# SMS: a novel approach for bacterial strain analysis in multiple samples

**Saidi Wang[1,*], Minerva Fatimae Ventolero[2,*], Haiyan Hu[1,3,#] and Xiaoman Li[2,#]**

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

tjwangsaidi@knights.ucf.edu, haihu@cs.ucf.edu

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

mventolero@knights.ucf.edu,xiaoman@mail.ucf.edu

[3]Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL, 32816;

haihu@cs.ucf.edu

*Contribute equally.

[#] Correspondence: haihu@cs.ucf.edu, 1-407-882-0134 (HH); xiaoman@mail.ucf.edu, 1-407-823-4811 (XL).

## Abstract

The analysis of the bacterial strains is important for understanding drug resistance. Despite the existence of dozens of computational tools for bacterial strain studies, most of them are for known bacterial strains. Almost all remaining tools are designed to analyze individual samples or local strain regions. With multiple shotgun metagenomic samples routinely generated in a project, it is necessary to create methods to infer novel bacterial strain genomes in multiple samples. To fill this gap, we developed a novel computational approach called SMS to de novo reconstruct bacterial Strain genomes in Multiple Samples. Tested on 702 simulated and 195 experimental datasets, SMS reliably identified the strain number, abundance, and polymorphisms. Compared with two existing approaches, SMS showed superior performance. The SMS source code and tool are available at https://github.com/UCF-Li-Lab/SMS.

## Keywords

bacterial strains; shotgun metagenomics; zero-inflated Poisson; strain genome reconstruction

# 1. Introduction

Bacteria are ubiquitous and play crucial roles in disease progression and human health [1-8]. Multiple strains of a bacterial species usually 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 [9]. Such a difference results in different fitness to survive or react to stimuli, which is often the cause of different host responses, drug resistance, mixed infection, etc. [10, 11]. It is thus important to study and reconstruct bacterial strain genomes.

Shotgun metagenomic sequencing is routinely employed to study microbes and reconstruct bacterial genomes [1, 6, 8, 12-14]. In shotgun metagenomics, the DNA of all species and strains in a clinical or environmental sample is randomly fragmented and sequenced. These sequenced DNA fragments called reads are then applied to infer the present species, their abundance, functionality, etc. Because reads are short and mixed from different species, it is still challenging to study low-abudnant species and strains in shotgun metagenomics [15-18]. Moreover, current assembly methods usually cannot distinguish different strains of the same species, which leaves most studies on taxons no lower than the species level and the strain analysis still at its infancy [18, 19]. On the other hand, with the sequencing cost dramatically decreasing, multiple shotgun metagenomic samples are often available from the same type of environments or clinical setups [20-22]. The multiple samples from the same or similar environmental niche are likely to share bacterial strains and thus provide an unprecedented opportunity to study and reconstruct bacterial strain genomes [20-22].

Dozens of computational methods are available to infer bacterial strains from shotgun metagenomic reads [16, 23-37]. Most of them rely on prior knowledge of known strains, and they have successfully identified known strains while cannot be applied to study new strains that commonly exist. A handful of methods

that do not depend on known strains are thus developed, which can be divided into two groups [27, 29, 31-33, 38, 39]. One group defines strain variations and strains based on species-specific marker genes, which can significantly speed up the process of analyzing a large number of species in a microbiome while depending on the quality and quantity of the marker genes [32, 38]. The other group considers the SNPs across the entire reference genomes of a species instead of only the marker gene regions, which can delineate the strain genomes in detail and are important for studying individual pathogen species [16, 26, 28, 29]. These 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. 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 [26, 40].

To accurately identify strains in shotgun metagenomic samples, we developed a novel method called SMS (Strains in Multiple Samples). Starting from a species genome, SMS de novo reconstructs its strain genomes from shtogun reads in multiple shotgun metagenomic samples. It models the coverage of every strain in individual samples by zero-inflated Poisson (ZIP) distributions and classifies SNPs with adaptively inferred centers, which enables it to identify low-coverage strains and predict strains with high accuracy. Tested on 702 simulated and 195 experimental datasets, SMS accurately predicted the strain number, abundance, and SNPs. Compared with two recent approaches, SMS showed much better performance.

## 2. Material and Methods

SMS reconstructs bacterial strain genomes with a reference genome and raw reads in multiple shotgun metagenomic samples (Fig. 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. Because of the species reference genome, SMS considers only the mixed reads in shotgun metagenomic samples that are mapped to the reference genome. In other words, SMS considers only the reads from one species in the *m* samples, as most reads mapped to the reference genome are likely from the reference species. Considering only one species makes sense because we often have a pathogen of interest and want to study its strains in clinical or environmental samples in practice. With the mapped reads, SMS 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 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.

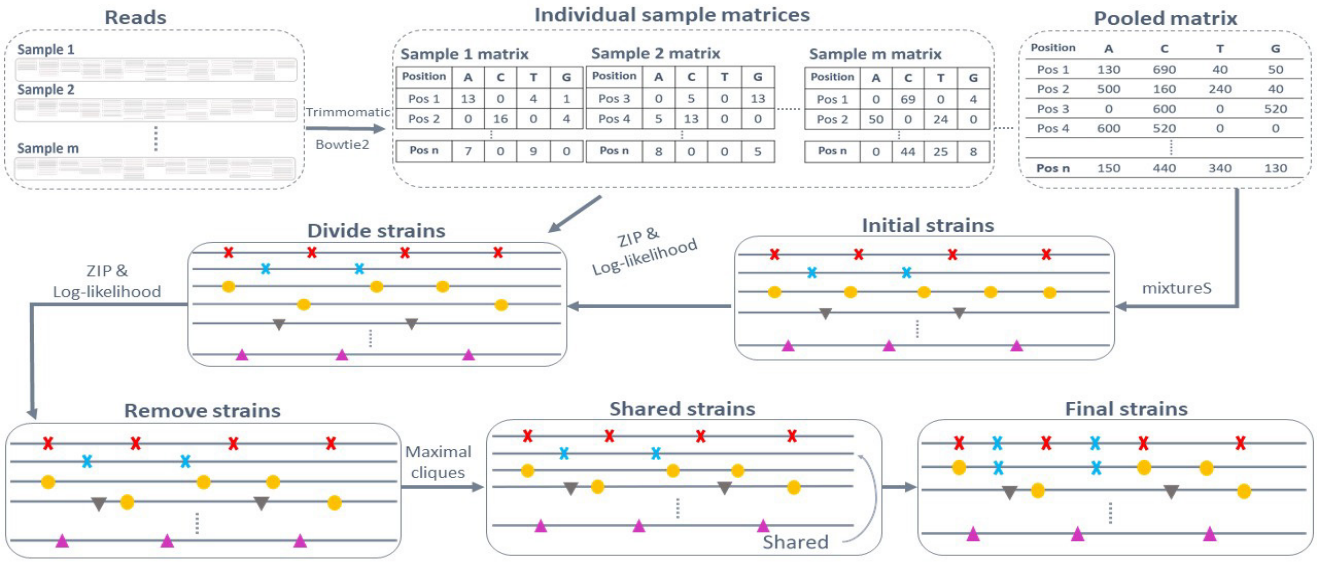


**Fig. 1. The SMS workflow.**

## 2.1. Identification of potential SNPs

With reads from the *m* samples, SMS trims reads, and filters low-quality reads with the tool trimmomatic

[41]. SMS then maps the cleaned reads to the reference genome by bowtie2 [42]. 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 potentially polymorphic if the following criteria are satisfied: 1). It has a coverage larger than 10% of the pooled coverage. 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, each with no smaller than 5% of the pooled coverage. Note that when the reference nucleotide at a position has fewer than 5% of the pooled 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 $n' \leq n1 \leq 3n'$. Note that despite the default requirement of at least 5% of the pooled coverage for any strain to be identified, SMS can identify low-abundance strains in multiple samples. A low-abundance strain may account for fewer than 0.01% of a metagenome. However, with a few dozen samples, its species may already have a reasonable coverage in the pooled sample, and SMS will identify each of its strains with at least 5% of the pooled species coverage in the pooled sample. As demonstrated in the following simulated studies, with the pooled species coverage 100X, SMS identified strains with a pooled coverage of 10X in 214 out of 216 datasets for three randomly chosen bacterial species.

**2.2. Prediction of the strain number and abundance**

With the *n1* potential SNPs, SMS infers the strain number and abundance in four steps.

First, SMS obtains an initial number of strains and their SNPs. SMS applies mixtureS to the above *n1*

SNPs with the pooled sample and outputs the predicted strains and their abundance. MixtureS reconstructs the strain genomes instead of local strain regions corresponding to marker genes from shotgun reads in one sample and has shown good performance previously [26, 40]. In this way, the strains with different pooled coverage are separated into *R* strains. *R* is automatically inferred.

Second, SMS refines the predicted strains so that almost all SNPs in an actual strain are assigned to one predicted strain. Since the coverage of SNPs from the same strain is 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 an 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 *n1* 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 an *m* by 1 coverage vector, the average of the coverage vectors of the SNPs currently assigned to this strain. SMS then reassigns each of the *n1* SNPs to the strain with the most similar coverage vector to the coverage vector of this SNP. With the reassigned 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 *n1* SNPs become more and more accurate, with almost all SNPs from an actual strain grouped together.

Third, SMS investigates whether there are more or fewer than *R* strains. 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 *n1* 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. 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, SMS removes this strain and reassigns its SNPs to all consistent strains in this 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.

### 2.3. The similarity of two coverage vectors

SMS calculates the similarity of two coverage vectors $(a_1, a_2, \dots, a_m)$ and $(b_1, b_2, \dots, b_m)$ 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 their Kendall rank correlation. 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, Euclendian distance, relative entropy, etc.

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

SMS models the coverage of the SNPs 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

$$ZIP(x,\pi,\lambda) = \begin{cases} \pi + (1-\pi) * \exp(-\lambda), for\ x = 0 \\ \dfrac{(1-\pi) * \lambda^x}{x!} * \exp(-\lambda), for\ x = 1, 2, 3, \ldots \end{cases}$$

Assume we have an *n1* by *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 this sample. When $y_{jr} = 1,$ we also define $b_{jr} = \sum_{i=1}^{n1} z_{ir} I_{x_{ij}=0}$, $n_{jr} = \sum_{i=1}^{n1} z_{ir}$, and $a_{jr} = \sum_{i=1}^{n1} z_{ir} x_{ij} \,/ n_{jr}$.

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 $s_{jr}^2 = \frac{\sum_{i=1}^{n1} z_{ir}(x_{ij} - a_{jr})^2}{\sum_{i=1}^{n1} z_{ir} - 1}$. SMS then uses the following iteration method to obtain the maximal likelihood estimation of $\pi_{jr}$ and $\lambda_{jr}$: first replaces $\pi_{jr}$ by $\pi_{jr} = \frac{n_{jr}(\lambda_{jr} - a_{jr}) e^{-\lambda_{jr}}}{\lambda_{jr} b_{jr} - n_{jr}(\lambda_{jr} - a_{jr})\left(1 - e^{-\lambda_{jr}}\right)}$ 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 $\lambda_{jr}$=$a_{jr}$.

**2.5. Log likelihood test**

Given $R$ strains, the full likelihood of observation the frequencies of these $n1$ SNPs in the $m$ samples is

$$L(X, Z|\pi, \lambda) = \prod_{i=1}^{n1} \prod_{j=1}^{m} (\sum_{r=1}^{R} z_{ir} y_{jr} ZIP(x_{ij}, \pi_{jr}, \lambda_{jr})).$$

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.

### 2.6. Simulated and experimental datasets

We simulated 702 datasets (Supplementary Table S1). As mentioned above, because SMS uses a species reference genome, we only need to consider reads from one species in a dataset. We thus simulated data with only one species in each dataset. In every dataset, a species reference genome was randomly chosen, 2 to 4 strains were simulated, and 5 to 35 samples were generated. For each reference genome, their four strains were generated by randomly choosing 0.01% of the genome positions and then randomly substituting the reference nucleotide with another nucleotide. This 0.01% mutation rate was from previous studies [16, 28], representing relatively more similar strains of a species (99.99% sequence identities) that are thus more challenging to distinguish from each other. The read coverage of a reference genome in a dataset was one of the following four coverage, 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 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 dwgsim (https://github.com/nh13/DWGSIM).

We tested SMS on 195 experimental datasets [11]. 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.

### 2.7. 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

## 3. Results

### 3.1. SMS correctly predicted the strain numbers

We studied the number of strains predicted in 702 simulated datasets (Supplementary 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 was the sum of the coverage of all strains of a 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 the configuration 10:20:30:40, the proportion of reads from the four strains was 10%, 20%, 30% and 40%, respectively.

Overall, SMS predicted the correct strain numbers in all but five datasets (Supplementary Tables S2-S5). Interestingly, SMS did not predict the correct strain number in at least one dataset for each of the three randomly selected species, implying that its performance was not species-specific. In each of the five datasets, a pair of strains shared 30% of their SNPs. In four of the five datasets, three strains shared 20% of their SNPs. These shared SNPs may have confused SMS when the coverage was 100X. When the coverage was increased, SMS predicted the correct strain number in each of the five corresponding datasets. These analyses 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).

### 3.2. SMS reliably estimated the strain abundance

We investigated how well SMS predicted the strain abundance. No matter whether the strain number was correctly predicted, the predicted strain abundance agreed well with the known strain abundance (Fig. 2, Supplementary Tables S2-S5). This agreement did not depend on the sample number, the pooled coverage, the strain number, etc.

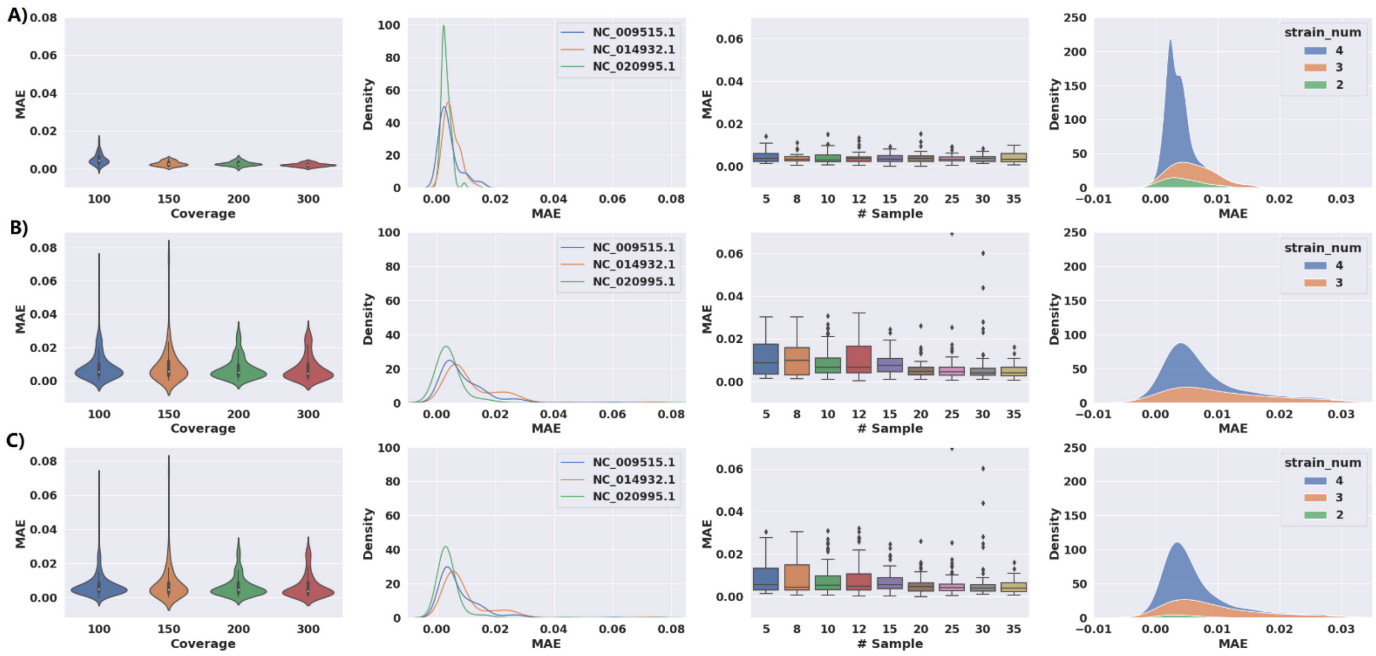


**Fig. 2. The predicted strain abundance. A) Unshared datasets; B) Shared datasets; and C) All datasets.** MAE is the average Maximal Absolute Difference between the predicted abundance and the corresponding true abundance across datasets.

In the 697 datasets SMS correctly predicted the strain number, 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 were 0.99 and 1.00, respectively. Even in the five datasets with the incorrectly predicted strain number, the predicted strain abundance was similar to the true abundance. For instance, SMS predicted four strains in three datasets with three strains (Supplementary Table S5). In two datasets, two strains had a predicted abundance of about 0.08 and 0.29, respectively, which were close to the corresponding true abundance of 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, SMS predicted the correct strain number and more similar abundance (Supplementary Table S5).

The accuracy was in general improved with more samples and a larger pooled coverage in a dataset (Fig. 2). For instance, when the sample number was larger, the median of the predicted abundance was closer

to the true abundance, and the variation of the maximal absolute difference (MAE) between the predicted abundance and the true abundance was smaller. The accuracy was not affected much by different species or the number of strains in a dataset (Fig. 2). For instance, the MAE 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 numbers of strains, etc.

### 3.3. SMS faithfully determined the SNPs

Existing methods mainly focus on the predicted strain number and only occasionally consider their 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 predict strain SNPs.

We studied the datasets without shared SNPs among strains (Supplementary Table S6). In all 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, they 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 precison 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.

We also studied the predicted strains on datasets with shared SNPs among strains (Supplementary Tables S7-S9). We again focused on the two most challenging configurations: 10:20:30:40 and 10:25:25:40. They 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 datasets (Supplementary Tables S7 and S8). The performance suggested that SMS could reconstruct the complicated evolutionary trajectories of strains with shotgun sequencing reads.

### 3.4. SMS performed well on experimental datasets

We tested SMS on 195 experimental datasets (Supplementary Table S10). We chose these datasets because their strain numbers were known. The strain abundance was also predicted previously [11]. Note that the datasets from the Critical Assessment of Metagenome Interpretation challenge did not provide the strain number, strain abundance and SNPs unique to strains, thus not suitable for the strain genome reconstruction here [18].

SMS identified two strains in each of these 195 datasets, which agreed well with the previous study [11]. 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 different datasets were the same, 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 [11]. 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 their similar abundance. Since the strain abundance was unknown, we compared the predicted abundance by SMS and the previous study. The difference between 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.

### 3.5 SMS reconstructed strain genomes better than existing methods

We compared SMS with mixtureS [26] and StrainFinder [29]. We did not compare other tools because mixtureS and StrainFinder showed better performance previously, and other tools may only work for marker gene regions instead of on the genome-scale [26]. Since mixtureS works on one sample, we ran it on the pooled sample in each dataset. Because StrainFinder cannot determine the strain numbers, we specified the known strain numbers in the corresponding datasets.

We compared the strain number, abundance and SNPs predicted by the three methods. SMS performed much better than others (Table 1). For instance, 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.

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

| Dataset | | SMS | | | mixtureS | | | StrainFinder | |
|---|---|---|---|---|---|---|---|---|---|
| | | # (%) of datasets | Precision, Recall, F1 | MAE | # (%) of datasets | Precision, Recall, F1 | MAE | Precision, Recall, F1 | MAE |
| 702 simulated datasets | Unshared | 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 |
| | Shared | 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 |
| | All | 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 74.87% | 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.

We also studied the running time of different methods (Supplementary Table S11). SMS took a little more time to run than mixtureS. However, the difference was not so evident. For all tools, the time cost mainly depended on the number of strains and SNPs, instead of the dataset sizes.

# 4. Discussion

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

SMS reconstructs bacterial strain genomes with a species reference genome and the raw sequencing reads. The reference is employed to map the cleaned reads. The chosen reference thus does not affect the predicted strain number and abundance, as they are inferred from the SNPs in strains that come from the mapped reads. 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 like SAMtools [43] to define SNPs in samples. Moreover, since reads are mapped to the reference genomes in advance, SMS can be applied to general metagenomic datasets instead of the simulated shotgun samples for individual species illustrated here.

SMS is not designed for the strain analysis of novel species. With more and more sequenced bacterial genomes, this issue may not be of 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 genomes. In this sense, what SMS reconstructs is similar to the strain core genomes but may include additional reference-specific regions. In the future, one may apply other machine learning and data mining methods [44-48] to further discover accessory genes in strains, with the inferred strain number and abundance in samples.

## Author Contributions

H.H. and X.L. designed the study. S.W., M.V., H.H. and X.L. analyzed data and wrote the manuscript.

## Conflict of interest

The authors declare no cometing interests.

## Acknowledgement

This work was supported by the National Science Foundation [1661414, 2015838, 2120907].

## References

[1] B.A. Methe, K.E. Nelson, M. Pop, H.H. Creasy, M.G. Giglio, C. Huttenhower, D. Gevers, J.F. Petrosino, S. Abubucker, J.H. Badger, A.T. Chinwalla, A.M. Earl, M.G. FitzGerald, R.S. Fulton, K. Hallsworth-Pepin, E.A. Lobos, R. Madupu, V. Magrini, J.C. Martin, M. Mitreva, D.M. Muzny, E.J. Sodergren, J. Versalovic, A.M. Wollam, K.C. Worley, J.R. Wortman, S.K. Young, Q. Zeng, K.M. Aagaard, O.O. Abolude, E. Allen-Vercoe, E.J. Alm, L. Alvarado, G.L. Andersen, S. Anderson, E. Appelbaum, H.M. Arachchi, G. Armitage, C.A. Arze, T. Ayvaz, C.C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M.J. Blaser, T. Bloom, V.R. Bonazzi, P. Brooks, G. Buck, C.J. Buhay, D.A. Busam, J.L. Campbell, S.R. Canon, B.L. Cantarel, P.S. Chain, I.M.A. Chen, L. Chen, S. Chhibba, K. Chu, D.M. Ciulla, J.C. Clemente, S.W. Clifton, S. Conlan, J. Crabtree, M.A. Cutting, N.J. Davidovics, C.C. Davis, T.Z. DeSantis, C. Deal, K.D. Delehaunty, F.E. Dewhisrst, E. Deych, Y. Ding, D.J. Dooling, S.P. Dugan, W.M. Dunne, A.S. Durkin, R.C. Edgar, R.L. Erlich, C.N. Farmer, R.M. Farrell, K. Faust, M. Feldgarden, V.M. Felix, S. Fisher, A.A. Fodor, L. Forney, L. Foster, V. Di Francesco, J. Friedman, D.C. Friedrich, C.C. Fronick, L.L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M.Y. Giovanni, J.M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, B.J. Haas, H.A. Hamilton, E.L. Harris, T.A. Hepburn, B. Herter, D.E. Hoffmann, M.E. Holder, C. Howarth, K.H. Huang, S.M. Huse, J. Izard, J.K. Jansson, H.Y. Jiang, C. Jordan, V. Joshi, J. Katancik, W. Keitel, S.T. Kelley, C. Kells, S. Kinder-Haake, N.B. King, R. Knight, D. Knights, H.H. Kong, O. Koren, S. Koren, K.C. Kota, C.L. Kovar, N.C. Kyrpides, P.S. La Rosa, S.L. Lee, K.P. Lemon, N. Lennon, C.M. Lewis, L. Lewis, R.E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu, C.C. Lo, C.A. Lozupone, R.D. Lunsford, T. Madden, A.A. Mahurkar, P.J. Mannon, E.R. Mardis, V.M. Markowitz, K. Mavrommatis, J.M. McCorrison, D. McDonald, J. McEwen, A.L. McGuire, P. McInnes, T. Mehta, K.A. Mihindukulasuriya, J.R. Miller, P.J. Minx, I. Newsham, C. Nusbaum, M. O'Laughlin, J. Orvis, I. Pagani, K. Palaniappan, S.M. Patel, M.

Pearson, J. Peterson, M. Podar, C. Pohl, K.S. Pollard, M.E. Priest, L.M. Proctor, X. Qin, J. Raes, J. Ravel, J.G. Reid, M. Rho, R. Rhodes, K.P. Riehle, M.C. Rivera, B. Rodriguez-Mueller, Y.H. Rogers, M.C. Ross, C. Russ, R.K. Sanka, P. Sankar, J.F. Sathirapongsasuti, J.A. Schloss, P.D. Schloss, T.M. Schmidt, M. Scholz, L. Schriml, A.M. Schubert, N. Segata, J.A. Segre, W.D. Shannon, R.R. Sharp, T.J. Sharpton, N. Shenoy, N.U. Sheth, G.A. Simone, I. Singh, C.S. Smillie, J.D. Sobel, D.D. Sommer, P. Spicer, G.G. Sutton, S.M. Sykes, D.G. Tabbaa, M. Thiagarajan, C.M. Tomlinson, M. Torralba, T.J. Treangen, R.M. Truty, T.A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D.V. Ward, W. Warren, M.A. Watson, C. Wellington, K.A. Wetterstrand, J.R. White, K. Wilczek-Boney, Y.Q. Wu, K.M. Wylie, T. Wylie, C. Yandava, L. Ye, Y. Ye, S. Yooseph, B.P. Youmans, L. Zhang, Y.J. Zhou, Y.M. Zhu, L. Zoloth, J.D. Zucker, B.W. Birren, R.A. Gibbs, S.K. Highlander, G.M. Weinstock, R.K. Wilson, O. White, H.M.P. Consortiu, A framework for human microbiome research, Nature, 486 (2012) 215-221.

[2] L.M. Proctor, H.H. Creasy, J.M. Fettweis, J. Lloyd-Price, A. Mahurkar, W.Y. Zhou, G.A. Buck, M.P. Snyder, J.F. Strauss, G.M. Weinstock, O. White, C. Huttenhower, I.H.i.R. Network, The Integrative Human Microbiome Project, Nature, 569 (2019) 641-648.

[3] L.M. Proctor, I.H.i.R. Network, The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease, Cell host & microbe, 16 (2014) 276-289.

[4] D.A. Rasko, V. Sperandio, Anti-virulence strategies to combat bacteria-mediated disease, Nat Rev Drug Discov, 9 (2010) 117-128.

[5] G.W. Tyson, J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, J.F. Banfield, Community structure and metabolism through reconstruction of microbial genomes from the environment, Nature, 428 (2004) 37-43.

[6] J.C. Venter, K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, I. Paulsen, K.E. Nelson, W. Nelson, D.E. Fouts, S. Levy, A.H. Knap, M.W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.H. Rogers, H.O. Smith, Environmental genome shotgun sequencing of the Sargasso Sea, Science (New York, N.Y, 304 (2004) 66-74.

[7] Y. Wang, S. Goodison, X. Li, H. Hu, Prognostic cancer gene signatures share common regulatory motifs, Sci Rep, 7 (2017) 4750.

[8] J.C. Wooley, A. Godzik, I. Friedberg, A primer on metagenomics, PLoS computational biology, 6 (2010) e1000667.

[9] T. Van Rossum, P. Ferretti, O.M. Maistrenko, P. Bork, Diversity within species: interpreting strains in microbiomes, Nat Rev Microbiol, 18 (2020) 491-506.

[10] D.W. Eyre, M.L. Cule, D. Griffiths, D.W. Crook, T.E. Peto, A.S. Walker, D.J. Wilson, Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in Clostridium difficile transmission, PLoS computational biology, 9 (2013) e1003059.
[11] B. Sobkowiak, J.R. Glynn, R.M.G.J. Houben, K. Mallard, J.E. Phelan, J.A. Guerra-Assuncao, L. Banda, T. Mzembe, M. Viveiros, R. McNerney, J. Parkhill, A.C. Crampin, T.G. Clark, Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data, BMC genomics, 19 (2018) 613.
[12] J.A. Eisen, Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes, PLoS biology, 5 (2007) e82.
[13] X. Li, S.A. Naser, A. Khaled, H. Hu, X. Li, When old metagenomic data meet newly sequenced genomes, a case study, PloS one, 13 (2018) e0198773.
[14] Y. Wang, H. Hu, X. Li, MBMC: An Effective Markov Chain Approach for Binning Metagenomic Reads from Environmental Shotgun Sequencing Projects, Omics : a journal of integrative biology, 20 (2016) 470-479.
[15] B. Cleary, I.L. Brito, K. Huang, D. Gevers, T. Shea, S. Young, E.J. Alm, Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning, Nature biotechnology, 33 (2015) 1053-1060.
[16] X. Li, S. Saadat, H.Y. Hu, X.M. Li, BHap: a novel approach for bacterial haplotype reconstruction, Bioinformatics (Oxford, England), 35 (2019) 4624-4631.
[17] O. Kyrgyzov, V. Prost, S. Gazut, B. Farcy, T. Bruls, Binning unassembled short reads based on k-mer abundance covariance using sparse coding, GigaScience, 9 (2020).
[18] A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Droge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T.S. Jorgensen, N. Shapiro, P.D. Blood, A. Gurevich, Y. Bai, D. Turaev, M.Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvociute, L.H. Hansen, S.J. Sorensen, B.K.H. Chia, B. Denis, J.L. Froula, Z. Wang, R. Egan, D. Don Kang, J.J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.W. Wu, S.W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M.D. Barton, T. Lingner, H.H. Lin, Y.C. Liao, G.G.Z. Silva, D.A. Cuevas, R.A. Edwards, S. Saha, V.C. Piro, B.Y. Renard, M. Pop, H.P. Klenk, M. Goker, N.C. Kyrpides, T. Woyke, J.A. Vorholt, P. Schulze-Lefert, E.M. Rubin, A.E. Darling, T. Rattei, A.C. McHardy, Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software, Nature methods, 14 (2017) 1063-1071.
[19] A.J. van der Walt, M.W. van Goethem, J.B. Ramond, T.P. Makhalanyane, O. Reva, D.A. Cowan, Assembling metagenomes, one community at a time, BMC genomics, 18 (2017).

[20] D. Gevers, S. Kugathasan, D. Knights, A.D. Kostic, R. Knight, R.J. Xavier, A Microbiome Foundation for the Study of Crohn's Disease, Cell host & microbe, 21 (2017) 301-304.
[21] D.H. Parks, C. Rinke, M. Chuvochina, P.A. Chaumeil, B.J. Woodcroft, P.N. Evans, P. Hugenholtz, G.W. Tyson, Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life, Nat Microbiol, 2 (2017) 1533-1542.
[22] E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M.C. Collado, B.L. Rice, C. DuLong, X.C. Morgan, C.D. Golden, C. Quince, C. Huttenhower, N. Segata, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle, Cell, 176 (2019) 649-+.
[23] D. Albanese, C. Donati, Strain profiling and epidemiology of bacterial species from metagenomic sequencing, Nat Commun, 8 (2017).
[24] C. Anyansi, T.J. Straub, A.L. Manson, A.M. Earl, T. Abeel, Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data, Front Microbiol, 11 (2020) 1925.
[25] C.J. Hong, S. Manimaran, Y. Shen, J.F. Perez-Rogers, A.L. Byrd, E. Castro-Nallar, K.A. Crandall, W.E. Johnson, PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples, Microbiome, 2 (2014).
[26] X. Li, H. Hu, X. Li, mixtureS: a novel tool for bacterial strain reconstruction from reads, Bioinformatics (Oxford, England), (2020).
[27] C. Luo, R. Knight, H. Siljander, M. Knip, R.J. Xavier, D. Gevers, ConStrains identifies microbial strains in metagenomic datasets, Nature biotechnology, 33 (2015) 1045-1052.
[28] S. Pulido-Tamayo, A. Sanchez-Rodriguez, T. Swings, B. Van den Bergh, A. Dubey, H. Steenackers, J. Michiels, J. Fostier, K. Marchal, Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations, Nucleic acids research, 43 (2015) e105.
[29] C.S. Smillie, J. Sauk, D. Gevers, J. Friedman, J. Sung, I. Youngster, E.L. Hohmann, C. Staley, A. Khoruts, M.J. Sadowsky, J.R. Allegretti, M.B. Smith, R.J. Xavier, E.J. Alm, Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation, Cell host & microbe, 23 (2018) 229-+.
[30] T.H. Ahn, J.J. Chai, C.L. Pan, Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance, Bioinformatics (Oxford, England), 31 (2015) 170-177.
[31] P.I. Costea, R. Munch, L.P. Coelho, L. Paoli, S. Sunagawa, P. Bork, metaSNV: A tool for metagenomic strain level analysis, PloS one, 12 (2017) e0182392.

[32] S. Nayfach, B. Rodriguez-Mueller, N. Garud, K.S. Pollard, An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography, Genome research, 26 (2016) 1612-1625.
[33] C. Quince, T.O. Delmont, S. Raguideau, J. Alneberg, A.E. Darling, G. Collins, A.M. Eren, DESMAN: a new tool for de novo extraction of strains from metagenomes, Genome biology, 18 (2017) 181.
[34] M. Roosaare, M. Vaher, L. Kaplinski, M. Mols, R. Andreson, M. Lepamets, T. Koressaar, P. Naaber, S. Koljalg, M. Remm, StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees, PeerJ, 5 (2017) e3353.
[35] A. Sankar, B. Malone, S.C. Bayliss, B. Pascoe, G. Meric, M.D. Hitchings, S.K. Sheppard, E.J. Feil, J. Corander, A. Honkela, Bayesian identification of bacterial strains from sequencing data, Microb Genom, 2 (2016) e000075.
[36] M. Scholz, D.V. Ward, E. Pasolli, T. Tolio, M. Zolfo, F. Asnicar, D.T. Truong, A. Tett, A.L. Morrow, N. Segata, Strain-level microbial epidemiology and population genomics from shotgun metagenomics, Nature methods, 13 (2016) 435-438.
[37] F.B. Tamburini, T.M. Andermann, E. Tkachenko, F. Senchyna, N. Banaei, A.S. Bhatt, Precision identification of diverse bloodstream pathogens in the gut microbiome, Nat Med, 24 (2018) 1809-1814.
[38] D.T. Truong, A. Tett, E. Pasolli, C. Huttenhower, N. Segata, Microbial strain-level population structure and genetic diversity from metagenomes, Genome research, 27 (2017) 626-638.
[39] C. Quince, S. Nurk, S. Raguideau, R. James, O.S. Soyer, J.K. Summers, A. Limasset, A.M. Eren, R. Chikhi, A.E. Darling, STRONG: metagenomics strain resolution on assembly graphs, Genome biology, 22 (2021) 214.
[40] M.F. Ventolero, S. Wang, H. Hu, X. Li, Computational analyses of bacterial strains from shotgun reads, Briefings in bioinformatics, 23 (2022).
[41] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics (Oxford, England), 30 (2014) 2114-2120.
[42] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nature methods, 9 (2012) 357-359.
[43] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, S. Genome Project Data Processing, The Sequence Alignment/Map format and SAMtools, Bioinformatics (Oxford, England), 25 (2009) 2078-2079.
[44] K. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.
[45] A. Talukder, S. Saadat, X. Li, H. Hu, EPIP: a novel approach for condition-specific enhancer-

promoter interaction prediction, Bioinformatics (Oxford, England), 35 (2019) 3877-3883.
[46] R. Tibshirani, P. Wang, Spatial smoothing and hot spot detection for CGH data using the fused lasso, Biostatistics, 9 (2008) 18-29.
[47] C. Zhao, X. Li, H. Hu, PETModule: a motif module based approach for enhancer target gene prediction, Sci Rep, 6 (2016) 30043.
[48] J. Zhou, J. Liu, V.A. Narayan, J. Ye, Modeling Disease Progression via Fused Sparse Group Lasso, KDD, 2012 (2012) 1095-1103.