SMS: A Novel Approach for Bacterial Strain Genome Reconstruction in Multiple Samples

Saidi Wang  
University of Central Florida

Minerva Fatimae Ventolero  
University of Central Florida

Haiyan Hu  
University of Central Florida

Xiaoman Li  
xiaoman@mail.ucf.edu  
University of Central Florida  
https://orcid.org/0000-0002-9209-458X

Methodology

Keywords: Bacterial strains, shotgun metagenomic samples, multiple samples

Posted Date: November 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1086425/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
SMS: a novel approach for bacterial strain genome reconstruction in multiple samples

Saidi Wang¹, Minerva Fatimae Ventolero², Haiyan Hu¹* and Xiaoman Li²*

¹Department of Computer Science, University of Central Florida, Orlando, FL, 32816

²Burnett School of Biomedical Science, University of Central Florida, Orlando, FL, 32816

*To whom correspondence should be addressed.

Contact: xiaoman@mail.ucf.edu, haihu@cs.ucf.edu
Abstract

Background

The analysis of the bacterial strains is important for the understanding of drug resistance. Despite the existence of dozens of computational tools for bacterial strain studies, most of them are developed for the limited number of known bacterial strains. Almost all remaining tools are designed to analyze individual samples or local regions instead of the entire bacterial genomes. With multiple shotgun metagenomic samples routinely generated in a project, it is necessary to create methods to reconstruct novel bacterial strain genomes in multiple samples.

Results

We developed a novel computational approach called SMS to de novo reconstruct bacterial Strain genomes from Multiple shotgun sequencing Samples. Tested on 702 simulated and 195 experimental datasets, SMS reliably identified the strain number, strain abundance, and strain polymorphisms. Compared with two existing approaches, SMS showed superior performance in terms of more accurate estimation of the strain number, abundance and variations.

Conclusions

SMS is a useful tool for novel bacterial strain reconstruction in multiple shotgun metagenomic samples.

Keywords

Bacterial strains; shotgun metagenomic samples; multiple samples
Background

Bacteria are ubiquitous in our environments and human body sites [1-3]. Because their genomes are constantly evolving, it is common that multiple strains of the same species coexist in an environmental niche. These strain genomes of the same species are different from each other, with small variations such as single nucleotide polymorphisms (SNPs), different gene contents, and/or different plasmid genes [4]. Such a difference results in different strengths and fitness to survive in an environment or react to a treatment, which is often the cause of drug resistance, mixed infection, reinfection, etc. [5, 6]. It is thus important to study bacterial strains and reconstruct their strain genomes.

Shotgun metagenomic sequencing is routinely used to study bacterial strains [7, 8]. In shotgun metagenomic sequencing, short DNA segments from different strains of all species in an environment are mixed and sequenced. These sequenced DNA segments called reads are then employed to determine the strains present and their abundance. By mapping reads to the marker genes, the reference genome(s), or the metagenome-assembled-genome(s) of a species, the variations buried in the mapped reads are identified, and their patterns are used to infer the presence and abundance of different strains [8, 9].

Dozens of computational methods are available to infer bacterial strains from shotgun metagenomic reads [7-22]. Most of them rely on prior knowledge of known strains of a species, such as the known gene contents of a strain or the SNPs in known strains. These methods have successfully identified known strains in a number of metagenomic shotgun samples. However, new strains and new SNPs accumulated in existing strains cannot be studied by these methods. A handful of methods that do not depend on prior strain information are thus developed, which can be further divided into two groups [8, 9, 16-18, 23]. One group defines strain variations and strains based on species-specific marker genes, which depend on the quality and quantity of the marker genes but can significantly speed up the
process to analyze a large number of species in a microbiome [17, 23]. The other group of methods consider the SNPs across the entire reference genomes of an individual species instead of only the marker gene regions, which can delineate the strain genomes of a species in detail and are important for the study of individual pathogen species [9, 12-14]. The two groups of methods have shed new light on bacterial strains in environmental samples. However, their performance is still suboptimal in terms of the predicted strain number and abundance accuracy. For instance, a recent method, StrainFinder, did not have good accuracy in predicting strain SNPs and strain abundance, even provided with the correct strain number in the sample(s) [12].

To accurately identify strains in environmental or clinical samples, we developed a novel strain identification method called SMS (bacterial Strain identification in Multiple metagenomic Samples). Starting from a reference genome, SMS de novo reconstructs bacterial strain genomes from shotgun reads in multiple samples, which thus does not rely on the prior knowledge of known strains. Unlike existing methods for novel strain identification, SMS models the coverage of every strain in individual samples by zero-inflated Poisson (ZIP) distributions and classifies SNPs with adatively inferred centers, which enables it to work well with low coverage of strains and predict strains with high accuracy. Tested on 702 simulated datasets and 195 experimental datasets, SMS accurately predicted the strain number, SNPs, and abundance. Compared with two recent approaches, SMS showed much better performance. The SMS tool is freely available at https://github.com/UCF-Li-Lab/SMS.

**Materials and methods**

SMS reconstructs bacterial strain genomes with a reference genome and raw reads in multiple shotgun sequencing samples (Figure 1). The basic assumption is that different SNPs from the same strain follow a common ZIP distribution in a sample, and SNPs from different strains follow different ZIP distributions in individual samples. Assume there are $R$ strains of a species of interest in $m$ samples.
Starting from the cleaned raw reads, SMS defines SNPs based on the reads mapped to the reference genome. It then determines the initial strains and their abundance with the pooled sample, the combined m samples. Next, SMS refines the initial strains and their abundance based on the SNP coverage patterns across samples. The rationale is that unique SNPs from the same strain will have more similar coverage patterns across samples than SNPs from different strains. Finally, SMS outputs the predicted strains and their abundance. The details are in the following sections.

**Figure 1. The SMS workflow.**

**Identification of potential SNPs**

With the reads from M samples, SMS trims reads and filters low-quality reads with the tool trimmomatic. SMS then maps the cleaned reads to the reference genome by bowtie2 [24]. In every sample, SMS obtains a 4 by n sample-specific matrix composed of the frequencies of A, C, G, and T in the mapped reads at each of the n reference genome positions. Similarly, SMS acquires a pooled matrix of 4 by n for the pooled sample, the sum of the m sample-specific matrices. SMS then determines the n' potential polymorphic positions based on these m+1 matrices. A reference genome position is selected as a potential polymorphic position if the following criteria are satisfied: 1). It has to have a coverage larger than 10% of the pooled coverage of the reference genome. The coverage of a genome (position or SNP) is calculated as the average number of reads mapped to this genome (position or SNP); 2). It has at least two nucleotides with coverage no smaller than 5% of the pooled coverage. Note that when the reference nucleotide at a position has fewer than 5% of the genome coverage, the reference nucleotide is replaced with the most frequent nucleotide at this position; 3). Each of its two most frequent nucleotides must occur in at least 5% of the m samples. Finally, SMS considers all n1
nucleotides with coverage larger than 5% of the genome coverage at these positions as potential SNPs, where $n' \leq n_1 \leq 3n'$.

**Prediction of the strain number and abundance**

With the $n_1$ potential SNPs, SMS infers the strain numbers and their abundance in four steps.

First, SMS obtains an initial number of strains and their SNPs. SMS applies mixtureS to the above $n_1$ SNPs with the pooled sample and outputs the predicted strains and their abundance. MixtureS reconstructs the strain genomes from shotgun reads in one sample and has shown good performance previously [12]. In this way, the strains with different pooled coverage are separated into $R$ strains, where $R$ is automatically inferred.

Second, SMS refines the predicted strains so that almost all SNPs in an original strain are assigned to one predicted strain. Since the coverage of SNPs from the same strain are expected to follow the same ZIP distributions in individual samples, the coverage vectors of two SNPs from the same strain are more similar than those of two SNPs from different strains. Here the coverage vector of a SNP is a vector composed of its coverage in the m samples. The similarity measurement of two vectors is described in the next section. Based on this observation, SMS iteratively regroups the $n_1$ SNPs into $R$ groups so that SNPs from the same group have more similar vectors. Starting from the predicted $R$ strains by mixtureS, the majority of SNPs in each of which are likely from the same strain, SMS represents each strain by a $m$ by 1 coverage vector, the average of the coverage vectors of the SNPs currently assigned to this strain. SMS then re-assigns each of the $n_1$ SNPs to the strain with the most similar coverage vector to the coverage vector of this SNP. With the re-assigned SNPs, the coverage vectors of the strains are recalculated. This process is repeated a given number of times or until the assigned SNPs to each strain do not change. In this way, the coverage vector of each predicted strain and the assignment of the $n_1$ SNPs become more and more accurate, with almost all SNPs from an
unknown strain grouped together.

Third, SMS investigates whether there are more or fewer than \( R \) strains in the \( m \) samples. SMS divides each strain into two strains, one strain at a time. To determine whether a strain should be divided, SMS models each strain in a sample by a ZIP distribution, estimates the parameters of the ZIP distributions, and calculates the likelihood ratio of observing the SNPs in this strain across the \( m \) samples to that in two divided strains. The details of the ZIP parameter estimation and the likelihood testing are in the following sections. A strain is divided only when its division significantly increases the likelihood (Chi-square test p-value<0.001). If a strain is divided, SMS considers whether the two new divided strains can be further divided similarly. This process is repeated until no strain can be further divided. With all possible divisions that significantly increase the likelihood, SMS obtains the updated \( R \) strains and repeats Step two to reassign the \( n_l \) SNPs to these \( R \) strains again. SMS then considers removing each strain, one strain at a time. The process is similar to dividing a strain based on the ZIP parameter estimation and the likelihood test.

Finally, SMS removes the predicted strains that are majorly composed of shared SNPs by multiple strains and reassigns their SNPs to the corresponding strains. To remove a strain, SMS identifies its consistent strains. Strain one is a consistent strain of strain two if every entry in the coverage vector of strain one is no large than the corresponding entry in the coverage vector of strain two plus a small cutoff, say 0.5. Similarly, multiple strains together are consistent with strain two if the sum of the corresponding entries in their coverage vectors is no large than the corresponding entry in the coverage vector of strain two plus the same cutoff. With the consistent strains of a strain, SMS constructs a graph, with each consistent strain as a node and edges connecting pairs of strains that are together still consistent with this strain. SMS then identifies the largest cliques in this graph with the corresponding groups of strains together consistent with this strain. With a clique identified, which could be a pair of strains, SMS removes this strain and reassigns the SNPs in this strain to all consistent strains identified
in this largest clique. In this way, SMS finalizes the predicted strains and their SNPs. The abundance of every strain is calculated as the average coverage of the SNPs unique to this strain.

**The similarity of two coverage vectors**

SMS calculates the similarity of two coverage vectors \((a_1, a_2, ..., a_m)\) and \((b_1, b_2, ..., b_m)\) by a pre-defined regression formula: \(79.25d + 43.06(c+c^3) - 0.04/(0.0025+d)\), where \(d\) is the distance between the two vectors, and \(c\) is the Kendall rank correlation of the two vectors. This formula was constructed based on a set of 18 pre-simulated training datasets. SMS chooses this similarity measurement, because it shows better performance than others, including correlation, Euclidian distance, relative entropy, etc. We also tried the following formula:

\[
\sum_{i=1}^{m} \left[ \frac{dpois(b_i, a_i + 1e-5)}{dpois(a_i, a_i + 1e-5) + dpois(a_i, b_i + 1e-5)} \right]^{2.0}
\]

where \(dpois\) is the Poisson density function in the \(R\) package. This formula works similarly to the above regression formula when the sample number is larger than 15.

**ZIP model of a strain in a sample**

SMS models the coverage of the sites from the \(p\)-th strain in the \(q\)-th sample by a ZIP distribution

\[
ZIP(x, \pi_{pq}, \lambda_{pq})
\]

when the \(p\)-th strain occurs in the \(q\)-th sample, where

\[
ZIP(x, \pi, \lambda) = \begin{cases} 
\pi + (1 - \pi) \times \exp(-\lambda), & \text{for } x = 0 \\
\frac{(1-\pi)\times \lambda^x}{x!} \times \exp(-\lambda), & \text{for } x = 1, 2, 3, ...
\end{cases}
\]

Assume we have an \(n1 \times m\) matrix, \(X = (x_{ij})\), which store the coverage of the above \(n1\) SNPs in the \(m\) samples. Assume \(Z = (z_{ir})\) is the indicator to tell whether the \(i\)-th SNP belongs to the \(r\)-th strain, where \(\sum_{r=1}^{R} z_{ir} = 1\) for all \(i\) from 1 to \(n1\) and \(z_{ir}\) can be only 0 or 1. Assume \(Y = (y_{jr})\) is the indicator to show whether the \(r\)-th strain occurs in the \(j\)-th sample, where \(y_{jr}\) can be only 0 or 1. If at least one SNP from a strain has a non-zero coverage in a sample, we tentatively claim that this strain occurs in
When \( y_{jr} = 1 \), we also define \( b_{jr} = \sum_{i=1}^{n_1} z_{ir} x_{ij} = 0 \), \( n_{jr} = \sum_{i=1}^{n_1} z_{ir} \), and \( a_{jr} = \sum_{i=1}^{n_1} z_{ir} x_{ij} / n_{jr} \).

To estimate the parameters in the ZIP, for a given strain that occurs in a given sample, say the \( r \)-th strain in the \( j \)-th sample (i.e., \( y_{jr} = 1 \)), SMS initializes \( \lambda_{jr} = \frac{s_{jr}^2 + a_{jr}^2}{a_{jr}} - 1 \), \( \pi_{jr} = \frac{s_{jr}^2 - a_{jr}}{s_{jr}^2 + a_{jr}^2 - a_{jr}} \), with \( s_{jr}^2 = \frac{\sum_{i=1}^{n_1} z_{ir}(x_{ij} - a_{jr})^2}{\sum_{i=1}^{n_1} z_{ir} - 1} \). SMS then uses the following iteration method to obtain the maximal likelihood estimation of \( \pi_{jr} \) and \( \lambda_{jr} \): first replaces \( \pi_{jr} \) by \( \pi_{jr} = \frac{n_{jr}(\lambda_{jr} - a_{jr})e^{-\lambda_{jr}}}{\lambda_{jr}^2 - n_{jr}(\lambda_{jr} - a_{jr})(1 - e^{-\lambda_{jr}})} \) in the equation \( \frac{n_{jr}a_{jr}}{\lambda_{jr}} - \frac{(1 - \pi_{jr})b_{jr}}{\pi_{jr} + (1 - \pi_{jr})e^{-\lambda_{jr}}} = 0 \) to obtain an equation of \( \lambda_{jr} \), then solves this equation by the Newton’s iteration method. Everywhere in this process, if \( \pi_{jr} = 0 \), you will directly estimate \( \lambda_{jr} = a_{jr} \).

**Log likelihood test**

Given \( R \) strains, the full likelihood of observation the frequencies of these \( nI \) SNPs in the \( m \) samples is

\[
L(X,Z|\pi,\lambda) = \prod_{i=1}^{n_1} \prod_{j=1}^{m} \prod_{r=1}^{R} z_{ir} y_{jr} ZIP(x_{ij}, \pi_{jr}, \lambda_{jr}).
\]

When SMS splits one strain into two or removes one strain, the likelihood can be similarly calculated.

To assess the significance of changing the current \( R \) strains, we calculate the ratio of the likelihood after changing (split or remove) to the likelihood before changing. The ratio approximately follows a Chi-square distribution with the degree of freedom equal to the difference of the parameters in the two models. If the Chi-square test p-value is smaller than a pre-defined cutoff, SMS correspondingly modifies the current \( R \) strains.

**Simulated and experimental datasets**

We simulated 702 datasets (Supplementary Table S1). In each dataset, a reference genome was chosen,
2 to 4 strains were simulated, and 5 to 35 samples were generated. We considered three different reference genomes (Bartonella clarridgeiae NC_014932, Enterococcus casseli flavus NC_020995, and Methanobrevibacter smithii NC_009515), which were randomly chosen from GenBank. For each reference genome, their four strains were generated by randomly choosing 0.01% of the genome positions and then randomly substituted the reference nucleotide with another nucleotide. The read coverage of a reference genome in a dataset was one of the following five coverage, 50x, 100x, 150x, 200x, and 300x. The number of strains and their relative abundance in a dataset were specified by one of the following five configurations: 10:20:30:40, 10:25:25:40, 10:30:60, 15:30:55, and 30:70. For instance, the configuration 10:20:30:40 defined that there were four strains with the relative abundance as 10%, 20%, 30% and 40%, respectively. For a dataset, with the chosen configuration and the number of samples, a subset of samples were randomly chosen for each strain and the coverage of this strain in one of the samples was then randomly determined so that the pooled coverage of this strain was the same as what was specified in the configuration. With the coverage of strains in a sample, paired reads of 100 base pairs long were randomly generated using the tool dwgsim (https://github.com/nh13/DWGSIM).

We tested SMS on 195 experimental datasets [6]. Each dataset is known to have two Mycobacterium tuberculosis strains with predicted abundance. The abundance is inferred from two different computational methods. The actual SNPs in each strain are unknown.

**Comparison with existing methods**

We compared SMS with mixtureS and StrainFinder in a desktop computer with the Intel Core i9-9900KF CPU (16 cores@3.6GHz) and 32 gigabytes memory. We used the following commands to run the three tools respectively:

```
SMS: python SMS/running.py --output_name %s --genome_len %s --genome_name %s --
```
genome_file_loc %s --bam_loc_file %s --res_dir %s

MixtureS: python mixtureS/mixture_model.py --sample_name %s --genome_len %s --genome_name %s --genome_file_loc %s --bam_file %s --res_dir %s

StrainFinder: python StrainFinder/StrainFinder.py --aln %s -N %s --max_reps 10 --dtol 1 --ntol 2 --max_time 3600 --coverage --em_out %s --out_out %s --log %s --n_keep %s --force_update --merge_out --msg

Results

**SMS correctly predicted the strain numbers**

We studied the number of strains predicted in 702 simulated datasets (Material and Methods, Table S1). There were 5 to 35 samples and 2 to 4 strains in every dataset, with the pooled coverage of strains from 100X to 300X. The pooled coverage of a species genome was the sum of the coverage of all strains of this species in all samples. The number of strains and their relative abundance are specified by one of the following five configurations in each dataset: 10:20:30:40, 10:25:25:40; 10:30:60, 15:30:55, and 30:70. For instance, for a dataset with a configuration of 10:20:30:40 and a pooled coverage of 200X, the first, second, third and fourth strain had a coverage of 20X, 40X, 60X and 80X, respectively. In other words, the proportion of reads from the four strains was 10%, 20%, 30% and 40%, respectively, in this dataset.

SMS predicts an initial strain number from the pooled sample and then iteratively determines the final strain number based on all individual samples in a dataset (Material and Methods). SMS correctly predicted the strain number in 282 datasets with only the pooled sample at the first step. With the refinement based on the information from individual samples, the strain number became correct in 415 additional datasets. Note that the correct strain number was changed after the first step and became
correct again in 93 datasets. The significant improvement at the refinement steps thus suggests that it is important to consider the coverage variation across individual samples to define strains.

Overall, SMS predicted the correct strain numbers in all but five datasets (Tables S2-S5). Interestingly, SMS did not predict the correct strain number in at least one dataset for each of the three randomly selected genomes, implying that its performance was not biased towards a specific genome/species. In each of the five datasets, a pair of strains shared 30% of their SNPs. In four of the five datasets, three strains were sharing 20% of their SNPs. These shared SNPs may have confused SMS when the coverage was 100X. As expected, when the coverage was increased from 100X to 200X or 300X, SMS predicted the correct strain number in each of the five corresponding datasets. These analyses on the five datasets suggested that SMS can accurately predict the strain number, even when the pooled coverage was 100X and there were only five samples in a dataset. Moreover, the predicted strain number was even more accurate with a larger pooled coverage (200X coverage for perfect prediction here).

**SMS reliably estimated the strain abundance**

We investigated how well SMS predicted the strain abundance in the above 702 simulated datasets. No matter whether the strain number was correctly predicted, the predicted strain abundance agreed well with the known strain abundance (Figure 2, Tables S2-S5). This agreement did not depend on the sample number, the total coverage, the strain number, whether the strains shared a portion of their SNPs, etc.

**Figure 2. The predicted strain abundance. A) Unshared datasets; B) Shared datasets; and C) All datasets.** MAE refers to the average Maximal Absolute Difference of the predicted abundance and the corresponding true abundance.
In the 697 datasets SMS correctly predicted the strain number, the predicted strain abundance agreed well with the known strain abundance. In these datasets, the predicted strain abundance was within 97.31% of the true abundance. The mean and median ratio of the predicted abundance to the true abundance in these datasets were 0.99 and 1.00, respectively. Even in the five datasets SMS did not correctly predict the strain number, the predicted strain abundance was similar to the true strain abundance. For instance, SMS predicted four strains in three datasets with three strains (Table S5). In two of these three datasets, two strains had the predicted abundance about 0.08 and 0.29, respectively, which were close to the corresponding true abundance 0.10 and 0.30. The two remaining predicted abundance were about 0.42 and 0.21, which differed from the third true abundance, 0.60. In the third dataset, one strain was predicted with an abundance of 0.31, close to the true abundance of 0.30. The wrong prediction of the strain number and strain abundance was likely due to the third strain’s uneven and relatively limited coverage. After increasing the coverage (200X or 300X), SMS predicted the correct strain number and more similar abundance (Table S5). The agreement between the predicted and true abundance suggested that SMS predicted the strain abundance well.

The accuracy was in general improved with more samples in a dataset (Figure 2). When the sample number was larger, the median of the predicted abundance was closer to the true abundance. Moreover, the variation of the difference between the predicted abundance and the true abundance was smaller. This improvement was especially evident in datasets with shared SNPs among strains. In datasets with no shared SNPs among strains, the same trend of improvement was still clear. However, the predicted abundance was not necessarily closer to the true abundance when the sample number was larger, which was likely because the coverage of a strain in individual samples became smaller when the sample number was larger.

We also studied how the accuracy of the predicted abundance was affected by the pooled coverage in a dataset. The accuracy was measured by the Maximal Absolute Difference (MAE) of the predicted
abundance and the corresponding true abundance. The prediction in general became more accurate with a larger pooled coverage (Figure 2). The median MAE was similar for different pooled coverage. However, the variation of the MAE became much smaller when the pooled coverage became 200X or larger. We thus would recommend that the pooled coverage should be at least 200X. The pooled coverage may need to be larger if it is expected to have more strains.

In addition, we examined how different bacterial genomes or the number of strains in a dataset would affect the predicted strain abundance. There was a slight variation when different genomes and/or different numbers of strains were considered (Figure 2). For instance, the difference between the predicted and true strain abundance was within a similar range and with a similar mean/median when there were different numbers of strains. The small variations suggested that the predicted abundance by SMS was robust to different bacterial genomes, different number of strains, etc.

Strains with similar abundance were challenging to distinguish from each other in previous studies [12, 13]. We thus specifically focused on the datasets with two configurations: 10:20:30:40 and 10:25:25:40, in which pairs of strains had coverage within two folds of each other and were difficult to separate previously. We found this was not the case since SMS predicted the strain abundance within 99% of the true abundance in 80 of the 81 datasets with the configuration 10:20:30:40. SMS also correctly predicted the strain abundance within 99% of the true abundance in 318 of the 324 datasets for the more challenging configuration 10:25:25:40. Moreover, the predicted strain abundance was within 98% of the true abundance in all 324 datasets. These observations indicated that the variation of the strain coverage across samples enabled SMS to separate strains of similar abundance better than previous methods that considered strain coverage in one sample.

Although the shared SNPs by multiple strains may affect the estimated strain abundance, we did not see their effect on the performance of SMS. In 486 of the 702 datasets with up to 30% of SNPs shared by a pair of strains, the predicted strain abundance was still within 97.0% of the true strain abundance.
This is because SMS employs only the SNPs unique to a strain to measure its abundance. Moreover, SMS can distinguish these shared SNPs and correctly reassign them to the original strains (Material and Methods).

**SMS faithfully determined the SNPs**

Existing methods mainly focus on the predicted strain number and only occasionally consider the strain abundance. Rarely do they mention the accuracy of the predicted strain SNPs. With the simulated datasets, we systematically evaluated the predicted SNPs. We found that SMS has a precision of 0.97 and a recall of 0.96 to assign SNPs to strains.

We studied the datasets without shared SNPs among strains (Table S6). In these 216 datasets, on average, SMS had a precision of 0.98 and a recall of 0.98. For a given species with a specified pooled coverage, the precision and recall were higher on datasets with more samples in general. Similarly, the precision and recall were generally higher on datasets with a larger pooled coverage when the species and the sample number were fixed. For instance, for the reference species genome NC_009515.1 and the sample number 20, the precision increased from 0.98 to 0.99 and the recall increased from 0.97 to 0.99 when the pooled coverage increased from 100X to 300X.

Since the shared SNPs by different strains complicated the strain prediction, we studied the predicted strains on datasets with shared SNPs among strains (Tables S7-S9). We again focused on the two most challenging configurations: 10:20:30:40 and 10:25:25:40. The datasets with the configurations were challenging because the shared SNPs among strains may have similar coverage across samples with SNPs unique to other strains. For instance, the shared SNPs between the first two strains in the configuration 10:20:30:40 had a relative abundance of 30%, the same as the relative abundance of the third strain. Even with such complexity, SMS on average had a precision of 0.97 and a recall of 0.96 on all 270 datasets (Tables S7 and S8). The performance was only slightly lower than that on datasets
without shared SNPs, suggesting that SMS could reconstruct the complicated evolutionary trajectories of strains with shotgun sequencing reads.

**SMS performed well on experimental datasets**

We tested SMS on 195 experimental datasets (Table S10). We chose these datasets because their strain numbers were known. The strain abundance was also predicted previously [6]. Note that the Critical Assessment of Metagenome Interpretation challenge provided the experimental datasets to evaluate taxonomical composition analysis and assembly [25]. They did not provide the strain number, strain abundance and SNPs unique to strains, thus not suitable for the strain genome reconstruction here.

SMS identified two strains in each of these 195 datasets, which agreed well with the previous study [6]. This study showed that there were at least 11 heterozygous sites in each of these 195 datasets. Interestingly, SMS showed that the two strains in each of these datasets were the same across these datasets, which was consistent with the fact that these datasets were from clinical samples collected from the same region. Moreover, SMS distinguished strains with similar abundance in these datasets. For instance, in the dataset ERR323056, there were 69 heterozygous sites observed in reads [6]. SMS predicted two strains with a relative abundance of 0.52 and 0.48. The previous study based on the SNP frequency identified only one strain, likely due to the similar abundance of the two strains [6]. In addition, SMS reliably predicted the strain abundance. The difference of the predicted strain abundance to the predicted abundance previously had a mean and median of 0.16 and 0.12 respectively, if we considered only the 186 datasets where the previous study correctly predicted the strain number.

**SMS reconstructed strain genomes better than existing methods**

We compared SMS with mixtureS [12] and StrainFinder [9]. Although several dozen methods are available for bacterial strain analysis, most of them are not for de novo strain identification [4]. mixtureS and StrainFinder showed better performance among the existing methods for novel strain
identifications previously [12]. Since mixtureS works on one sample, we ran it on the pooled sample in each dataset. Because StrainFinder is unable to determine the strain numbers, we specified the known strain numbers in the corresponding datasets.

We compared the strain number, SNPs, and strain abundance predicted by the three methods (Table 1). For simulated datasets with no shared SNPs among strains, SMS predicted the correct strain number in all 216 datasets while mixtureS correctly predicted the strain number in 98 datasets. On average, the predicted SNPs by SMS had a precision of 0.97 and a recall of 0.98, larger than those of mixtureS and SStrainFinder. Moreover, the predicted strain abundance by SMS had an average MAE of 0.004, compared with 0.08 by mixtureS and 0.07 by StrainFinder. The average MAE measured the average difference between the predicted and the known strain abundance. For simulated datasets with shared SNPs among strains, SMS and mixtureS predicted the correct strain numbers in 481 (98.97%) and 184 (37.86%) datasets, respectively. SMS had an average precision of 0.97 and recall of 0.96, which were larger than those of mixtureS and StrainFinder. For the strain abundance, SMS had an MAE of 0.008, compared with 0.07 by mixtureS and 0.06 by StrainFinder. On experimental datasets, SMS also showed superior performance to other methods (Table 1). For instance, SMS correctly predicted two strains in all 195 experimental datasets, while mixtureS predicted two strains in 84 datasets.

Table 1 The performance of the three tools.

We also studied the running time of different methods. We collected the tool running time on nine simulation datasets with the same genome (NC009515.1) and same strain configuration (10:30:60) (Table S11). As mixtureS is applied in the first step of SMS, SMS took a little longer to run than mixtureS. However, the difference was not so evident. For instance, for a dataset with 926,975 100 base pairs long paired reads, it took mixtureS and SMS 28 and 31 seconds to predict strains, while it
took StrainFinder 1212 seconds. For all tools, we observed that the time cost had little relation with the size of the datasets. Instead, the time cost mainly depended on the number of strains and the number of SNPs in strains.

**Discussion**

SMS reconstructs bacterial strain genomes with a species reference genome and the raw sequencing reads. The reference genome is employed to map the cleaned reads. The chosen reference genome thus does not affect the identified strains much, as the SNPs in strains come from the mapped reads. In other words, the reference genome only provides the regions to be considered in the species and does not affect the predicted strain number and abundance. SMS defines SNPs with an in-house procedure, which may affect the quality of individual SNPs. However, we do not think that the potential false SNPs will affect the predicted strain number and abundance, as they are determined by the coverage of the majority of SNPs in individual strains. Users may choose existing tools such as SAMtools [26] to define SNPs in individual samples. In addition, since reads are mapped to the reference genomes to predict bacterial strains, SMS can be applied to general metagenomic datasets instead of the shotgun samples for individual species illustrated here.

SMS is not designed for the strain analysis of novel bacterial species. With more and more sequenced bacterial genomes, this issue will not be of a concern in the future. Moreover, SMS considers only the reference genomic regions to reconstruct bacterial strain genomes. It thus does not consider accessory genes that are not represented in the chosen reference genome. In this sense, what SMS reconstructs is the strain core genomes. In the future, one may develop methods to further discover accessory genes in strains, with the inferred strain number and abundance in samples.

**Conclusions**
SMS reconstructs bacterial strain genomes with multiple shotgun sequencing samples. It considers the coverage variation of individual strains across samples to distinguish strains. As demonstrated in simulated and experimental datasets, SMS is able to separate strains with similar abundance in samples. The capability to separate strains with similar abundance by SMS is in general improved with more samples and larger pooled sequencing coverage.

**Availability and requirements**

Project name: SMS

Project home page: https://github.com/UCF-Li-Lab/SMS

Operating system: Linux

Programming languages: Python 3.6

Other requirements: Conda; Hard drive space and memory are recommended to run the software, dependent on databases and input file sizes used.

License: GPLv3

Restrictions to use by non-academics: No

**Abbreviations**

SNPs: single nucleotide polymorphisms

ZIP: zero-inflated Poisson

MAE: Maximal Absolute Difference

**Acknowledgements**
Funding

This work was supported by the National Science Foundation [1661414, 2015838].

Availability of data and materials

SMS is available at https://github.com/UCF-Li-Lab/SMS. Dwgsim is available at https://github.com/nh13/DWGSIM. The 195 experimental datasets are from EBI European Nucleotide Archive: ERP000436 at https://www.ebi.ac.uk/ena/data/view/PRJEB2358 and ERP0001072 at https://www.ebi.ac.uk/ena/data/view/PRJEB2794. The detail of the simulation data are in Table S1. Reference geneomes are available at NC_014932, NC_020995, and NC_009515.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.
Author information

Affiliations

Department of Computer Science, University of Central Florida, Orlando, FL, 32816

Saidi Wang & Haiyan Hu

Burnett School of Biomedical Science, University of Central Florida, Orlando, FL, 32816

Minerva Fatimae Ventolero & Xiaoman Li

Contributions:

H.H and X.L. conceived the idea. S. W. implemented the idea and generated results. S.W and M.F.V. analysis the result. M.F.V., H.H and X.L. wrote the manuscript. All authors reviewed the manuscript.

Corresponding author:

Corresponding to Xiaoman Li or Haiyan Hu

References

1. Methe, B.A., et al., A framework for human microbiome research. Nature, 2012. 486(7402): p. 215-221.

2. Human Microbiome Project, C., Structure, function and diversity of the healthy human microbiome. Nature, 2012. 486(7402): p. 207-14.

3. Proctor, L.M., et al., The Integrative Human Microbiome Project. Nature, 2019. 569(7758): p. 641-648.

4. Van Rossum, T., et al., Diversity within species: interpreting strains in microbiomes. Nat Rev Microbiol, 2020. 18(9): p. 491-506.
5. Eyre, D.W., et al., Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in Clostridium difficile transmission. PLoS Comput Biol, 2013. 9(5): p. e1003059.

6. Sobkowiak, B., et al., Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. Bmc Genomics, 2018. 19(1): p. 613.

7. Albanese, D. and C. Donati, Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nature Communications, 2017. 8.

8. Luo, C., et al., ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol, 2015. 33(10): p. 1045-52.

9. Smillie, C.S., et al., Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell Host & Microbe, 2018. 23(2): p. 229-+.

10. Anyansi, C., et al., Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data. Front Microbiol, 2020. 11: p. 1925.

11. Hong, C.J., et al., PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome, 2014. 2.

12. Li, X., H. Hu, and X. Li, mixtureS: a novel tool for bacterial strain reconstruction from reads. Bioinformatics, 2020.

13. Li, X., et al., BHap: a novel approach for bacterial haplotype reconstruction. Bioinformatics, 2019. 35(22): p. 4624-4631.

14. Pulido-Tamayo, S., et al., Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. Nucleic Acids Res, 2015. 43(16): p. e105.
15. Ahn, T.H., J.J. Chai, and C.L. Pan, Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics, 2015. 31(2): p. 170-177.

16. Costea, P.I., et al., metaSNV: A tool for metagenomic strain level analysis. PLoS One, 2017. 12(7): p. e0182392.

17. Nayfach, S., et al., An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Research, 2016. 26(11): p. 1612-1625.

18. Quince, C., et al., DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol, 2017. 18(1): p. 181.

19. Roosaare, M., et al., StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees. PeerJ, 2017. 5: p. e3353.

20. Sankar, A., et al., Bayesian identification of bacterial strains from sequencing data. Microb Genom, 2016. 2(8): p. e000075.

21. Scholz, M., et al., Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods, 2016. 13(5): p. 435-8.

22. Tamburini, F.B., et al., Precision identification of diverse bloodstream pathogens in the gut microbiome. Nat Med, 2018. 24(12): p. 1809-1814.

23. Truong, D.T., et al., Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research, 2017. 27(4): p. 626-638.

24. Langmead, B. and S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat Methods, 2012. 9(4): p. 357-9.

25. Sczyrba, A., et al., Critical Assessment of Metagenome Interpretation-a benchmark of
metagenomics software. Nat Methods, 2017. 14(11): p. 1063-1071.

26. Li, H., et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics, 2009. 25(16): p. 2078-9.
Figure 1. The SMS workflow.
Figure 2. The predicted strain abundance. A) Unshared datasets; B) Shared datasets; and C) All datasets. MAE is the average Maximal Absolute Difference of the predicted abundance and the corresponding true abundance.
### Tables

**Table 1** The performance of the three tools.

| Dataset | SMS | mixture | StrainFinder |
|---------|-----|---------|--------------|
|         | # (%) of datasets | # (%) of datasets | # (%) of datasets |
|         | Precision, Recall, F1 | MAE | Precision, Recall, F1 | MAE | Precision, Recall, F1 | MAE |
| 702 simulated datasets | 216 (100%) | 0.97, 0.98, 0.98 | 0.004 | 98 (45.37%) | 0.81, 0.83, 0.80 | 0.08 | 0.66, 0.56, 0.53 | 0.07 |
|         | 481 (98.97%) | 0.97, 0.96, 0.96 | 0.008 | 184 (37.86%) | 0.83, 0.58, 0.63 | 0.07 | 0.68, 0.56, 0.56 | 0.06 |
|         | 697 (99.29%) | 0.97, 0.96, 0.96 | 0.007 | 282 (40.17%) | 0.82, 0.66, 0.68 | 0.07 | 0.68, 0.56, 0.55 | 0.06 |
| 195 experimental datasets | 195 (100%) | NA | 0.16 | 146 | NA | 0.12 | NA | 0.26 |

The three columns for each tool are the number (percentage) of datasets where the tool predicted the correct strain number; the precision, recall and F1 score of the predicted strain SNPs; and the average MAE of the predicted strain abundance.
Additional Files

Additional file1: Supplementary Tables

**Table S1. Simulated datasets.** There are 27 datasets generated in each row, each of which corresponds to one of the three different bacterial species (species 1: *Bartonella clarridgeiae* NC_014932, species 2: *Enterococcus casseli flavus* NC_020995, and species 3: *Methanobrevibacter smithii* NC_009515) and one of the nine different sample numbers (5, 8, 10, 12, 15, 20, 25, 30, 35). The number and order of strains in each dataset are specified by the strain configuration. For instance, the configuration 10:20:30:40 in the first row specifies that there are four strains, with the coverage of the 1st, 2nd, 3rd and 4th strain as 10X, 20X, 30X and 40X, respectively. There is no shared SNPs among strains in Type 1 datasets. In each type 2 dataset, the first two strains share 30% of their SNPs, the first three strains share 20% of their SNPs, and the fourth strain shares no SNP with other strains. In each type 3 dataset, the first two strains share 30% of their SNPs, the first three strains share 10% of their SNPs, and there is no SNP shared between the first three strains and the last strain. In each type 4 dataset, the first two strains share 30% of their SNPs, and the third strain share no SNP with other strains.

**Table S2. Prediction on Type 1 simulated datasets.**

**Table S3. Prediction on Type 2 simulated datasets.**

**Table S4. Prediction on Type 3 simulated datasets.**

**Table S5. Prediction on Type 4 simulated datasets.**

**Table S6. Precision, recall and F1 score of the predicted SNP by three tools on simulated Type 1 datasets**

**Table S7. Precision, recall and F1 score of the predicted SNP by three tools on simulated Type 2 datasets**

**Table S8. Precision, recall and F1 score of the predicted SNP by three tools on simulated Type 3 datasets**
Table S9. Precision, recall and F1 score of the predicted SNP by three tools on simulated Type 4 datasets

Table S10. 195 TB abundance result from three tools.

Table S11. Tool running time on nine simulated datasets.
Supplementary Files

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

- SupplementaryTables.docx