METHODS
published: 09 July 2019
doi: 10.3389/fmicb.2019.01560

Using QC-Blind for Quality Control and Contamination Screening of Bacteria DNA Sequencing Data Without Reference Genome

Wang Xi†, *Yan Gao*†, *Zhangyu Cheng, Chaoyun Chen, Maozhen Han, Pengshuo Yang, Guangzhou Xiong and Kang Ning**

Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular Imaging, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

OPEN ACCESS

Edited by:
Feng Gao,
Tianjin University, China

Reviewed by:
Xuefeng Cui,
Tsinghua University, China
Xiaofeng Song,
Nanjing University of Aeronautics and Astronautics, China

***Correspondence:**
Kang Ning
ningkang@hust.edu.cn

†*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *05 November 2018*
Accepted: *21 June 2019*
Published: *09 July 2019*

Citation:
Xi W, Gao Y, Cheng Z, Chen C, Han M, Yang P, Xiong G and Ning K (2019) Using QC-Blind for Quality Control and Contamination Screening of Bacteria DNA Sequencing Data Without Reference Genome. Front. Microbiol. 10:1560. doi: 10.3389/fmicb.2019.01560

Quality control for next generation sequencing (NGS) has become increasingly important with the ever increasing importance of sequencing data for omics studies. Tools have been developed for filtering possible contaminants from species with known reference genome. Unfortunately, reference genomes for all the species involved, including the contaminants, are required for these tools to work. This precludes many real-life samples that have no information about the complete genome of the target species, and are contaminated with unknown microbial species. In this work we proposed QC-Blind, a novel quality control pipeline for removing contaminants without any use of reference genomes. The pipeline merely requires the information about a few marker genes of the target species. The entire pipeline consists of unsupervised read assembly, contig binning, read clustering, and marker gene assignment. When evaluated on *in silico*, *ab initio* and *in vivo* datasets, QC-Blind proved effective in removing unknown contaminants with high specificity and accuracy, while preserving most of the genomic information of the target bacterial species. Therefore, QC-Blind could serve well in situations where limited information is available for both target and contamination species.

Keywords: quality control, contamination screening, metagenome, next generation sequencing (NGS), novel pipeline

IMPORTANCE

At present, many sequencing projects are still performed on potentially contaminated samples, which bring into question their accuracies. However, current reference-based quality control methods are limited as they need either the genome(s) of target species or contaminations. In this work we propose QC-Blind, a novel quality control pipeline for removing contaminants without any use of reference genome. Evaluations performed on *in silico*, *ab initio* and *in vivo* datasets proved that QC-Blind is effective in removing unknown contaminants with high specificity and accuracy, while preserving most of the genomic information of the target bacterial species. Therefore, QC-Blind is suitable for real-life samples where limited information is available for both target and contamination species.

INTRODUCTION

As next generation sequencing (NGS) techniques become popular, such as those based on the Illumina platform (Salter et al., 2014), the need for accurate analysis of these sequencing data has likewise become increasingly urgent. At present, many sequencing projects are still performed on potentially contaminated samples, which bring into question their accuracies. This situation calls for high-performance quality control (QC) tools for high-throughput sequencing data.

In most research scenarios, a target species is considered as the organism under study, while other species are seen as contaminants. These contaminants often include the microbial species found in the environment of the molecular biology laboratories (Concolino et al., 2014; Strong et al., 2014; Rütti and Widmann, 2015). Their interference with the sequencing would compromise the precision and reproducibility of the analysis (Strong et al., 2014). Usually, these contaminations consist of not one, but a mixture of microbial species (Jung et al., 2016). Previous studies in the removal of contaminants are more concerned with water or soil samples, where relatively comprehensive bacteria genera profiles exist (Salter et al., 2014; Strong et al., 2014). These QC tools (Paszkiewicz et al., 2014) include Kontaminant (Leggett et al., 2013), FastQC (Brown et al., 2017), and many other tools (Guo et al., 2014; Coleman et al., 2015; Chen et al., 2017; Sheng et al., 2017). However, contaminants removal becomes difficult when the profile of the sample is unknown (Tanner et al., 1998; Henderson et al., 2013). The problem becomes much more difficult when the reference genomes for both the target and the contaminations are not available (Jung et al., 2016; Vikram et al., 2016).

One strategy to identify and remove contaminants is the metagenomic approach (Simon and Daniel, 2011), which facilitate taxonomical and functional analyses of the contaminating microbial genomes. A few methods based on this strategy have already been proposed: SourceTracker (Knights et al., 2011) applies Bayesian inference to estimate the composition and abundance of microbial contaminations, while DeconSeq (Schmieder and Edwards, 2011) deals with possible contamination from human through long reads alignment. Previously we have also published a method, QC-Chain, to differentiate the reads from target species and contaminations (Zhou et al., 2013), based on contig clustering (Strous et al., 2012). However, the false positive rate of read assignment remained high, and potentially valuable information was not considered in that method. For instance, knowledge of the abundance of a certain target species among multiple samples (with similar contaminations) were not utilized (Zhou et al., 2013).

In this study, we proposed QC-Blind, a pipeline for bacteria NGS data quality control and contamination screening with high specificity and accuracy. This pipeline requires only a few marker genes for differentiating reads from target bacteria species and bacteria contaminations. QC-Blind could identify and remove contaminations that were introduced during sample handling, and recover genomes from mixed cultures/environmental samples. Extensive downstream performance evaluations based on *in silico*, *ab initio*, and *in vivo* datasets, showed the method to be effective. As most microbial contaminations could be removed and almost complete genomic information of target bacteria species could be preserved after processing, this pipeline is shown to be a fairly good solution for quality control and contamination screening of bacteria DNA sequencing data.

MATERIALS AND METHODS

The general process flow of QC-Blind is as follows. First, reads are assembled into contigs. The contigs are clustered into species-level groups by species abundance and sequence features. Then, the marker genes of the target species [generated through MetaPhlAn2 (Truong et al., 2015) and manual curation] are utilized to identify the contig clusters for the target species. Here, the main issues include the assembly and clustering accuracies, as well as the specificity of the contig clusters for target species. To perform this fine-tuning, we put the method to very thorough tests on simulated, *ab initio* and *in vivo* datasets. The simulated and sequencing data were deposited to NCBI SRA with project access number PRJNA491366.

Simulated and Real Datasets

Three types of metagenomic datasets have been utilized in this study: simulated, *ab initio* and *in vivo* (**Figure 1**; **Table 1**).

Metagenomic Data Simulation

For *in silico* simulated datasets, reads of target, and contamination species were generated by NeSSM (Jia et al., 2013). In this study, we assume only one target bacteria species is present in each sample. The target bacteria species used in this study include three model organisms: *Bacillus subtilis, Staphylococcus aureus, Escherichia coli* (dataset A, dataset B, dataset C, **Table 1**). Their reads were mixed with reads generated from the genome of 5 or 10 representative species in human oral microbial community [referred to as HOB(5/10)], which were used as possible human contaminations (Hasan et al., 2014). Gradient proportions of reads from target species were set to 5, 35, 65, 95%. We also combined *Saccharomyces cerevisiae* with *B. subtilis* and 10 oral bacteria to simulate a special condition with eukaryotic contamination (dataset D, **Table 1**). In each dataset, over 10 million pair-end reads with 100X coverage were generated at the length of 120 bp every 200 bp bin. All other parameters were set as default (Jia et al., 2013).

ab initio Dataset Preparation

For *ab initio* datasets, we mixed the real sequencing data of *B. subtilis* strain 168 with real metagenomic sequences from human saliva samples (dataset E, **Table 1**), with the relative proportion of reads from target species (*B. subtilis*) set at 35, 65, 95% for different datasets. These samples were named AB_BS_35%, AB_BS_65%, AB_BS_95%, respectively. In order to maintain them in their natural state, we did not perform any filtration of any human contamination from saliva samples (Hasan et al., 2014), even though the lack of filtration may adversely affect the contig assembly and clustering process.

For both *ab initio* and *in vivo* dataset preparations, real DNA extraction and sequencing were needed. Their procedures are

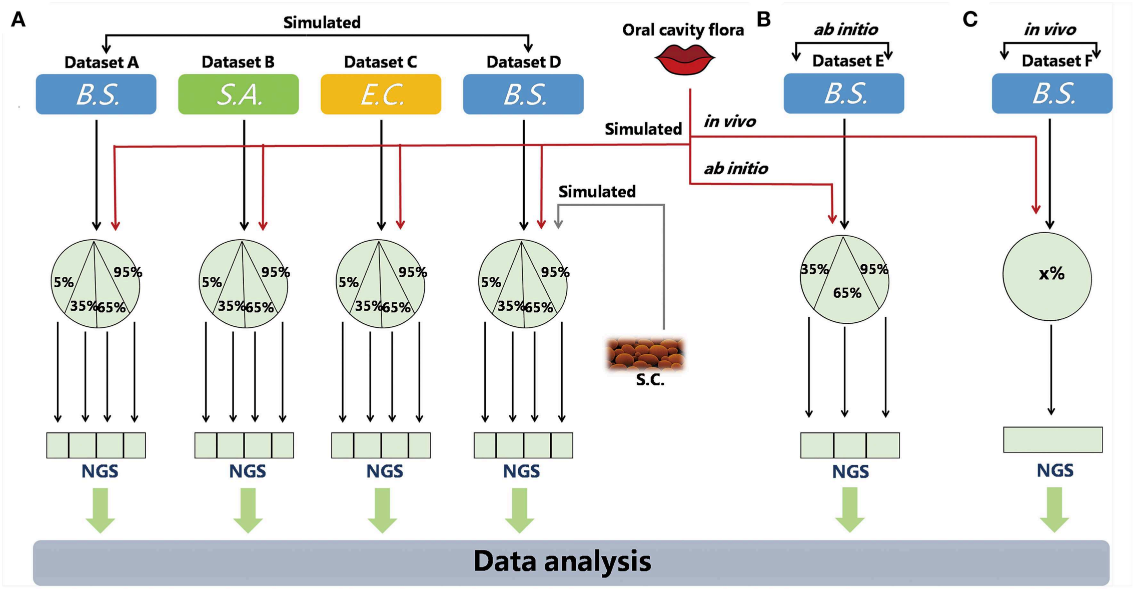

FIGURE 1 | Simulated and real datasets. QC-Blind were tested on three types of data, simulated, *ab initio*, and *in vivo* in this study. *B.S., S.A., E.C* represent the three target species (*Bacillus subtilis, Staphylococcus aureus*, and *Escherichia coli*.) *S.C.* represents *Saccharomyces cerevisiae*, a source of contamination. The red lines symbolize sequence data from oral cavity flora, while the black lines represent sequence data from target species. **(A)** In simulated datasets, the target species were mixed with human saliva (or S.C.) at 5, 35, 65, and 95%. The gray line represents *Saccharomyces cerevisiae* sequence data. **(B,C)** For *ab initio* and *in vivo* datasets, *B. subtilis* (extracted DNA solution and bacteria culture) were mixed with human saliva, the DNA proportions of which were set to 35, 65, 95, or unknown ratio (x%). The DNA proportions of which were set to 35, 65, 95, or unknown (x%).

TABLE 1 | Information about simulated and real metagenomic datasets.

Dataset	Dataset A	Dataset B	Dataset C	Dataset D	Dataset E	Dataset F
Type	Simulated	Simulated	Simulated	Simulated	*ab initio*	*in vivo*
Target	*Bacillus subtilis*	*Staphylococcus aureus*	*Escherichia coli*	*Bacillus subtilis*	*Bacillus subtilis*	*Bacillus subtilis*
Sample Names	Simu_BS_5% Simu_BS_35% Simu_BS_65% Simu_BS_95%	Simu_SA_5% Simu_SA_35% Simu_SA_65% Simu_SA_95%	Simu_EC_5% Simu_EC_35% Simu_EC_65% Simu_EC_95%	Simu_BS_5%(SC) Simu_BS_35%(SC) Simu_BS_65%(SC) Simu_BS_95%(SC)	AB_BS_35% AB_BS_65% AB_BS_95%	Real_BS
Contamination	HOB(5/10)	HOB(5/10)	HOB(5/10)	HOB(5/10) and *S.C.*	HOB(5/10)	HOB(5/10)

Dataset A-D were simulated datasets, Dataset E was ab initio dataset, and Dataset F was in vivo dataset. Naming style: for simulated datasets, the target species and the relative proportion of reads from target species were provided. For example, "Simu_BS_5%" means that Bacillus subtilis was the target species, and reads from this target species compose of 5% of all reads in this sample. For ab initio datasets, the sample names were defined similarly. The reference genomes of all species were downloaded from NCBI Microbial Genomes website (https://www.ncbi.nlm.nih.gov/genome/microbes/).

detailed in the sub-section "*in vivo* sample preparation, DNA extraction and sequencing".

In vivo Sample Preparation, DNA Extraction, and Sequencing

The *in vivo* datasets used in this study were metagenomic (not 16s rRNA) datasets from real community samples prepared as follows: after being activated, strain *B. subtilis* 168 was cultured overnight till its OD_{600} value reaches between 0.6 and 0.8. All the *B. subtilis* were centrifuged at 12,000 rev min^{-1} (12114 g) for the following experiments. The fresh saliva was collected from three healthy adults abstaining from drinking water or gargling for about 30 min before sample collection. Then 200 μl fresh saliva was added to the *B. subtilis* culture before DNA extraction, amplification and sequencing. This sample was named Real_BS (dataset F, **Table 1**).

Modified CTAB method (Porebski et al., 1997; Cheng et al., 2014a,b) was chosen for obtaining high MW metagenomic DNA of samples. Five milliliter lysis buffer (Cetyl Trimethyl Ammonium Bromide, 1% w/v; EDTA, 100 mM; NaCl, 1.5 mol l^{-1}; Sodium phosphate, 100 mmol l^{-1}; Tris-Cl pH 8.0, 100 mmol l^{-1}) and 20 μl Proteinase K was added into 15 ml liquid sample (*B.S.* culture or its mixture with saliva), followed by gentle shaking at 100rev min^{-1}. SDS was added to a final concentration of 1% and the reaction was incubated at 65°C for 30 min with intermittent shaking. After the above steps, an equal volume of saturated phenol, chloroform and isoamyl alcohol (25: 24:1) was added to the mixture and centrifuge at 12,000 rev min^{-1}

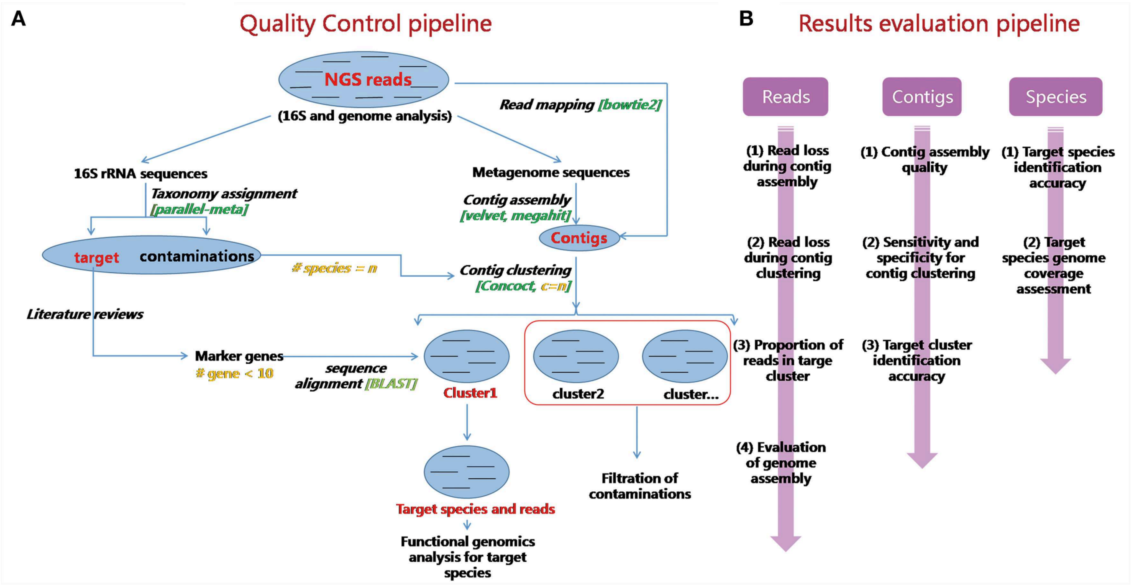
A
Quality Control pipeline
NGS reads
(16S and genome analysis)
Read mapping [bowtie2]
16S rRNA sequences
Taxonomy assignment [parallel-meta]
Metagenome sequences
Contig assembly [velvet, megahit]
Contigs
target
contaminations
# species = n
Contig clustering [Concoct, c=n]
Literature reviews
Marker genes
# gene < 10
sequence alignment [BLAST]
Cluster1
cluster2
cluster...
Target species and reads
Filtration of contaminations
Functional genomics analysis for target species
B
Results evaluation pipeline
Reads
Contigs
Species
(1) Read loss during contig assembly
(2) Read loss during contig clustering
(3) Proportion of reads in targe cluster
(4) Evaluation of genome assembly
(1) Contig assembly quality
(2) Sensitivity and specificity for contig clustering
(3) Target cluster identification accuracy
(1) Target species identification accuracy
(2) Target species genome coverage assessment

FIGURE 2 | Schematic demonstration and evaluations of QC-Blind. **(A)** An overview of the QC-Blind pipeline. Based on NGS data, on one hand, 16S rRNA genes were extracted and compared with bacteria 16S rRNA database to count the number of species; on the other hand, raw reads were assembled into contigs, binned into species-level groups, and then target clusters were identified through mapping marker genes of target species onto them. Within these steps, the marker genes mapping step would finally determine which cluster of contigs and reads belong to the target species. **(B)** Evaluations for QC-Blind were performed at three levels: read, contig, and species. For clustering quality, the purity, and concentration of target and other clusters were measured. For contamination removal, sensitivity, and specificity were calculated. For data loss, target reads and contigs that fail to pass at each step were counted. For functional analysis, the coverages of target genomes at base and gene level were calculated.

(12,114 g) for 10 min to collect supernatant that are free from protein. This step is then repeated once. Metagenomic DNA was precipitated with 0.6 volumes of isopropanol for 30 min at −20°C and pelleted by centrifugation at 12,000 rev min^{-1} (12,114 g) for 10 min. DNA was washed twice with 70% ethanol and finally dissolved into a 200 µl of TE (1X), pH 8.0. The genomic DNA of *B. subtilis* 168 and the mixture of *B. subtilis* 168 with human saliva were extracted with Soil Genomic DNA kit(CWBIO).

Before sequencing with Illumina Miseq-2000, DNA samples were quantified using a Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA) and its quality was checked on a 0.8% agarose gel. 5–50 ng metagenomic DNA in high quality was used as the template for amplifying the V3-V4 diameter region of 16S rRNA genes for each individual sample, with ″5′-CCTACGGRRBGCASCAGKVRVGAAT-3′″ used as the forward primer and ″5′-GGACTACNVGGGTWTCTAATCC-3′″ as the reverse primer. The sequencing library was constructed using a MetaVxTM Library Preparation kit (GENEWIZ, Inc., South Plainfield, NJ, USA). Then indexed adapters were added to the ends of 16S rDNA amplicons by limited-cycle PCR. Verified by Agilent 2100 Bioanalyzier (Agilent Technologies, Palo Alto, CA, USA), and quantified by Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA) and real-time PCR (Applied Biosystems, Carlsbad, CA, USA), the DNA libraries were normalized for sequencing. All sequencing reactions were performed on the Illumina MiSeq platform using paired-end sequencing technology (2*300 bp).

ANALYTICAL PROCEDURE

An overview of our quality control pipeline is shown in **Figure 2A**. First, real sequencing reads are trimmed by Trimmomatic-0.36 to remove low quality bases and reads (Bolger et al., 2014). Three leading/trailing bases are cut if their quality scores are below a certain quality threshold. Reads with lengths that are too short (<50 bp as default) are also discarded. 16S rRNA genes are extracted from remaining reads, for species identification and quantification. Then read assembly, contig binning, and marker gene mapping are performed in sequential order.

Identification of Target and Contamination Species

The taxonomical profiles were generated by the Parallel-Meta pipeline (version 2.0) (Jia et al., 2013). 16S rRNA sequences were extracted from raw sequencing data through an HMM model, and these sequences were searched against the Greengene database (http://greengenes.secondgenome.com/) to identify their species.

At the contig binning step, the total number of species identified was used to set the initial cluster number, which aims to provide better accuracy for clustering. An additional eukaryotic18S rRNA database was used as reference when processing the dataset with *S. cerevisiae.* Though choosing the number of clusters become difficult for unknown contaminants whose information are not recorded in 16S RNA or 18S rRNA database, this approach is practical and performs well for target identification and contamination filtration.

Assembly of Contigs From Community Data

Two assemblers were applied to assemble contigs from community reads. One of the assemblers selected was Velvet (Jia et al., 2013), which could deal with *de novo* genomic assembly and short sequencing reads alignment. For Velvet, we use the *velveth* command to construct the dataset as preparation work, and *velvet* command to build the *de Bruijn* graph from the *k-mers* obtained by *velveth* and extract the contigs. We use $k = 12$, and set other parameters to auto or default. The other one was MEGAHIT (Li et al., 2015), an assembler designed specifically for complex metagenomics via succinct *de Bruijn* graph. It is worth mentioning that by using these two assemblers, the abundance information has been intrinsically taken into consideration.

For simulated metagenomic datasets, assembly was performed on two assemblers to compare their performance. Basic assembly statistics were extracted and compared. As MEGAHIT has been shown to be superior to Velvet on simulated data, only MEGAHIT was used to process *ab initio* and *in vivo* dataset.

Contig Binning With Concoct

Contig binning is another key step of QC-Blind. Out of all the existing and evaluated binning algorithms, CONCOCT is selected because: firstly, both sequence composition and coverage across multiple samples were considered in contig binning; and secondly, it could handle both single sample and multiple samples. These make CONCOCT a suitable choice for batch processing of possibly contaminated samples (Knights et al., 2011). For multiple species, co-assembly is a necessary prerequisite to running CONCOCT, as CONCOCT takes contigs as input to maximize the number of genomes that could potentially be resolved. Contigs are limited to lengths from 1,000 to 10,000 bp, the lower limit filters low quality contigs while the upper limit cuts fragments for more statistical weight. The number of clusters was precisely determined with 16S rRNA method by Parallel-Meta (Su et al., 2014). Contigs would be clustered into species-level groups after CONCOCT processing. Again, we have to emphasize that by using CONCOCT, the abundances (read depth) of contigs in each of the clusters are similar.

Marker Gene Selection and Mapping

Utilizing marker genes for target species cluster identification helps us overcome the situation where we do not have the complete or partial reference genome. This is realistic for most of the contemporary sequencing tasks, in which we only know a few of the targets' marker genes (Davey et al., 2011; Sunagawa et al., 2013).

Except for a few well-studied species that have had their whole genome sequenced, we have knowledge on only a few genes for most of the other species, such as important regulators in their metabolism, proliferation, or 16S rRNA genes. Therefore, we can only utilize these genes as markers to identify contig clusters that belong to the target species. Marker genes were generated by the combination of MetaPhlAn2 (Truong et al., 2015) and manual curation. The majority of these marker genes are obtained from MetaPhlAn2's unique clade-specific marker genes identified from ∼17,000 reference genomes (http://huttenhower.sph.harvard.edu/metaphlan2). For those species (newly sequenced) not included in MetaPhlAn2's marker gene sets, we have extracted marker genes manually and ensure they are clade-specific.

The clusters generated in the previous step were mapped to marker genes of target species by BLAST (e-value cutoff = 1e-20). The number of usable marker genes are species-dependent. The more unique the genes, the more specific the identification would be. Through MetaPhlAn2 selection and literature review, marker genes ftsZ, lytF, nsrR, spo0A, ygxB, yjbH, yjbI were selected for *B. subtilis* (Amann et al., 1995; Lindgreen et al., 2016; Vikram et al., 2016), acpP, casA, cof, dxs, fabB, fabF, leuO, tesA, uidA were chosen for *E. coli* (Coleman et al., 2015; Katta et al., 2015; Freedman et al., 2016; Yan et al., 2016; Yu et al., 2017).

After marker gene assignment, contigs containing marker genes for target species are identified as belonging to the target species (defined as target contigs). Based on these assignments, raw reads were mapped to contigs identified as belonging to the targets with BOWTIE2 (Langmead and Salzberg, 2012) (defined as target reads).

EVALUATION METHODS

Statistics of total reads number and target reads number in every step could be evaluated with reference genome used. For *ab initio* and *in vivo* datasets, only the target reads or contigs were classified through mapping to reference of *B. subtilis*, since it is impractical to classify each read from contaminations to their source species, especially when many of them do not yet have their whole genome sequenced (Hasan et al., 2014).

Our assessment of QC-Blind is based on the purity of the clusters, target distribution, sensitivity, specificity, data loss and coverage (**Figure 2B**).

Purity of Clusters

DS (dominant species) is defined as the species whose contigs (and reads) outnumber other species' in the cluster. In each cluster, all the contigs were mapped to their reference genome database consisted of both target bacteria and contaminations by BLAST, to identify their source species. Reads were mapped to contigs by BOWTIE2 (Langmead and Salzberg, 2012) and thus inheriting taxonomical information from contigs. Purity was defined as the proportion of contigs or reads of the DS in each cluster, after contig binning through CONCOCT (Knights et al., 2011; Formulas 1, 2).

$$Purity_{\_contig} = \frac{\text{\# of } DS \text{ contigs in a cluster}}{\text{\# of all contigs in a cluster}} \times 100\% \quad (1)$$

$$Purity_{\_read} = \frac{\text{\# of } DS \text{ reads in a cluster}}{\text{\# of all reads in a cluster}} \times 100\% \quad (2)$$

Purity of each cluster was categorized and evaluated at three levels, 100%, 90%+, 80%+. The proportion of clusters exceeding these thresholds among all the clusters reflects the quality of the clustering.

Target Contig and Target Read Concentration and Distribution

We define clusters that contain contigs (and reads) from target species (TS) as target clusters (TC), no matter how many contigs are in the TC. The proportion of target contig or target read in each clusters, as well as their distributions among all TC are measured in Formulas (3, 4).

$$Targetconcentration_{\_contig} = \frac{TS \text{ contigs in one } TC}{TS \text{ contigs in all } TCs} \times 100\% \quad (3)$$

$$Targetconcentration_{\_read} = \frac{TS \text{ reads in one } TC}{TS \text{ reads in all } TCs} \times 100\% \quad (4)$$

Target concentrations in all clusters would thus form a distribution, which is named "Target distribution." A more biased target distribution indicates a better clustering result, and *vice versa*.

Sensitivity and Specificity

"Target/Contamination Dichotomization" were also measured by sensitivity and specificity, which provided quantitative information about contaminations correctly removed and target sequences successfully preserved (Formulas 5–8). Here, the contigs or reads of target species in the clusters found by marker genes alignment were considered as true positive (TP), while those of other species were considered as false positive (FP). The contigs or reads mapping on the target genome were considered as ground truth (GT), which contained those identified or unidentified contigs or reads in target clusters.

$$Sensitivity_{\_contig} = \frac{\text{Target contigs in } TCs \text{ identified by marker genes}}{\text{All target contigs}} \times 100\% \quad (5)$$

$$Sensitivity_{\_reads} = \frac{\text{Target reads in } TCs \text{ identified by marker } genes}{\text{All target reads}} \times 100\% \quad (6)$$

$$Specificity_{\_contig} = \frac{\text{Target contigs in } TCs \text{ identified by marker genes}}{\text{All contigs in } TCs \text{ identified by marker genes}} \times 100\% \quad (7)$$

$$Specificity_{\_reads} = \frac{\text{Target reads in } TCs \text{ identified by marker genes}}{\text{All reads in } TCs \text{ identified by marker genes}} \times 100\% \quad (8)$$

Coverage Evaluation

We performed functional evaluations to examine how much of the target bacteria species sequencing data has been kept after the quality control process, and whether they retain the functional genomics of the target species. The coverage in target genome was calculated at both base and gene levels (Formulas 9, 10). At the base level, the number of bases in the genome that have been mapped for one or more times was measured (mapped base, MB) and compared with the length of the whole genome (total base, TB). Statistical data were summarized after target reads alignment. At the gene level, genes covered by contigs in our target clusters were viewed as mapped genes (MG), while the genes for the all species were considered as total gene (TG). Genes preserved after the processing were identified by the *intersect* command of BEDTools (Quinlan and Hall, 2010). Obviously, the gene and base coverage for unprocessed simulated data are both 100%.

$$BaseCoverage = \frac{\text{Mapped base}}{\text{Total bases}} \times 100\% \quad (9)$$

$$GeneCoverage = \frac{\text{Mapped genes}}{\text{Total genes}} \times 100\% \quad (10)$$

RESULTS AND DISCUSSION

Simulated metagenomic datasets, each consisting of 4 samples with different multi-species configurations, were selected to benchmark the performance of QC-Blind for target species identification (**Figure 1**). Here we present the results for dataset consisting of *B. subtilis* as target species (read proportion of 5, 35, 65, and 95%) and 10 human oral bacteria as contaminations (Hasan et al., 2014) (**Table 1**). Results for target species of *S.A.*, *E.C.*, and *B.S* mixed with *S.C*, were shown in **Supplementary File 1**.

Target and Contamination Species Identification

Target species can be completely identified at genus level, along with 88% of contaminations identified on average (**Table 1** in **Supplementary File 2**). The setting of number of species acquired in this step could increase the purity of clustering, although the unsupervised contig binning method utilized in QC-Blind does not necessarily require this parameter (Knights et al., 2011). However, it was also found that most eukaryotic organisms (usually not contaminations) would not be identified at species level by this method.

Read Assembly and Contig Binning

For read assembly, results based on Velvet (Zerbino and Birney, 2008) and MEGAHIT (Li et al., 2015) were compared, showing that MEGAHIT would generate less contigs with longer N50 (e.g., for BS 5%, 603 contigs were generated with N50 = 154,200

TABLE 2 | Assembly result summary for *ab initio* and *in vivo* datasets.

Dataset	AB_BS 35%	AB_BS 65%	AB_BS 95%	Real_BS
Type	*ab initio*	*ab initio*	*ab initio*	*in vivo*
Contig number	1136824	182802	333868	90457156
N50 (bp)	490	413	432	540
Average length (bp)	489	424	519	532

Contig number, N50, and average length of contigs assembled by MEGAHIT from real datasets are shown. More information of AB_BS_35%, AB_BS_65%, AB_BS_95%, and Real_BS are shown in ***Table 1***.

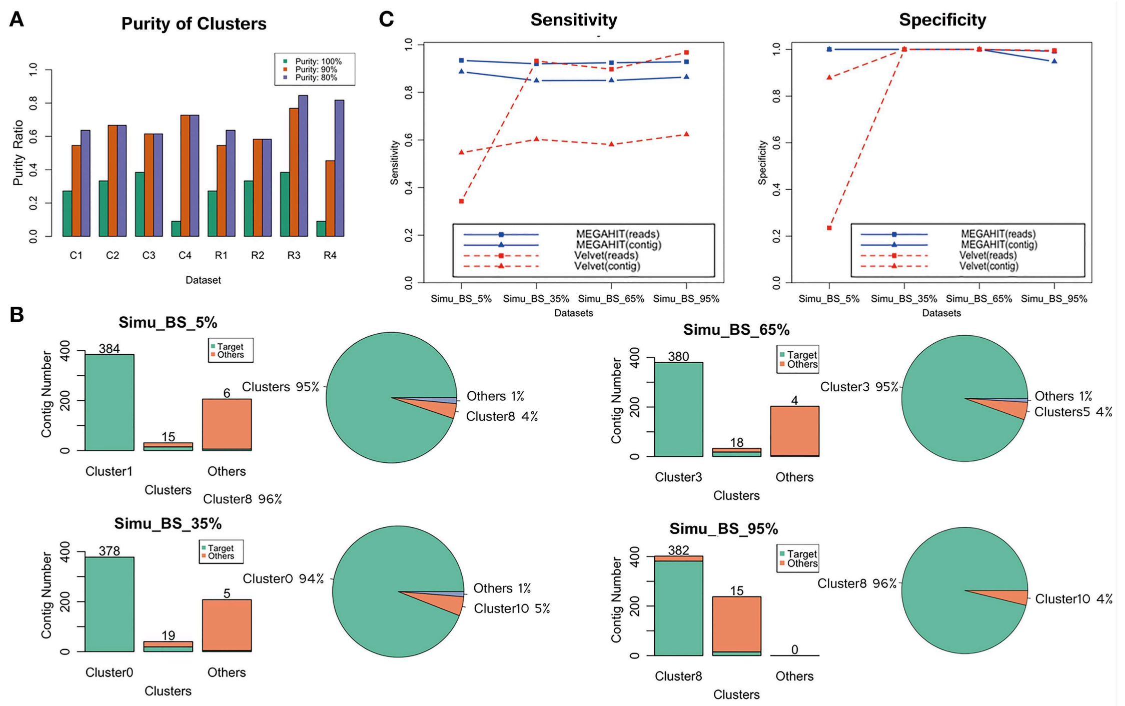

FIGURE 3 | Evaluation of contig binning and contaminations removal. **(A)** 11, 12, 13, 11 clusters were generated, respectively for the four datasets with target reads fraction of 5, 35, 65, 95%. Bar plots show the purity ratio of reads and contigs in all clusters in each sample. C1–C4 and R1–R4 stands for contigs and reads of the four samples, respectively. Purity were measured at 100, 90%+, 80%+ level, in which green bar represents 100% purity, the orange bar represents purity over 90%, the violet bar represents purity over 80%. **(B)** Target distribution and concentration in four samples. Clusters with the largest and second largest number of target contigs were shown, combined with all other clusters that has at least one target contig. The green bar represents target species, while the orange bar represents other species. Distribution of all target contigs were also shown in pie charts. **(C)** Sensitivity and specificity of the results of the four groups for MEGAHIT-BLAST (blue lines) and Velvet-BLAST (red lines) pipeline. The squares stand for results based on read level and the triangles stand for results based on contig level.

by MEGAHIT, while 10,667 contigs were assembled by Velvet with N50 = 39,369), thus we deemed contigs from MEGAHIT outperform those from Velvet for downstream analysis (**Table 2** in **Supplementary File 2**).

Purity of cluster evaluation indicated that most of these contigs generated by QC-Blind were dominated by single species. At the read level, 59.6% (28 of 47) of the clusters reached 90% purity on average, and 72.3% clusters reached 80% purity (34 of 47) (**Figure 3A**). At contig level, more than 50% of the clusters reached 90% purity. For target concentration, each dataset had a single main cluster that contained over 94% target contigs. Noticeably, the dominant clusters in three simulated datasets were of 100% purity, except that purity of Simu_BS_95.0% dominant cluster 8 was 95% (**Figure 3B**).

Taken together, the contig binning method could resolve single highly concentrated and pure target cluster from multiple species. Considering possible artifacts produced during read mapping on the simulated datasets, we anticipated that the method would actually perform better for real datasets.

Evaluation of Sensitivity and Specificity for Target Species Read Assignment

The sensitivity and specificity values for target species read assignment of MEGAHIT-processed data were both high (**Figure 3C**). Sensitivity values were 92.7% on average in four samples, while specificity values of those were even higher for both target contigs and reads: 100% assignment specificity in Simu_BS_5%, Simu_BS_35%, and Simu_BS_65%, showing that the target information in target cluster can be extracted with very few contaminations remaining. However, the sensitivity and specificity evaluation of Velvet-processed data were extremely low at the dataset with 5% target reads (34.3%, compared to 93.5% in MEGAHIT), which raised question on the ability of Velvet to deal with severely contaminated data. Velvet's sensitivity at the contig level was also not optimistic.

This evaluation of sensitivity and specificity for target species read assignments showed the superiority of using MEGAHIT in QC-Blind method. Thus, in the following analyses, we adopted MEGAHIT in the QC-Blind method as default.

Data Loss in Screening Process

The information loss that we generally experience as the target information progress from raw reads, to read assembly, contig binning, and then marker gene mapping, decreases as the proportion of target species increases (**Figure 4A**). The greatest data loss on read level occurred in marker gene mapping. the proportion of reads loss were up to 5.31, 5.91, 5.32, 4.87% for each simulated dataset with target fraction from 5 to 95%. Samples

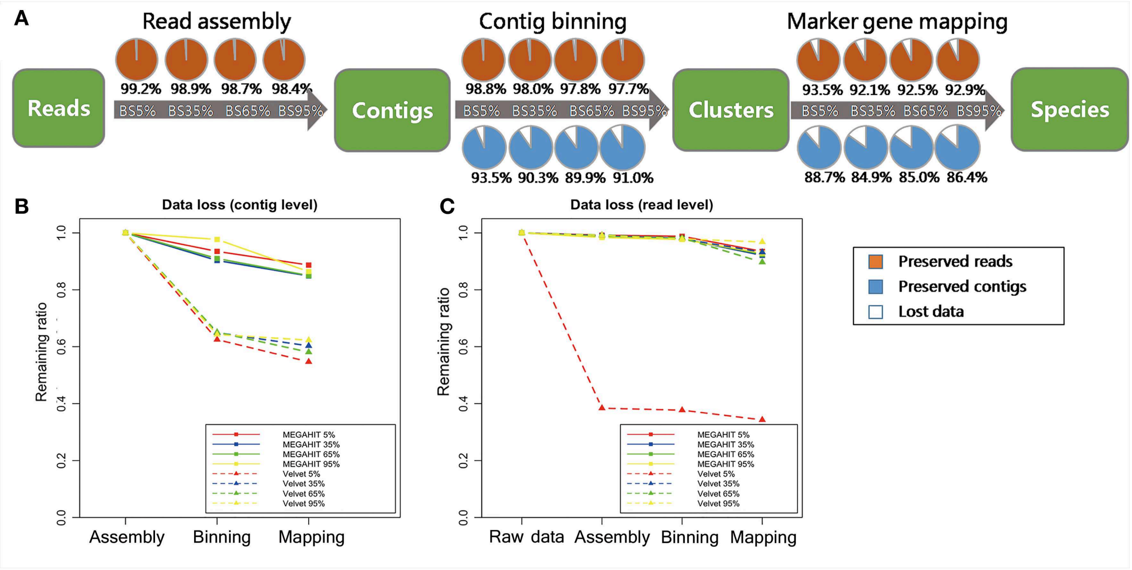

FIGURE 4 | Data loss assessment for simulated data. **(A)** Percentage of target contigs and reads preserved in each step (read assembly, contig binning, marker gene mapping), for MEGAHIT pipeline. The four circles in a row represent groups in which the frequency of target species increases from 5 to 95%. The orange portions represent preserved reads, the blue portions represent preserved contigs, while the white portions represent lost data. **(B)** Comparison of data loss in the steps of read assembly, contig binning, and marker gene mapping between MEGAHIT-BLAST (full squares) and Velvet-BLAST (dot lines and triangles) pipeline at contig level. The red, blue, green, and yellow lines represent the samples with proportion of target species from 5 to 95%. **(C)** Data loss comparison between MEGAHIT-BLAST (full squares) and Velvet-BLAST (dot lines and triangles) pipeline at read level. The red, blue, green, and yellow lines represent the samples with proportion of target species from 5 to 95%.

without a dominant species (e.g., Simu_BS_5%, Simu_BS_35%) encountered difficulty on assigning all reads correctly, as there may not be enough unique reads from them to reconstruct a complete genome for species identification (Knights et al., 2011). Certain proportions of short contigs (6.47, 9.66, 10.07, 9.05%, for 4 samples, respectively) were also filtered out during contig binning (**Figure 4B**). Data loss at read level is almost negligible (<2%) in reads assembly and contig binning (**Figure 4C**).

By contrast, analysis results on the data loss issue by Velvet was unsatisfactory, indicating its inability to deal with highly contaminated data. At contig level, over 35% contigs were lost after contig binning and mapping of Velvet-processed data in all four samples. At read level, Velvet performed comparably as MEGAHIT in three samples, except that in Simu_BS_5%, 61.6% reads were lost after contig assembly by Velvet.

Base and Gene Coverage Analysis

QC-Blind results have nearly perfect coverage of genomic information regardless of the different complexities of the samples. Through mapping these reads back to the genome, on average 93.5% of the bases could be covered for one or more times, indicating the potential of this method to reconstruct a complete genome (**Figure 5A**). More notably, this coverage is consistent among the four samples (94.1, 92.9, 93.2, 93.8%), which suggests that QC-Blind is able to work with samples of different complexities. Similarly, most of the annotated genes are also covered with processed reads (on average 93.8%). The Arginine biosynthesis pathway, a critical pathway for *B. subtilis*'s metabolism (Kunst et al., 1997), was selected as an example for examination (**Figure 5B**); all the 20 genes that are involved in this pathway could be found in the processed reads (after quality control) of the four datasets (**Figure 5C**).

The above assessments of QC-Blind based on simulated data have not only demonstrated the possibility but also the high fidelity of the reference-free QC. This performance can be attributed to the fact that the vast majority of target contigs were binned into a single cluster. Certainly, the selection of marker genes was very crucial, as their uniqueness among microbial community would assure high specificity in target/contamination classification.

Hence for simulated data, the resultant near-perfect coverage proved that the additional work performed on screening is worthwhile. We have put other simulation settings and results in **Supplementary File 1** including analytical results of Datasets B, C, and D (**Figure 1A**).

Analysis Based on *ab initio* and *in vivo* Datasets

Before bacterial contamination screening, QC-Blind was able to capture the genetic information of target species in *ab initio* and *in vivo* datasets that were contaminated by large proportion of human-oriented reads (416,679,339 reads were identified from bacteria floras of human saliva in this study). This might affect the contig assembly and clustering process.

Single dominant cluster for the target species *B. subtilis* were successfully determined in each *ab initio* and *in vivo* dataset,

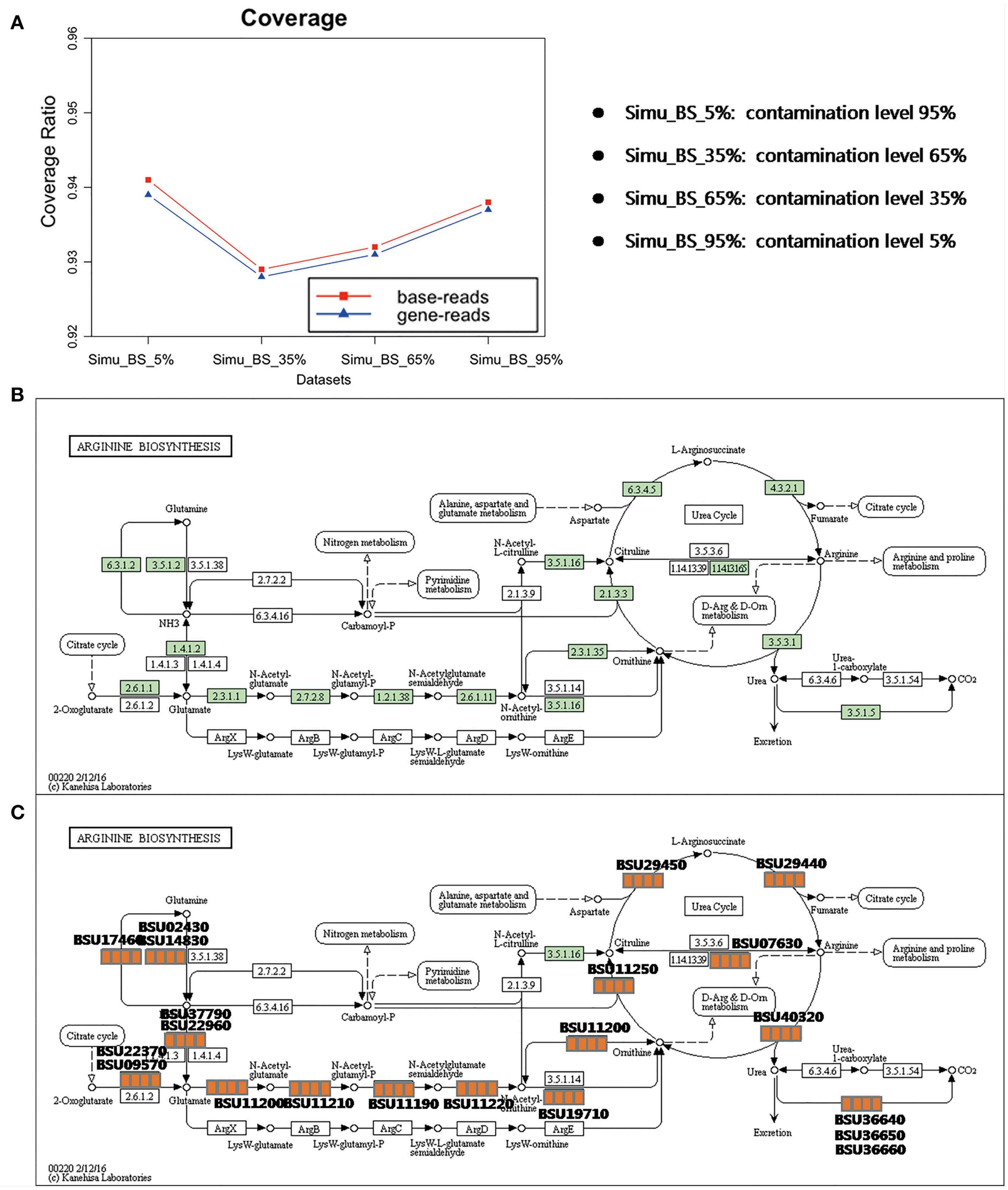

FIGURE 5 | Coverage and pathway reconstruction based on simulated data. **(A)** Y-axis represents base coverage (red line, squares) and gene coverage (blue line, triangles) of target genome as described in Materials and Methods section of this paper. X-axis stands for four samples with target species from 5 to 95%. **(B)** Arginine biosynthesis pathway of *B. subtilis* from KEGG. **(C)** Genes successfully found in processed data marked with orange squares, and four squares stands for each sample, in which orange and white stands for exist and non-exist, respectively.

with cluster 34 in AB_BS 35% containing 99.6% target reads, cluster 33 in AB_BS 65% containing 99.9% target reads, cluster 32 in AB_BS 95% containing 99.5% target reads and cluster 14, 65, 78 together containing 59.7% of target reads (**Figure 6A**), while a lot of contigs from contamination species with very few sequences were not classified into independent cluster in AB_BS 65% and AB_BS 95%. All three dominant clusters were identified by marker genes with high specificity (**Figure 6B**). However, the sensitivity of AB_BS 65% at read level dropped to 47.5% while the sensitivity of AB_BS 35% and AB_BS 95% and Real_BS remained high.

For data loss ratios, our method's performance on read datasets remained high level (**Figure 7A**), except that <20% contigs remained in AB_BS 65% with <30% reads after

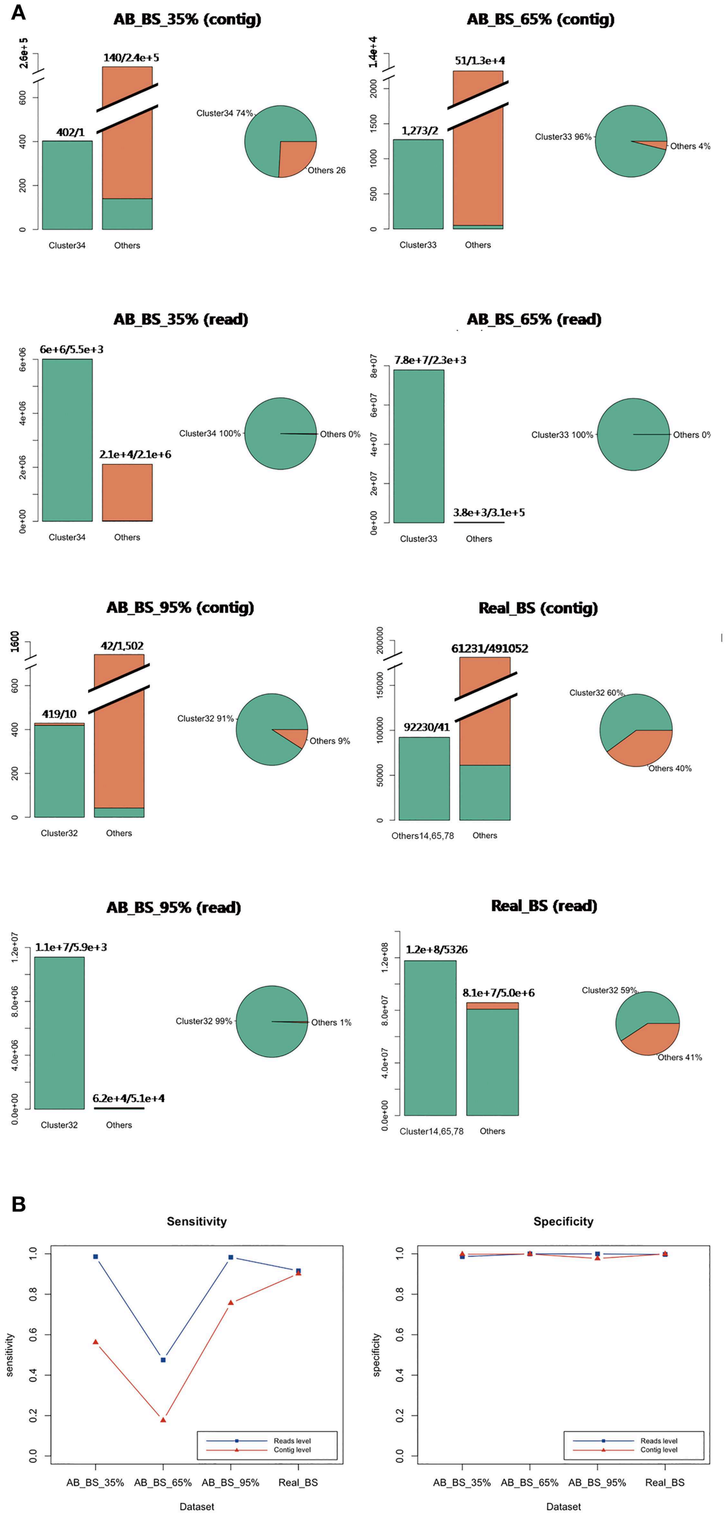

FIGURE 6 | Evaluation of contig binning and contaminations removal for real data. **(A)** Target distribution and concentration in four samples. In the bar charts, clusters with the largest number of target contigs were shown, combined with all other clusters that has at least one target contig. The green bar represents target species, while the orange bar represents other species. Distributions of all target contigs were also shown in pies. The green portion stands for the cluster with the largest number of target contigs, while the orange portion stands for other clusters containing target contigs. The clusters are marked by original cluster ID from QC-Blind pipeline (e.g., Cluster34). **(B)** Sensitivity and specificity of the results of the three groups at MEGAHIT-BLAST. The blue lines with squares stand for read level results and the red lines with triangles stand for contig level results.

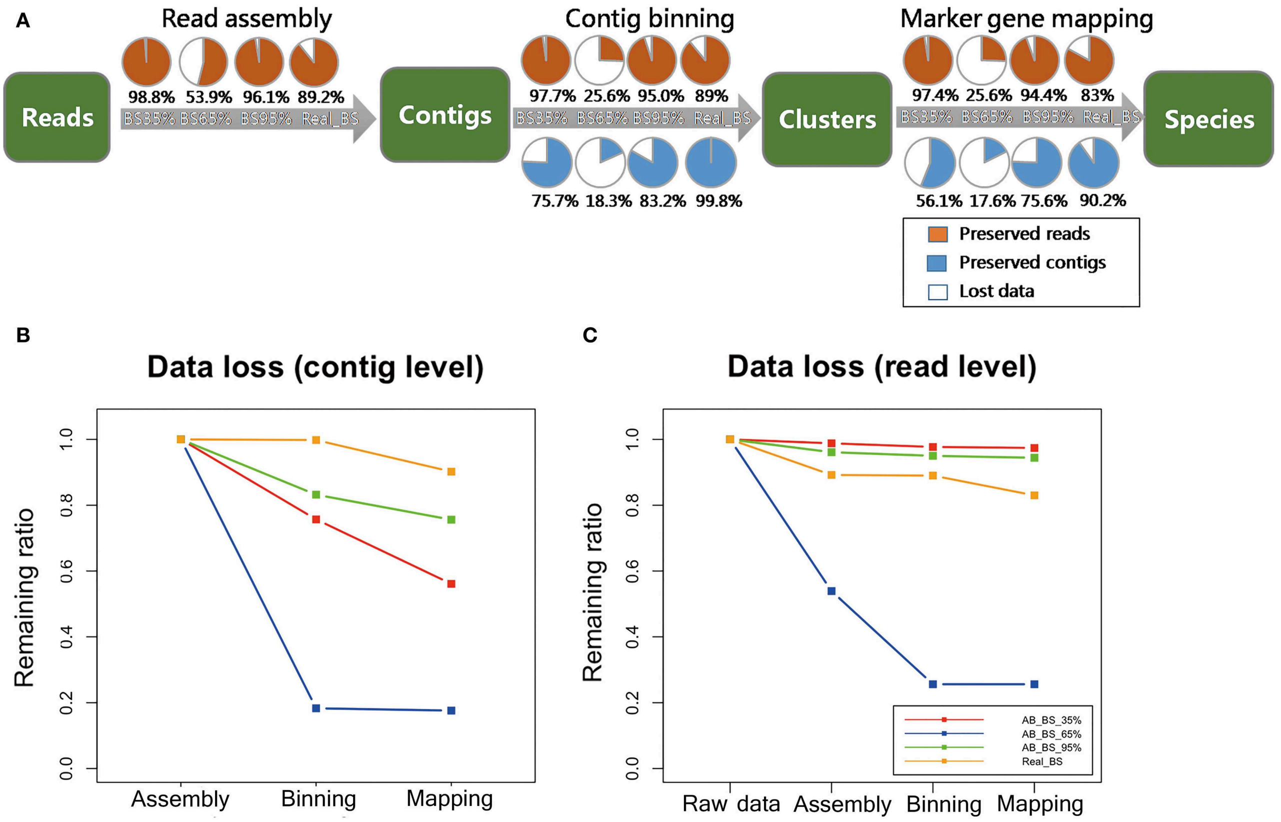

FIGURE 7 | Data loss assessment for real data. **(A)** Proportion of target contigs and reads preserved in each step (read assembly, contig binning, marker gene mapping), for MEGAHIT pipeline. The four circles in a row represent groups in which the frequency of target species increases from 35, 65 to 95%, as well as real BS sequencing data. The orange portions represent preserved reads, the blue portions represent preserved contigs, while the white portions represent lost data. **(B)** Data loss in the steps of read assembly, contig binning, and marker gene mapping of MEGAHIT-BLAST pipeline at contig level. The red, blue, green, and yellow lines represent the samples in which the frequency of target species increases from 35, 65 to 95%, as well as real BS sequencing data. **(C)** Data loss of MEGAHIT-BLAST pipeline at read level. All results were calculated based on simulated dataset A.

assembly (**Figures 7B,C**). For possible explanation for this abnormal phenomenon, we found that the N50 and average contig length of AB_BS 65% were the lowest among the three (**Table 2**), a large number of its contigs were filtered due to the 600 bp cutoff of CONCOCT. This stringent cutoff was set to remove low quality reads and keep specificity at high level, which if set lower, might recover the data loss (**Supplementary File 3**). No more than 6% of reads were lost in marker gene mapping.

For base- and gene-level coverage, the performance of QC-Blind on real datasets was consistent with that on simulated datasets, as the analytical results of AB_BS 35%, AB_BS 95% and Real_BS datasets all show coverage of over 98%, except for AB_ BS 65% (**Figure 8A**). The results reaffirmed the potential of QC-Blind to reconstruct genome from real sequencing data with contaminations (**Figures 8B,C**).

Comparison With Existing QC Tools

There is no existing reference-free QC tool similar to QC-Blind up till now, so it is difficult to conduct a completely fair performance comparison with QC-Blind. Here we briefly describe three reference-based methods, followed by a simple comparison study.

Firstly, Kontaminant is a k-mer-based contamination screening tool (Ramirez-Gonzalez, 2013), which has proved to be effective in host filtering for novel viral discovery. Compared to QC-Blind, Kontaminant is limited by the completeness of existing k-mer database, which lead to its inability to work on unknown contaminations. Another general QC analytical tool is FastQC, which is able to generate a comprehensive report on the quality profile of the reads (Andrews, 2010; Brown et al., 2017). With respect to contaminant identification, it could detect the overrepresented sequence through assessing per base sequence content, per base GC content and per sequence GC content, which can be used as evidence that the library is contaminated (Andrews, 2010). Although FastQC is able to detect unknown contaminations, it cannot remove contaminants well, especially in situations where there are multiple contaminating species. Compared to FastQC, QC-blind implements species-based contigs binning method (ideally one cluster is from one species), which performs better in the case that the contaminant is a mixture of multiple species. The third reference-based method is QC-Chain (Zhou et al.,

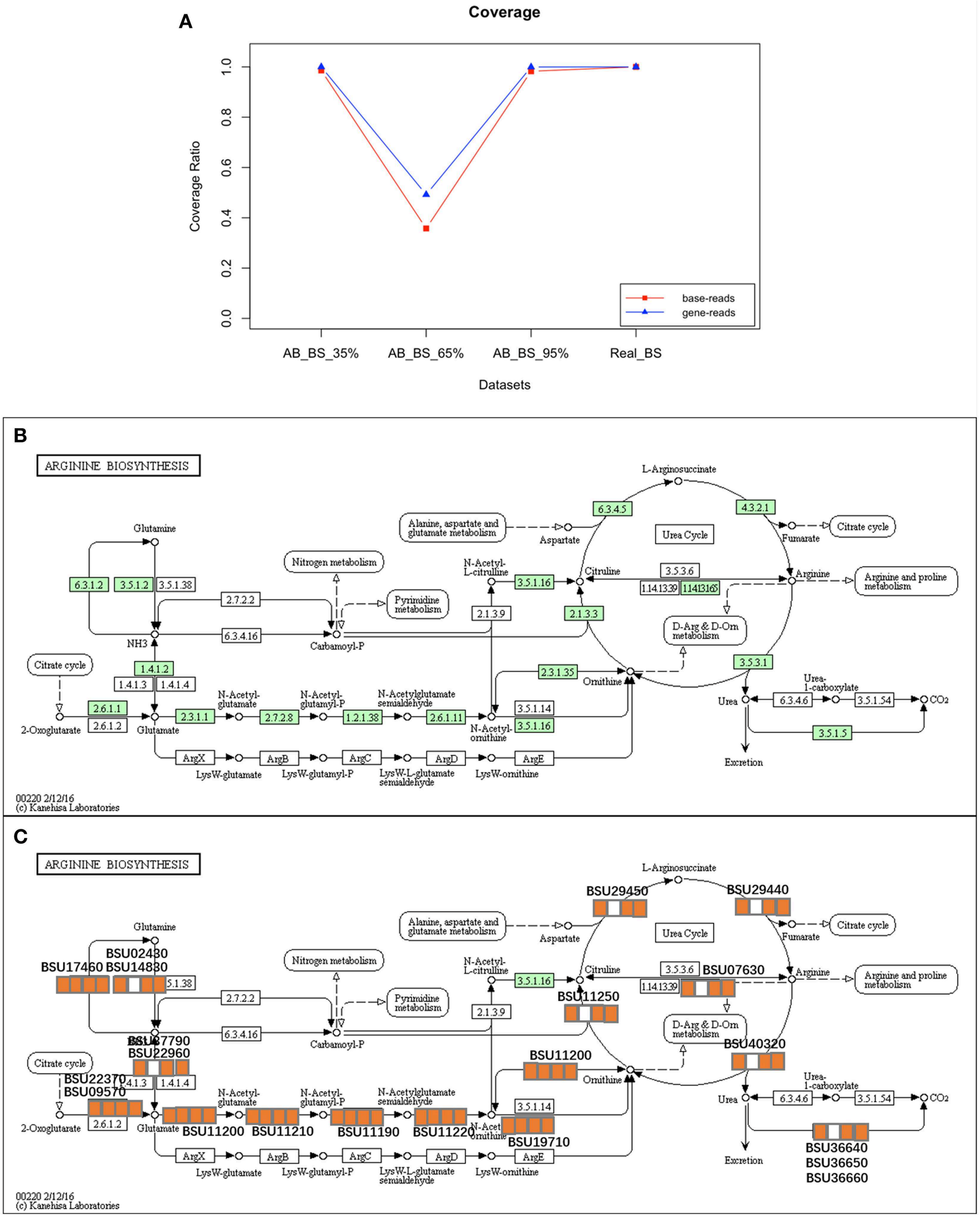

FIGURE 8 | Coverage and pathway reconstruction for real data. **(A)** Y-axis represents base coverage (red line, squares) and gene coverage (blue line, triangles) of target genome as described in Materials and Methods section of this paper. X-axis stands for four samples with proportion of target species from 5 to 95%. **(B)** Arginine biosynthesis pathway of *B. subtilis* from KEGG. **(C)** Genes successfully found in processed data marked with orange squares, and three squares stands for each sample, in which orange and white stands for exist and non-exist, respectively.

2013), which has good performance in identifying and removing contaminations. However, the false positive rate remains high. Compared to this method, the purity of target clusters in QC-Blind is much higher, and marker genes can map the target clusters accurately, which could ensure higher true positive rate.

As it is still valuable to compare QC-Blind with commonly used reference-based QC tools in terms of contaminant identification and filtration, while the above-mentioned reference-based QC tools are all based on certain assumptions, here we have conducted a simple yet direct comparison of QC-Blind with reference-based Bowtie2 method (https://sourceforge.

net/projects/bowtie-bio/files/bowtie2/). Since Bowtie2 is simply a set of mapping tools, it is obvious that if all reference genomes are provided, Bowtie2 would yield the best specificity for target identification and read assignment.

For a direct comparison with reference-based method, we have utilized reference genome for strain *B. subtilis* 168, as well as the most abundant oral microbes (HOB10, as detailed in **Table 1**) for the assessment. The Bowtie2 mapping tools with default parameters was used. Results have shown that based on simulated dataset A (**Table 1**), when using reference genome for strain *B. subtilis* 168, almost perfect quantities of reads could be assigned to target successfully, which is not surprising. However, when using HOB10 as reference genome for contaminations, 0.3% of error could still occur. On *in vivo* datasets, it was observed that 88.6 ± 1.5% read-level assignment specificity could be achieved by using reference genome for strain *B. subtilis* 168. Such results on *in vivo* datasets are close to the results of QC-Blind (**Figure 6**). Therefore, it is still true that reference-based method is superior to reference-free method in specificity. However, the gap between the reference-based method and the reference-free method is much closer when QC-Blind is adopted, especially on *in vivo* sequencing data.

Efficiency Evaluation

On a computational server with Intel Xeon CPU E7-4820 v2 (each core 2.00 GHz, 8 core, 16 processors—we used only 1 processor) and 512GB RAM, QC-Blind ran in <12 h on datasets with 40 million paired-end reads, with varying time depending on the sequencing quality and contig number. The greatest proportion of time was consumed on contig assembly and contig binning. Due to the time complexity of the clustering algorithm, the running time increased significantly when working with larger number of contigs, which result from the use of lower contig length cutoff values. Thus, a reasonable cutoff, an improved clustering algorithm, as well as the utilization of multiple processors could be taken into consideration to reach higher efficiency.

CONCLUSION

In this study, we proposed the QC-Blind pipeline for assigning reads to target species. QC-Blind first uses the well-established 16S rRNA approach for species identification. It then employs read assembly (Velvet or MEGAHIT) and contig binning (CONCOCT) sequentially to cluster contigs and reads. Finally, it uses marker genes to identify contig clusters.

A systematic assessment of QC-Blind based on *in silico*, *ab initio*, and *in vivo* datasets with different fraction of reads from contaminations, showed QC-Blind performs well on screening microbial contamination. After being processed by QC-Blind, contigs and reads were highly concentrated in target clusters and were easily identified through their marker genes. The clusters are almost homogeneous even when samples are contaminated by more than 50% heterogeneous reads. The reads which QC-Blind recovered consists of a large proportion of the genome of the target species. However, the performance of QC-Blind on simulation datasets, *ab initio* mixture of species and *in vivo* real datasets showed certain differences: data loss was more significant in a few real datasets such as BS 65%.

Our tests show that QC-Blind is able to acquire high-quality sequencing data, as well as reduce the widely presented "batch effects" (Weiss et al., 2014) caused by experimental procedure and human factors (exemplified with human saliva in this study). We have also shown that reference-free methods work well, and such performance is not very much dependent on the choice of assembler. However, it is very hard to completely separate reads from target and contamination even when using reference-based method, as exemplified by our comparison study, as identical sequence fragment usually occur, between target, and contamination species. Having this in mind, it is clear that with the application of QC-Blind, the performance gap between reference-based and reference-free approaches further shrink. Additionally, unlike traditional reference-based method that highly depend on reference genome, QC-Blind could accurately identify and filter sequencing reads from target species utilizing only a small number of marker genes (Leek et al., 2010). Moreover, the selection of marker genes is flexible and context-dependent, thus providing a lot of room for improvement during actual application.

As a future work, we are considering putting QC-Blind to the task where both target and contamination species are unidentified (and without biomarker genes other than 16S rRNA or a few genes) before sequencing. This problem is equivalent to the problem of metagenomic read binning, in which the reads of both target species and contaminations should be clustered into separate clusters. A more complex situation is with multiple species as target species, and multiple known/unknown contaminant species (but without multiple samples for comparison). Theoretically, contig binning can be directly applied to this multiple-species problem, and there are two points worth noticing: First, to increase the accuracy of marker gene identification, sequences from the same species have to be precisely clustered. Second, the performance of assembly and contig binning method should be stable in different situations, especially when there is a lack of dominant species. All these would be considered in our future explorations on the QC issues of high-throughput sequencing data.

AUTHOR CONTRIBUTIONS

This study was designed by KN. WX and YG designed the simulated data and performed the data analysis. PY and CC collected the real data. MH made the contribution in the data analysis aspect. WX, ZC, GX, and KN wrote the initial draft of the manuscript. All authors revised the manuscript.

FUNDING

This work is partially supported by Ministry of Science and Technology's high-tech grant 2018YFC0910502, and National Science Foundation of China grant 31871334 and 31671374.

ACKNOWLEDGMENTS

We thank Francis Ng for his careful proof-reading.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.01560/full#supplementary-material

Supplementary File 1 | The clustering results and evaluations of simulated datasets of BS_1_5 (contaminated by HOB5), BS_1_10 (contaminated by HOB10), BS+SC_1_10 (contaminated by HOB10), Ecoli_1_5 (contaminated by HOB5), Ecoli_1_10 (contaminated by HOB10), SA_1_5 (contaminated by HOB5), SA_1_10 (contaminated by HOB10).

Supplementary File 2 | (Table S1) Identification Sensitivity of target and contamination species for QC-Blind-processed simulated datasets. **(Table S2)** Assembly Statistics for simulated datasets.

Supplementary File 3 | Realdata_cluster_BS.xlsx.
The clustering results of real datasets of AB_BS when cutoff of CONCOCT was set to 500 bp.
Realdata_evaluation of BS.xlsx.
The evaluations of real datasets of AB_BS when cutoff of CONCOCT was set to 500 bp.

REFERENCES

Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol. Rev.* 59, 143–169.

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Babraham Institute. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 30, 2114–20. doi: 10.1093/bioinformatics/btu170

Brown, J., Pirrung, M., and McCue, L. A. (2017). FQC dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics.* (2017) 33, 3137–3139. doi: 10.1093/bioinformatics/btx373

Chen, S., Huang, T., Zhou, Y., Han, Y., Xu, M., Gu, J., et al. (2017). AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformat.* 18(Suppl 3):80. doi: 10.1186/s12859-017-1469-3

Cheng, X., Chen, X., Su, X., Zhao, H., Han, M., Bo, C., et al. (2014a). DNA extraction protocol for biological ingredient analysis of Liuwei Dihuang Wan. *Genom. Proteom. Bioinformat.* 12, 137–143. doi: 10.1016/j.gpb.2014.03.002

Cheng, X., Su, X., Chen, X., Zhao, H., Bo, C., Xu, J., et al. (2014b). Biological ingredient analysis of traditional Chinese medicine preparation based on high-throughput sequencing: the story for Liuwei Dihuang Wan. *Sci. Rep.* 4, 5147–5147. doi: 10.1038/srep05147

Coleman, C., Quinn, E. M., and McManus, R. (2015). Quality control procedures for high-throughput genetic association studies. *Methods Mol. Biol.* 1326, 203–215. doi: 10.1007/978-1-4939-2839-2_17

Concolino, P., Costella, A., Minucci, A., Scaglione, G. L., Santonocito, C., Salutari, V., et al. (2014). A preliminary Quality Control (QC) for next generation sequencing (NGS) library evaluation turns out to be a very useful tool for a rapid detection of BRCA1/2 deleterious mutations. *Clin. Chim. Acta.* 437, 72–77. doi: 10.1016/j.cca.2014.06.026

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12:499–510. doi: 10.1038/nrg3012

Freedman, Z. B., Upchurch, R. A., Zak, D. R., and Cline, L. C. (2016). Anthropogenic N deposition slows decay by favoring bacterial metabolism: insights from metagenomic analyses. *Front. Microbiol.* 7:259. doi: 10.3389/fmicb.2016.00259

Guo, Y., Ye, F., Sheng, Q., Clark, T., and Samuels, D. C. (2014). Three-stage quality control strategies for DNA re-sequencing data. *Brief Bioinform.* 15, 879–889. doi: 10.1093/bib/bbt069

Hasan, N. A., Young, B. A., Minard-Smith, A. T., Saeed, K., Li, H., Heizer, E. M., et al. (2014). Microbial community profiling of human saliva using shotgun metagenomic sequencing. *PLoS ONE* 9:e97699. doi: 10.1371/journal.pone.0097699

Henderson, G., Cox, F., Kittelmann, S., Miri, V. H., Zethof, M., Noel, S. J., et al. (2013). Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. *PLoS ONE* 8:e74787. doi: 10.1371/journal.pone.0074787

Jia, B., Xuan, L., Cai, K., Hu, Z., Ma, L., and Wei, C. (2013). NeSSM: a next-generation sequencing simulator for metagenomics. *PLoS ONE* 8:e75448. doi: 10.1371/journal.pone.0075448

Jung, J., Philippot, L., and Park, W. (2016). Metagenomic and functional analyses of the consequences of reduction of bacterial diversity on soil functions and bioremediation in diesel-contaminated microcosms. *Sci. Rep.* 6:23012. doi: 10.1038/srep23012

Katta, M. A., Khan, A. W., Doddamani, D., Thudi, M., and Varshney, R. K. (2015). NGS-QCbox and raspberry for parallel, automated and rapid quality control analysis of large-scale next generation sequencing (Illumina) data. *PLoS ONE* 10:e0139868. doi: 10.1371/journal.pone.0139868

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., et al. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* 8, 761–763. doi: 10.1038/nmeth.1650

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., et al. (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* 390, 249–256. doi: 10.1038/36786

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. doi: 10.1038/nrg2825

Leggett, R. M., Ramirez-Gonzalez, R. H., Clavijo, B. J., Waite, D., and Davey, R. P. (2013). Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics. *Front. Genet.* 4:288. doi: 10.3389/fgene.2013.00288

Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31, 1674–1676. doi: 10.1093/bioinformatics/btv033

Lindgreen, S., Adair, K. L., and Gardner, P. P. (2016). An evaluation of the accuracy and speed of metagenome analysis tools. *Sci. Rep.* 6:19233. doi: 10.1038/srep19233

Paszkiewicz, K. H., Farbos, A., O’Neill, P., and Moore, K. (2014). Quality control on the frontier. *Front. Genet.* 5:157. doi: 10.3389/fgene.2014.00157

Porebski, S., Bailey, L. G., and Baum, B. R. (1997). Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Rep.* 15, 8–15. doi: 10.1007/BF02772108

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26, 841–842. doi: 10.1093/bioinformatics/btq033

Ramirez-Gonzalez, R. (2013). *Kontaminant, a k-mer Based Contamination Screening and Filtering Tool*. Available online at: http://www.tgac.ac.uk/kontaminant

Rütti, S., and Widmann, C. (2015). Genetics and molecular biology: HDL plasticity and diversity of functions. *Curr. Opin. Lipidol.* 26, 596–597. doi: 10.1097/MOL.0000000000000242

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z

Schmieder, R., and Edwards, R. (2011). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6:e17288. doi: 10.1371/journal.pone.0017288

Sheng, Q., Vickers, K., Zhao, S., Wang, J., Samuels, D. C., Koues, O., et al. (2017). Multi-perspective quality control of Illumina RNA sequencing data analysis. *Brief Funct. Genom.* 16, 194–204. doi: 10.1093/bfgp/elw035

Simon, C., and Daniel, R. (2011). Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.* 77, 1153–1161. doi: 10.1128/AEM.02345-10

Strong, M. J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., et al. (2014). Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog.* 10:e1004437. doi: 10.1371/journal.ppat.1004437

Strous, M., Kraft, B., Bisdorf, R., and Tegetmeyer, H. E. (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front. Microbiol.* 3:410. doi: 10.3389/fmicb.2012.00410

Su, X., Pan, W., Song, B., Xu, J., and Ning, K. (2014). Parallel-META 2.0: enhanced metagenomic data analysis with functional annotation, high performance computing and advanced visualization. *PLoS ONE* 9:e89323. doi: 10.1371/journal.pone.0089323

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10:1196. doi: 10.1038/nmeth.2693

Tanner, M. A., Goebel, B. M., Dojka, M. A., and Pace, N. R. (1998). Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Appl. Environ. Microbiol.* 64, 3110–3113.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., et al. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* 12, 902–903. doi: 10.1038/nmeth.3589

Vikram, S., Guerrero, L. D., Makhalanyane, T. P., Le, P. T., Seely, M., and Cowan, D. A. (2016). Metagenomic analysis provides insights into functional capacity in a hyperarid desert soil niche community. *Environ. Microbiol.* 18, 1875–1888. doi: 10.1111/1462-2920.13088

Weiss, S., Amir, A., Hyde, E. R., Metcalf, J. L., Song, S. J., and Knight, R. (2014). Tracking down the sources of experimental contamination in microbiome studies. *Genome Biol.* 15:564. doi: 10.1186/s13059-014-0564-2

Yan, X., Luo, X., and Zhao, M. (2016). Metagenomic analysis of microbial community in uranium-contaminated soil. *Appl. Microbiol. Biotechnol*. 100, 299–310. doi: 10.1007/s00253-015-7003-5

Yu, J., Feng, Q., Wong, S. H., Zhang, D., Liang, Q. Y., Qin, Y., et al. (2017). Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* 66, 70–78. doi: 10.1136/gutjnl-2015-309800

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18, 821–829. doi: 10.1101/gr.074492.107

Zhou, Q., Su, X., Wang, A., Xu, J., and Ning, K. (2013). QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PLoS ONE* 8:e60234. doi: 10.1371/journal.pone.0060234

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.