The better performance of CleanSeqU was due to its higher specificity and PPV than those of the other algorithms (Supplementary Fig. S1). CleanSeqU correctly removed more contaminant ASV features and reduced the estimates of alpha diversity. Specifically, the majority of contaminant ASV features that were removed by CleanSeqU but remained in Decontam, Microdecon, and SCRUB correspond to category 3 ASVs, which do not exist in the blank extraction controls (Supplementary Fig. S2). Because contaminants present at low abundances may not be adequately represented in negative controls, low-abundance contaminants in the sample of interest are not effectively removed by the algorithms that relies solely on negative controls for contaminant removal.



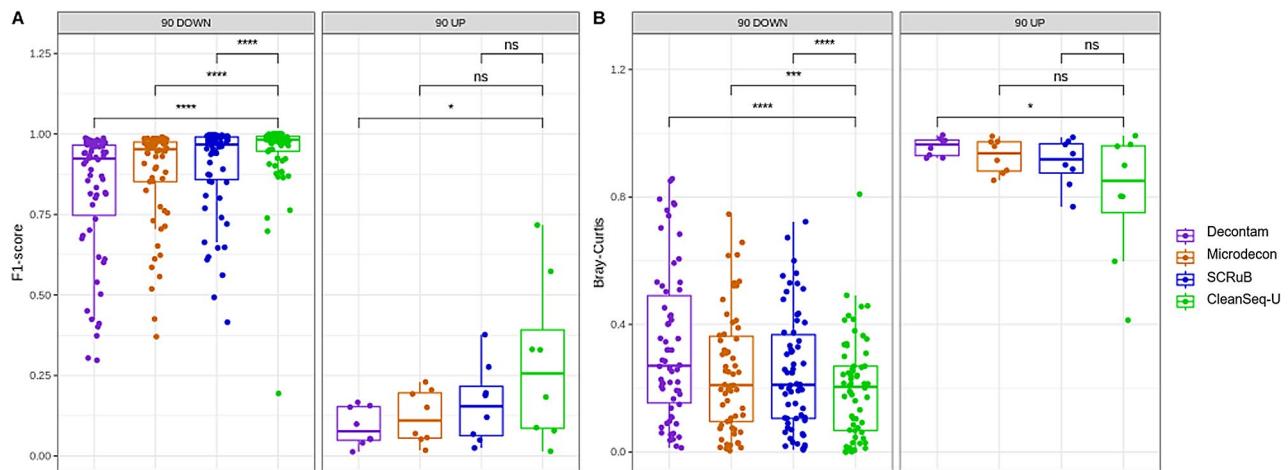
**Fig. 6.** Stacked bar plot for classification of ASVs in each decontamination algorithms. (A) Decontam stacked bar plot, (B) Microdecon stacked bar plot, (C) SCRuB stacked bar plot, (D) CleanSeqU stacked bar plot. TP (true positive), undiluted vaginal community ASVs correctly classified; TN (true negative), contaminant ASVs correctly classified; FP (false positive), contaminant ASVs incorrectly classified; FN (false negative), undiluted vaginal community ASVs incorrectly classified.



**Fig. 7.** Comparison of accuracy, F1-score, and Bray–Curtis dissimilarity of four algorithms according to the dilution samples in each dilution series set. (A) Accuracy, (B) F1-score, (C) Bray–Curtis dissimilarity. CleanSeqU had higher accuracy and F1-score and showed more similar results to the ground truth than the other algorithms in most of dilution samples.



**Fig. 8.** Trend of changes in accuracy, F1-score, and Bray–Curtis dissimilarity depending on contaminant proportion using all dilution samples. (A) Accuracy, (B) F1-score, (C) Bray–Curtis dissimilarity. Both of accuracy and F1-score gradually decreased and beta-dissimilarity gradually increased as the contaminant proportion increased. Especially, the values of F1-score and beta-dissimilarity tend to change sharply in more highly contaminated samples.



**Fig. 9.** Difference in F1-score and beta-dissimilarity between decontam, microdecon, SCRuB and CleanSeqU based on the group which was divided by the contaminant proportion 90%. (A) F1-score, (B) Bray–Curtis dissimilarity. CleanSeqU showed a significantly better F1-score and beta-dissimilarity in the group with a contaminant proportion less than 90%, in contrast, there was no significant difference in those parameters in the group with a contaminant proportion more than 90%. An asterisk (\*), (\*\*\*) and (\*\*\*\*) indicates that Wilcoxon rank  $P < 0.1$ ,  $P < 0.001$  and  $P < 0.0001$ , respectively.

Although contaminants present in low proportions have minimal effect on microbial composition, their cumulative impact can significantly alter the overall bacterial profile as the number of contaminants increases. We found that a substantial portion of the contamination does not appear in the control samples. The use of a predefined database, such as CleanSeqU, can enhance the effectiveness of contaminant removal.

Although CleanSeqU showed overall better performance compared to SCRuB, there was one dilution sample where SCRuB outperformed CleanSeqU. Specifically, this occurred in dilution series set 4, in the D6 sample. *S. agalactiae* in the D6 sample of dilution series set 4 was a true feature; however, the feature was removed as contamination in CleanSeqU because its adjusted modified Z-score did not reach the cutoff threshold. If a true biological feature is present at a low proportion in one sample but occurs at a higher prevalence in other samples from the same batch, CleanSeqU's modified Z-score algorithm may falsely remove the feature due to its low abundance in the sample.

While CleanSeqU demonstrated superior performance compared to other algorithms at lower contamination rates, its effectiveness diminishes when the contamination rate becomes higher. CleanSeqU outperforms the other algorithms when the contamination rate is less than 90%, showing a significant difference in performance

(Figure. 9). However, when contamination rates exceed 90%, no significant performance differences are observed among the four algorithms. This suggests that CleanSeqU, like other decontamination methods, may struggle to maintain high accuracy in highly contaminated samples.

We evaluated the performance of four algorithms using datasets from different low-biomass environments, including a publicly available mock community dilution series. In samples with a low contamination rate, CleanSeqU demonstrated the highest accuracy among the four algorithms, reaching close to 100%. However, in samples with a contamination rate of 50–60% or higher, its performance was comparable to or lower than that of Decontam and Microdecon. This was due to the higher proportion of ASV features classified as false negatives and false positives in CleanSeqU compared to Decontam and Microdecon.

The higher false negative rate of CleanSeqU can be inferred from the fact that most ASV features classified as false negatives belonged to Category 2 (Supplementary Fig. S6). Since certain taxa in the mock community, such as *Bacillus*, *Pseudomonas*, and *Staphylococcus*, could be frequently detected as low-abundance contaminants in negative controls, these taxa likely failed to exceed the modified Z-score cutoff in high-dilution samples and were consequently removed as contaminants in CleanSeqU.

The higher false positive rate of CleanSeqU also appears to be due to the Category 2 ASV features, considering that the proportion of Category 2 ASV features classified as false positives in CleanSeqU is much higher than that in Microdecon. CleanSeqU applies the modified Z-score when an ASV is detected in three or more samples. In a small-scale mock community dilution series, contaminants present in only one or two samples may not meet this criterion and thus remain unfiltered.

SCRuB exhibited high proportions of both false-negatives and false-positives. The predominant false negative ASVs identified by the SCRuB algorithm fall into Category 2, similar to CleanSeqU. However, their proportion remains consistently high across samples with both low and high contamination rates. This may be attributed to SCRuB's inability to distinguish genuine taxa when they are detected in the sample and the negative control, both, despite their relative abundances are different. Meanwhile, most of false-positive ASV features identified by the SCRuB algorithm belonged to Category 3, suggesting that low-abundance contaminants not detected in the negative control were still not effectively removed.

Taken together, these findings suggest that at a low contamination rate, CleanSeqU consistently exhibits clearly superior decontamination performance compared to other algorithms. CleanSeqU performs well when the batch contains a sufficiently large number of samples. And, the performance of CleanSeqU may decrease in samples with a high contamination rate due to the false removal of ASVs as contaminants, as their abundance and modified Z-score may not reach the cutoff in highly contaminated samples.

We offer several additional considerations regarding the use of CleanSeqU. First, in the case of taxa that were both contaminants and truly present in the ecosystem of interest, the proportion originating from contamination was not completely removed in CleanSeqU. Second, the in-house blacklist applied in our algorithm was created using data specifically generated in our laboratory. To use CleanSeqU, this blacklist needs to be customized to each lab experiment's unique data similar to how it was developed for the current algorithm. Third, this algorithm was developed to investigate catheterized urine microbiome samples and further investigation is required to determine whether this algorithm can be applied to other low-biomass microbiome samples with different microbiome compositions. Fourth, the cutoff value of the Euclidean distance and adjusted modified Z-score may depend on existing data. Since the performance may vary depending on the number of samples and the microbial distribution in the batch, it is necessary to optimize these cutoff values for the specific analysis environment. Fifth, CleanSeqU may not be effective for samples with extremely high contamination proportions; therefore, caution should be exercised when interpreting such results. Sixth, this algorithm was specifically designed and optimized for Illumina sequencing data, as it is the most widely used platform for 16S rRNA gene sequencing in microbiome studies. Given the inherent differences in error profiles, read lengths, and biases between sequencing technologies, additional validation would be necessary to assess the performance of the algorithm on other platforms such as MinION or PacBio Sequel.

## Conclusions

CleanSeqU outperformed the previously reported decontamination algorithms including decontam, microdecon, and SCRuB using a dilution series of human vaginal microbial communities and a publicly available mock community dilution series, especially in samples with low contamination rate. It is anticipated that CleanSeqU will advance urine microbiome research by providing accurate decontaminated results, particularly for low-biomass catheterized urine samples. It's thought that study on the catheterized urine microbiome using CleanSeqU is further required.

## Data availability

The decontamination tool used in this study is available on GitHub at <https://github.com/BITsmyoon/CleanSeqU>. The original datasets analyzed, including metadata and ASV counts, are publicly accessible in the GitHub repository at <https://github.com/BITsmyoon/CleanSeqU/tree/master/data>. The raw FASTQ files generated from this study have been deposited in the European Nucleotide Archive (ENA) under the project accession number PRJEB86132.

Received: 23 October 2024; Accepted: 15 April 2025

Published online: 02 June 2025

## References

- Turnbaugh, P. J. et al. The Human Microbiome Project. *Nature* **449** (7164), 804–810 (2007).

2. Brubaker, L. & Wolfe, A. J. The new world of the urinary microbiota in women. *Am. J. Obstet. Gynecol.* **213** (5), 644–649 (2015).
3. Theis, K. R. et al. Does the human placenta delivered at term have a microbiota? Results of cultivation, quantitative real-time PCR, 16S rRNA gene sequencing, and metagenomics. *Am. J. Obstet. Gynecol.* **220** (3), 267 (2019). e1-267 e39.
4. Aho, V. T. E. et al. The Microbiome of the human lower airways: A next generation sequencing perspective. *World Allergy Organ. J.* **8** (1), 23 (2015).
5. Pohl, H. G. et al. The urine microbiome of healthy men and women differs by urine collection method. *Int. Neurourol. J.* **24** (1), 41–51 (2020).
6. Neugent, M. L. et al. Advances in understanding the human urinary microbiome and its potential role in urinary tract infection. *mBio* **11** (2), e00218-20 (2020).
7. Whiteside, S. A. et al. The microbiome of the urinary tract—A role beyond infection. *Nat. Rev. Urol.* **12** (2), 81–82 (2015).
8. Magistro, G. & Stief, C. G. The urinary tract microbiome: The answer to all our open questions? *Eur. Urol. Focus.* **5** (1), 36–38 (2019).
9. Bschleipfer, T. & Karl, I. Bladder microbiome in the context of urological disorders—Is there a biomarker potential for interstitial cystitis? *Diagnostics* **12** (2), 281–290 (2022).
10. Lee, H. Y. et al. The impact of urine microbiota in patients with lower urinary tract symptoms. *Ann. Clin. Microbiol. Antimicrob.* **20** (1), 23 (2021).
11. Brubaker, L. & Wolfe, A. J. The female urinary microbiota, urinary health and common urinary disorders. *Ann. Transl. Med.* **5** (2), 34 (2017).
12. Li, K. et al. Interplay between bladder microbiota and overactive bladder symptom severity: A cross-sectional study. *BMC Urol.* **22** (1), 39 (2022).
13. Hiergeist, A. & Gessner, A. Clinical implications of the microbiome in urinary tract diseases. *Curr. Opin. Urol.* **27** (2), 93–94 (2017).
14. Patel, S. R. et al. The microbiome and urolithiasis: Current advancements and future challenges. *Curr. Urol. Rep.* **23** (3), 47–56 (2022).
15. Jayalath, S. & Magana-Arachchi, D. Dysbiosis of the human urinary microbiome and its association to diseases affecting the urinary system. *Indian J. Microbiol.* **62** (2), 153–166 (2022).
16. Bae, S. & Chung, H. *The Urobiome and Its Role in Overactive Bladder*, 190–200 (2022).
17. Shim, J. H. et al. Clinical implications of urinary microbiome in bladder cancer. *Korean J. Urol. Oncol.* **19** (2), 71–78 (2021).
18. Choi, H. W., Lee, K. W. & Kim, Y. H. Microbiome in urological diseases: Axis crosstalk and bladder disorders. *Investig. Clin. Urol.* **64** (2), 126–139 (2023).
19. Glassing, A. et al. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* **8**, 1–12 (2016).
20. Grahn, N. et al. Identification of mixed bacterial DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3 variable regions by pyrosequencing of cloned amplicons. *FEMS Microbiol. Lett.* **219** (1), 87–91 (2003).
21. Laurence, M., Hatzis, C. & Brash, D. E. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PLoS ONE* **9** (5), e97876 (2014).
22. Salter, S. J. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12** (1), 87 (2014).
23. Nagler, M. et al. Extracellular DNA in natural environments: Features, relevance and applications. *Appl. Microbiol. Biotechnol.* **102** (15), 6343–6356 (2018).
24. Motley, S. T. et al. Improved multiple displacement amplification (iMDA) and ultraclean reagents. *BMC Genom.* **15** (1), 443 (2014).
25. Kulakov, L. A. et al. Analysis of bacteria contaminating ultrapure water in industrial systems. *Appl. Environ. Microbiol.* **68** (4), 1548 (2002).
26. Shen, H., Rogejl, S. & Kieft, T. L. Sensitive, real-time PCR detects low-levels of contamination by *Legionella pneumophila* in commercial reagents. *Mol. Cell. Probes.* **20** (3), 147–153 (2006).
27. Mohammadi, T. et al. Removal of contaminating DNA from commercial nucleic acid extraction kit reagents. *J. Microbiol. Methods.* **61** (2), 285 (2005).
28. Lo, S. C. et al. Presence of bacterial phage-like DNA sequences in commercial Taq DNA polymerase reagents. *J. Clin. Microbiol.* **42** (5), 2264 (2004).
29. Rand, K. H. & Houck, H. Taq polymerase contains bacterial DNA of unknown origin. *Mol. Cell. Probes* **4** (6), 445 (1990).
30. Corless, C. E. et al. Contamination and sensitivity issues with a real-time universal 16s rRNA PCR. *J. Clin. Microbiol.* **38** (5), 1747 (2000).
31. Dyrhovden, R. et al. Managing contamination and diverse bacterial loads in 16S rRNA deep sequencing of clinical samples: Implications of the law of small numbers. *mBio* **12** (3), e0059821 (2021).
32. Rabin, M. Inference by believers in the law of small numbers. *Q. J. Econ.* **117** (3), 775–816 (2002).
33. Reimer, L. C. et al. BacDive in 2022: The knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res.* **50** (D1), D741–D746 (2022).
34. Tang, Z. et al. Soil bacterial community as impacted by addition of rice straw and biochar. *Sci. Rep.* **11** (1), 22185 (2021).
35. Wang, Q. et al. Metagenomic insight into patterns and mechanism of nitrogen cycle during biocrust succession. *Front. Microbiol.* **12**, 633428 (2021).
36. Piro, V. C. & Renard, B. Y. Contamination detection and microbiome exploration with GRIMER. *Gigascience* **12**, giad017 (2023).
37. Komesu, Y. M. et al. Defining the relationship between vaginal and urinary microbiomes. *Am. J. Obstet. Gynecol.* **222**, 154.e1–154.e10 (2020).
38. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37** (8), 852–857 (2019).
39. Callahan, B. J. et al. DADA2: High-resolution sample inference from illumina amplicon data. *Nat. Methods* **13** (7), 581–583 (2016).
40. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11** (12), 2639 (2017).
41. Nicholas, A. B. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6** (1), 1–17 (2018).
42. Davis, N. M. et al. Simplified statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6** (1), 226 (2018).
43. McKnight, D. T. et al. MicroDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environ. DNA* **1** (1), 14–25 (2019).
44. Austin, G. I. et al. Contamination source modeling with scrub improves cancer phenotype prediction from Microbiome data. *Nat. Biotechnol.* **41** (12), 1820–1828 (2023).
45. Karstens, L. et al. Controlling for contaminants in low-biomass 16S rRNA gene sequencing experiments. *mSystems* **4** (4), e00290–e00219 (2019).
46. An, U. et al. TENSIL: Microbial source tracking with environment selection. *mSystems* **7** (5), e00995–e00921 (2022).
47. Shenhai, L. et al. FEAST: Fast expectation-maximization for microbial source tracking. *Nat. Methods* **16** (7), 627–632 (2019).
48. Weyrich, L. S. et al. Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* **19** (4), 982–996 (2019).

## Acknowledgements

I would like to thank YEJI KANG for conducting the entire experiments.

## Author contributions

SM and JS conceived of the presented idea and designed the model. SM designed the computational framework and analysed the data. JS wrote the manuscript. JS is in charge of overall direction. All authors have read and approved the final version of the manuscript.

## Declarations

### Competing interests

The authors declare no competing interests.

### Ethics approval and consent to participants

All procedures involving the leftover of human vaginal samples and the study were approved by the Ethics Committee of GC Labs (GCL-2023-1075-02). All authors have provided their consent for publication.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-98875-3>.

**Correspondence** and requests for materials should be addressed to S.M.Y. or J.S.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025