Computational analyses of bacterial strains from shotgun reads

Minerva Fatimae Ventolero[1*], Saidi Wang[2*], Haiyan Hu[2,3,#] and Xiaoman Li[1,#]
[1] Burnett School of Biomedical Science, University of Central Florida, Orlando, FL, 32816
[2] Department of Computer Science, University of Central Florida, Orlando, FL, 32816
[3] Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL, 32816

*Contribute equally.
[#]To whom correspondence should be addressed.

Abstract

Shotgun sequencing is routinely employed to study bacteria in microbial communities. With the vast amount of shotgun sequencing reads generated in a metagenomic project, it is crucial to determine the microbial composition at the strain level. This study investigated twenty computational tools that attempt to infer bacterial strain genomes from shotgun reads. For the first time, we discussed the methodology behind these tools. We also systematically evaluated six novel-strain-targeting tools on the same datasets and found that BHap, mixtureS and StrainFinder performed better than other tools. Because the performance of the best tools is still suboptimal, we discussed future directions that may address the limitations.

Keywords: metagenomics, shotgun sequencing, tool comparison, bacterial strains, bacterial strain genome reconstruction.

Introduction

Shotgun metagenomic sequencing is routinely applied to study microbes [1-12]. In shotgun metagenomic sequencing, mixed DNA from all microbial species and strains in an environmental/clinical sample are randomly cut into pieces. These short DNA fragments are then sequenced to generate reads. Ideally, these mixed shotgun reads from different species/strains in a sample proportionally represent the abundance of the original DNA, with more reads from more abundant species/strains, and vice versa. With the shotgun reads and other data, one may then answer key questions such as which species are present, how abundant each species is, how these species work together in the community under a specific condition, and so on [13].

Many computational methods have been developed to analyze shotgun metagenomic reads to study microbial genomes [14-16]. These methods can be classified into two main categories: assembly-based and binning-based. The assembly-based methods assemble the mixed reads into individual microbial genomes [15-28]. These approaches face great challenges such as the mixed reads from a large number of unknown species, sequencing errors in reads, uneven sequencing coverage across microbial genomes, horizontal gene transfers (HGTs) that produce "duplicated" genes in different species, etc. Moreover, when there are billions of reads involved in a project, rarely can existing assembly approaches address the assembly problem even on commercial computer servers [23, 29]. In addition, there are usually misassembled regions in the resulted genomes and/or contigs, which are challenging to identify and resolve [16, 26]. Despite these challenges, the assembly-based methods have successfully reconstructed a number of microbial genomes. The other main category of computational methods is the binning-based methods. These methods cluster reads into groups based on sequence composition and/or sequence

homology so that each group is likely to represent reads from a taxon at a specific taxonomic level (phylum, class, order, family, genus, species) [16, 30-34]. Later, with the sequencing cost dramatically decreased, multiple shotgun metagenomic samples from the same environments or human body sites are widely available [35-45]. It becomes popular to bin pre-assembled contigs from multiple metagenomic samples, resulting in a large number of metagenome-assembled-genomes (MAGs) [35, 36]. Currently, contig-binning is arguably the best strategy to reconstruct bacterial genomes. Although contig-binning barely produces close-to-complete genomes, certain ones are of good quality [35, 36].

Despite the existence of many computational methods to analyze shotgun metagenomic reads, most of them are developed to study taxons at the species or higher levels [16]. They are not designed to analyze microbial strains, especially bacterial strains. Bacterial strains play critical roles in our environments and human health [46-52]. Such potential roles can be seen in bioremediation [53-55], their utility in probiotics development [56], and their diagnostic potential for health and disease states [57, 58], etc. For instance, *Acinetobacter calcoaceticus* GK2 strain [59], *Mucor hiemalis* strains [60], and *Lactobacillus plantarum* MF042018 [61] have shown potential for diesel or metal bioremediation. Lactic acid bacterial strains of *Lactobacillus, Pediococcus* and *Weissella* have the potential for probiotics development [62]. *Lactobacillus paracasei* CP133, *Lactobacillus plantarum* CP134, and *Bacillus subtilis* CP350 have been found to have probiotic potential when used as additives in feeds [63]. In human health and disease research, *Staphylococcus hominis* A9 strain has been explored for its bacteriotherapy potential on Atopic Dermatitis [64]. In cancer therapeutics studies, specific strains of *Bifidobacterium bifidum* [65] and the attenuated recombinant *Listeria* strains [66] have shown their potential in tumor reduction. Because of the important roles of bacterial strains, it is imperative to create computational methods to study bacterial strains from shotgun reads.

It is still a daunting task to reconstruct bacterial strain genomes from shotgun metagenomic reads. The complexity lies in the fact that the existing methods often mistake the variations in strains for sequencing errors in reads, in addition to their inability to handle the low coverage of species and the mixture nature of reads from a large number of species and strains [17, 67-70]. Even for the analysis of shotgun reads from an individual species, it is still not trivial to reconstruct the strain genomes of this species from reads. The non-trivial nature of the bacterial strain genome reconstruction is because bacterial DNA constantly evolves, with the mutation rate from one to one hundred per billion base pairs per generation [71]. Such a small mutation rate results in no pair of variant sites within the same read, and two strains of the same bacterial species may share more than 99.9% sequence identities in their shared genomes [72, 73]. This low mutation rate makes bacterial genomes different from viral genomes, in which the mutation rate is so high that multiple variant sites in the same reads enable a ladder to connect adjacent variant sites through read overlapping to reconstruct the viral strain genomes [72, 73].

Because of the importance of bacterial strains, we surveyed twenty tools on bacterial strain analysis (Figure 1). The criteria to choose a tool to survey here were: (1) The tool does the strain analysis with next-generation sequencing reads and a reference; (2). The tool is claimed to identify known bacterial strains and/or predict novel bacterial strains; and (3) The tool is freely available to the public. Unlike a couple of existing survey papers on bacterial strain analyses [46,

48], we focused more on the practical issues of the methodology involved and emphasized the scenarios for their usage in the following sections.

Figure 1: A classification of the existing tools for bacterial strain analysis.

The tools we surveyed can be broadly classified into two categories: known-strain-based [74-82] and novel-strain-targeting [47, 49, 72, 73, 83-89]. The known-strain-based tools rely on prior knowledge of known strains to discover known strains, although they may detect or identify novel strains. The novel-strain-targeting tools de novo reconstruct local or genome-wide novel strains with shotgun metagenomic reads and a species reference. These tools thus do not require any strain information and can be further divided into two categories, local or genome-wide novel-strain-targeting tools. The former reconstruct novel strains at the marker gene regions with the species marker genes, while the latter reconstruct novel strains in the entire genome with a species reference genome. Because the current knowledge of known strains is limited and novel strains are routinely present in different studies due to the constant mutations in bacterial genomes, we emphasized the genome-wide novel-strain-targeting tools in this survey.

Although computational tools are available for novel strain genome reconstruction, their performance has not been systematically evaluated on the same datasets. To understand these tools better, we evaluated all five genome-wide novel-strain-targeting tools (BHap, EVORhA, MetaSNV, mixtureS, and StrainFinder [72, 73, 87-89]). Although StrainPhlAn is a local novel-strain-targeting tool, we included it because of its higher citations [49]. We also tried to evaluate other local novel-strain-targeting tools such as MIDAS [83] and PStrain [86]. However, MIDAS did not report the strains and their abundance on our test data, and we failed to run PStrain.

In the following, first, we provided an overview of the twenty tools and categorized them into three types (Figure 1). We then presented the details of six local novel-strain-targeting tools and five genome-wide novel-strain-targeting tools. Next, we evaluated the performance of the aforementioned six tools for novel strain reconstruction. Finally, we commented on the future directions of bacterial strain analyses.

Overview of the computational bacterial strain analysis tools

Similar to typical metagenomic analyses, bacterial strain analyses from shotgun reads attempt to identify the strains present, the abundance of each strain, the function and evolution advantages of different strains, etc. [90-95]. To answer these questions, it often needs to reconstruct the bacterial strain genomes [46]. In some sense, this is equal to identifying the variant sites or regions in different strains with respect to the species reference genomes [80, 88].

Many existing computational tools for bacterial strain analyses are known-strain-based [74, 75, 77, 78, 80-82, 96] (Figure 1, Table 1). For these tools, the strain genome sequences or strain variant sites, k-mers, or gene contents are known for the species under consideration. The known strain variant sites are often derived from genome-typing of individual culture isolates, such as the Multi Locus Sequence Typing [82]. The known strain genomes are usually obtained from whole-genome shotgun sequencing of culture isolates by Sanger sequencing [97], from which the strain k-mers and gene contents can be inferred. With the known strain information, these methods usually map reads to the strain or species genomes to define variant sites, k-mers, gene

contents, etc. They then compare the variants sites, k-mers, gene contents, etc., with the known ones to determine whether a strain is present. They next measure the abundance of a known strain by the coverage of its corresponding variant sites, unique k-mers, genes, etc., based on the mapped reads. These methods, therefore, intend to determine which subset of known strains are present and their abundance. Although this type of analysis is valuable to understand bacterial strains, they are mostly for discovering known strains, instead of reconstructing novel strains. In practice, many mutations are likely to have accumulated in bacterial strains. Novel instead of known strain genomes are thus expected in samples. In other words, the known strain information may be limited in practice [98, 99].

Table 1: The twenty tools surveyed.

Without prior knowledge of bacterial strains, almost all computational tools that reconstruct novel strain genomes are reference-based (Figure 1, Table 1). In literature, reference-dependent sometimes refers to tools based on known strains, while novel-strain-targeting tools are called reference-independent [46]. Here we emphasize that the latter are also reference-dependent, although no strain information is required. References are indispensable for computational strain analysis because alternatively, one has to de novo assemble reads to reconstruct the species or strain genomes, which rarely have the desired quality [84]. In addition, this becomes a formidable task when we study low-abundance species [29]. A couple of methods claimed to reconstruct strain genomes of low-abundance species, which are not widely tested or supported by the literature [29, 100].

The novel-strain-targeting tools have a common workflow, which maps reads to the references, identifies variant sites based on the mapped reads, and infers the mixture of strains based on the frequency difference of the variant sites. These tools differed from each other mainly at the steps of defining variant sites and assigning them into different strains. The rationale behind this workflow is that the DNA segments from all strains are sampled and proportionally represented in the shotgun reads. The variant sites in strains can be identified by mapping reads to the reference, which may mistake sequencing errors in reads for variant sites. Because the sequencing error rate is low, the frequency of the "variant sites" caused by sequencing errors in the mapped reads is presumably much smaller than the coverage of true variant sites in identifiable strains. Moreover, since different strains of a species often have different abundance in a sample, variant sites of different strains have different read coverage and can thus be separated from each other. Here the coverage of a genome, a genomic fragment, a genomic position, or a variant site is defined as the average number of reads mapped to every position in the corresponding region. The number of mapped reads containing any site unique to a strain is thus similar due to the assumed uniform coverage of this strain genome in shotgun sequencing. With the assumption of uniform coverage of every strain and different abundance of different strains, one thus can classify the mixed sites based on their coverage so that the sites from individual strains are clustered into separate groups. This uniform coverage and abundance difference of strains are the basic assumptions and rationale behind the existing strain analysis tools that aim to de novo reconstruct the bacterial strain genomes [47, 84, 88].

In summary, almost all existing computational strain analysis tools are reference-based, in the sense that known-strain-based tools call for reference strain genomes or strain variant

information while novel-strain-targeting tools require species genomes or marker gene sequences. In the following, we present the details of the novel-strain-targeting tools in two sections and leave the details of the known-strain-based in Supplementary File S1.

Local novel-strain-targeting tools

This type of tools discover novel strains based on variant sites in local marker gene regions. Identifying novel strains based on marker genes is fast and efficient, especially to pinpoint potentially biologically or clinically important species. It is the way to have a global picture of which species and strains are present in a microbiome. Here we described six such tools that analyze bacterial strains based on marker genes.

ConStrains [47]. ConStrains can identify strains in one or multiple samples. In at least one sample, the coverage of the species needs to be >=10X. It applies MetaPhlAn [101] to determine the existence of a species in a sample. For every present species, ConStrains maps reads by bowtie2 [102] to its PhyloPhlAn marker gene set stored in a custom database [103]. ConStrains then processes the mapped reads with the SAMtools to identify single nucleotide polymorphisms (SNPs) [104]. SNPs are clustered based on a flow algorithm to construct different candidate strain models, assuming that SNPs from different strains have a different frequency pattern in the sample(s). A Markov Chain Monte Carlo procedure then optimizes the candidate strain models to obtain the abundance of the strain models. The Akaike information criterion then selects the final strain models and their corresponding abundance.

MIDAS [83]. MIDAS requires a database of reference genomes of all species under consideration. This database also has the universal single-copy genes in each of these species. The universal single-copy genes are believed to be single-copy genes in a bacterial genome. A few dozen universal single-copy genes were identified previously [105-107]. When multiple reference genomes are available for a species, the gene content in its different strains is also included in the MIDAS database. MIDAS maps reads in one or multiple samples to these genes to determine the abundance of a species. For species with an abundance of at least 10X, MIDAS maps reads to the corresponding species reference genome to define SNPs, minor and major alleles, etc. MIDAS also provides gene content analysis for the abundant species with multiple sequenced reference genomes. In summary, MIDAS does not reconstruct or determine the presence of a strain but identifies the abundant species, their gene content, and SNPs. Because of this, the authors acknowledge that strain abundance estimation is not possible with the tool. Additionally, the strain level analysis relies on a majority rule to define a consensus allele for a given position considered. The strain composed of the consensus alleles is likely to be a mixture of multiple unknown strains.

StrainPhlAn [49]. StrainPhlAn maps reads to marker genes of each species defined in MetaPhlAn2 [101] by bowtie 2 [102]. In general, 200 species-specific marker genes per species are stored in MetaPhlAn2. The mapped reads are used to infer the consensus sequence for each marker gene in each species. The consensus sequences for different marker genes in a species are concatenated to represent this species genome, the marker gene portion of the species genome. StrainPhlAn then defines polymorphic sites based on binomial testing, with the sequencing error rate as 0.01. StrainPhlAn defines a dominant strain as the concatenated most frequent nucleotide at each polymorphic site. Only one strain can be defined in this way, although StrainPhlAn can

detect the existence of more than one strain. Because the frequency of the dominant SNPs at different locations could be very different, these SNPs may be from different strains instead of one dominant strain, and there may be way more than two strains.

DESMAN [84]. DESMAN can de novo identify strains from multiple shotgun metagenomic samples. It is claimed to be able to do so without a reference genome. This claim is somewhat misleading as it starts with a MAG, an assembled draft genome. Although MAGs are obtained from direct read assembly and contig binning, the quality of the MAGs may greatly affect the identified strains. In this study, DESMAN has combined several MAGs to represent the *E. coli* genome since the E. coli species genomes are grouped into multiple MAGs during assembly. With the pre-assembled contigs in the merged MAG, DESMAN applies RPS-BLAST to map contigs to the 36 universal single-copy genes to identify the corresponding genes in the MAG under consideration. Then it maps reads to these identified genes and identifies SNPs. Note that since the 36 genes may have multiple copies in a strain genome due to HGTs, it removes these genes and only considers the remaining genes. From the SNPs in these remaining genes, it considers them as a mixture multinomial distribution. It infers the strains based on Gibbs sampling [108, 109], assuming that there are a fixed number of strains in input samples, and these strains may have different coverage in different samples. Because DESMAN calls for MAGs and contigs, it may not work well with the low-abundant species in samples.

Strain-GeMS [85]. Strain-GeMs is similar to ConStrains [47]. As ConStrains, it applies MetaPhlAn to identify the present species and selects species with coverage >10X. Moreover, it also maps read to PhyloPhlAn marker genes for each selected species to identify SNPs. Unlike ConStrains, it defines and clusters SNPs in each sample with multiGeMs [110]. Finally, it applies ConStrains to identify strains. There is another difference between Strain-GeMS and Constrains in that Strain-GeMS requires more than 10X coverage per sample for a species, while ConStrains requires the coverage of 10X in at least one sample.

PStrain [86]. PStrain infers the strains and their abundance in individual samples. It maps reads to the MetaPhlAn2 marker gene database [101] with Bowtie 2 [102]. It then chooses species with at least 5X coverage for strain analysis. It defines SNPs based on the mapped reads to the MetaPhlAn2 marker genes with the GATK tool [111]. Assume there are k strains in a sample. PStrain iteratively infers k strains (assigns SNPs to the k strains) and their abundance by dynamic programming and linear programming. PStrain starts from k=1 and increases k by one each time to determine the optimum k through the elbow method.

Genome-wide novel-strain-targeting tools

When the goal is to thoroughly understand the strains of a species and how a strain causes drug resistance, one may want to reconstruct the entire strain genomes instead of local regions around marker genes. Only a handful of methods are developed for this purpose, most of which are designed to analyze individual samples, with essentially only StrainFinder developed for strain core genome reconstruction in multiple samples [88] (Figure 1, Table 1). In this study, strain core and accessory genomes are slightly different from those in literature. Here a strain core genome refers to the genomic portion of this strain mapped to the species reference genome. Similarly, a strain accessory genome is the unique genomic portion of this strain not represented

in the species reference genome. To our knowledge, the following are the only tools that can reconstruct bacterial strain core genomes based on a species reference genome.

EVORhA [73]. EVORhA is arguably the first tool for de novo bacterial strain genome reconstruction from shotgun reads. It maps reads to the reference genome of interest, reconstructs the local strains based on the shared SNPs in overlapping reads, and extends these local strains as long as possible. Next, it groups the extended local strains by a mixture of Gaussian modeling, with each distribution for the coverage of the local strains in a group. Finally, it connects each group of local strains with the remaining extended local strains that are not grouped and represent the shared regions by multiple strains into genome-wide strains, assuming that different strains evolved from a common ancestor. EVORhA was tested on simulated and experimental datasets and showed reliable performance.

MetaSNV [87]. MetaSNV requires a database of reference genomes, which users can also provide. Shotgun reads in samples are mapped to the reference genomes to generate the alignment file. The strain analysis from the alignment file involves the following three-step procedure: First, the genome coverage in samples is determined. Second, MetaSNV determines the genomic variation sites position-by-position, where at least four reads containing the variant at a given position are considered a SNP. Variants are categorized as either population or individual variants based on whether they are observed within the default threshold of >1% frequency across samples or observed in more than three reads in at least one sample, respectively. Third, the coverage and SNP information are used for taxon determination and downstream analyses, such as the computation of distance matrices and nucleotide diversity. The tool provides the identified variant sites in the reference genomes and does not reconstruct the strain genomes as expected.

StrainFinder [88]. StrainFinder reconstructs strain genomes in one or multiple samples with a reference genome of interest. It maps reads in one or multiple samples to the reference genome. In the case of multiple samples, reads in each sample are mapped and counted separately. StrainFinder considers the frequency of nucleotides in the mapped reads in every position of the entire genome as a mixture of multinomial distributions. The multinomial distribution is employed because the observed nucleotide frequency at a given locus is the sum of the corresponding nucleotide frequency from different strains or independently and identically distributed sequencing errors. StrainFinder applies an EM algorithm to model and infer the present strains in the sample(s) to decompose this multinomial mixture. StrainFinder needs to specify the strain number to run on a dataset.

BHap [72]. BHap reconstructs bacterial strain genomes in individual samples. It requires a reference genome, and the shotgun reads in the sample. BHap applies Velvet [27] to assemble reads into contigs without error correction, which prevents SNPs from being removed in the assembly by Velvet. It then constructs a flow network with these uncorrected contigs that have coverage larger than a threshold, which is likely to remove contigs that contain sequencing errors. In this network, contigs are nodes, and weighted edges connect two contigs if they are mapped to the overlapping regions in the reference genome. In this way, the strain genome reconstruction problem becomes the flow decomposition problem. BHap applies an EM

algorithm to infer the strains and their abundance [112, 113]. Based on the testing on simulated and experimental data, BHap can determine the strain number and the abundance of strains reliably.

mixtureS [89]. Like BHap, mixtureS reconstructs bacterial strain genomes with a species reference genome and the shotgun reads in individual samples. It does not have the read assembly procedure in BHap and is thus much simpler. It maps reads to the reference genome and then considers only the biallelic SNPs to infer the strain genomes. By assuming different strains have different coverage in the sample, mixtureS models SNP frequency by the binomial distribution and reconstructs the strain genomes by an EM algorithm. Because of the consideration of the biallelic SNPs, the inference is fast while accurate. It performs better than other strain genome reconstruction tools [72, 73, 88] in terms of a more accurate predicted number of strains and the strain abundance. However, mixtureS does not consider Triallelic/Quatra-allelic SNPs.

Material and Methods for tool comparisons

We tested the six tools on 108 simulated datasets and 196 experimental datasets on a desktop computer with the following configuration: Intel core i9-9900KF CPU @3.6HZ with 16 cores, 32 gigabytes memory, and Ubuntu 18.04 Long Term Support. We ran each tool except StrainFinder with their default parameters. We ran StrainFinder with the input of the correct strain number in each dataset because otherwise StrainFinder predicted an incorrect strain number and had a poor performance.

We tested the tools on datasets with reads from an individual species in most cases below. This is because all tools we surveyed infer bacterial strain genomes from reads with a reference genome. The majority of reads from a species of interest can thus be separated from other reads in a microbiome by mapping reads to the given reference. This reference prerequisite may prevent these tools from being applied to the study of strains of unknown species. However, we usually have the known genomes for the pathogens that scientists are concerned about. Moreover, these tools can still be applied to new species with the pre-assembled MAGs. In this case, the inferred strain number and abundance are still reasonable, while the actual SNPs and genomes may be different from the true ones because of the lower quality of the MAGs compared with the sequenced reference genomes.

To evaluate the accuracy of the predicted SNPs on datasets with known strain SNPs, we calculated the precision, recall and F1 scores of the predicted SNPs in strains. In brief, for each known strain in a dataset, we identified its corresponding predicted strain as the predicted strain that shared the largest number of SNPs with this known strain. One and only one predicted strain was selected for each known strain, and vice versa. We then calculated how many percent of the SNPs were predicted in the corresponding predicted strain (recall), how many percent of the SNPs in a predicted strain were in the corresponding known strain (precision). We averaged the recall and precision across all known strains in a dataset to obtain the final recall and precision in this dataset. The F1 score was then calculated as $\frac{2*precision*recall}{precision+recall}$.

To assess the tool performance on the predicted strain abundance on datasets with known strain abundance, we calculated the mean absolute difference (MAE) of the abundance of the predicted

strains and the abundance of their corresponding known strains in every dataset (Figure 2). Assume there were n datasets, and m_i strains in the i-th dataset for i from 1 to n. Assume the known abundance were $a_1, a_2, \ldots, a_{m_i}$ and the corresponding predicted abundance were $b_1, b_2, \ldots, b_{m_i}$ in the i-th dataset. Then, the MAE in the i-th dataset is calculated as $MAE = (\sum_{j=1}^{m_i} |a_j - b_j|)/m_i$. When there were more predicted strains than the known strains in the i-th dataset, m_i predicted strains that share most SNPs to the known strains were selected, with one and only one predicted strain for every known strain. When there were fewer predicted strains than the known strains in the dataset, each predicted strain was assigned to a different known strain that shared most SNPs and the abundance was set to be 0 for unpredicted known strains.

We created the 108 simulated datasets with the tool dwgsim [114] and three randomly selected bacterial reference genomes, namely *Bartonella clarridgeiae* (NC_014932), *Enterococcus casseliflavus* (NC_020995) and *Methanobrevibacter smithii* (NC_009515). Each dataset was for a specific reference genome, with 2 to 4 strains, a specified relative strain abundance, and a coverage from 50X to 500X. See the commands used to simulate the data in Supplementary File S2. Two types of simulated datasets, 36 unshared and 72 shared datasets (Supplementary Tables S1-S5), were generated. The strains did not share any SNP in each unshared dataset, while at least two strains shared a specified portion of their SNPs (10% to 30%) in every shared dataset. In other words, the shared datasets consider the evolutionary relationship among strains, while the unshared datasets do not consider such an evolution relationship.

We collected 196 experimental datasets from two previous studies: 195 datasets from an investigation on mixed *Mycobacterium tuberculosis* infections [51] and a pooled dataset from the genomics analysis of *Staphylococcus epidermidis* in infant gut microbiome [43]. We call the datasets the MTB datasets and the Staph dataset, respectively. These datasets can be retrieved with the provided reference numbers in Supplementary Tables S6 and S7. The MTB datasets were the single-species-based datasets, in each of which reads from multiple strains of one species were mixed. The Staph dataset was the typical metagenomics dataset, where multiple strains of different species were mixed. In each MTB dataset, the number and abundance of strains are inferred previously by two computational methods, while the SNPs in strains are unknown [51]. We evaluated how well the tools predict the number and abundance of strains on the MTB datasets. The Staph dataset consisted of paired-end reads of 100 base pairs in 18 samples (~1.5G base pairs per sample). A previous study assembled the reads and obtained two close-to-complete strain genomes and one partially reconstructed strain genome [43]. Compared with the reference *Staphylococcus epidermidis* genome NZ_CP035288.1, there are 7342, 10836, and 76 SNPs in these three assembled strains, respectively. We trimmed the adaptor sequences in the raw reads, filtered low-quality reads with the tool Trimmomatic [115] and then mapped the reads to the reference genome NZ_CP035288.1 with the SAMtools [116]. The read coverage of the reference genome is about 830, and the average coverage of the unique SNPs in the three strains is 304, 296, and 271, respectively. Because the reference genome coverage is close to the sum of the coverage of the three strains, there are likely three strains in this dataset. We thus evaluated the tools on the predicted SNPs and strain abundance in this dataset.

Data Availability

All codes used for the generation, procession of the simulation data, including those used for the creation of the figures are available at https://github.com/UCF-Li-Lab/StriainToolSurvey. Since the size of the simulated data (184 GB) was too large to upload online, we provided one shared dataset and one unshared dataset at https://doi.org/10.6084/m9.figshare.18092846, together with the code to generate all simulated data. The accession number to the experimental datasets were provided in the Supplementary Tables S6 and S7.

Tool comparison on simulated datasets

We studied the precision, recall and F1 scores of the predicted SNPs in strains by the six tools on simulated datasets (Supplementary Tables S3 and S4). We found that mixtureS, StrainFinder and BHap were the tools that consistently showed higher precision, recall and F1 scores compared with EVORhA, MetaSNV and StrainPhlAn on all simulated datasets (Figure 2). The three tools perform better because they had more accurate strain numbers to reconstruct the strains, with more accurate prediction of the strain numbers by BHap and mixtureS and the actual strain number input into StrainFinder. MixtureS had a larger F1 score than other tools on unshared datasets, while BHap and StrainFinder performed similarly, with StrainFinder slightly better. MixtureS and StrainFinder had a similar performance on the shared datasets, slightly better than BHap. Note that the better performance of StrainFinder relies on the input of the actual strain number to predict strains, while other tools do not require such an input.

Figure 2: Comparative performance of tools on unshared and shared simulated datasets. (**A**) Precision, recall and F1 scores of the predicted strain SNPs. (**B**) Comparison of predicted strain abundance with known strain abundance by MAE. (**C**) The difference between the number of known strains and the number of predicted strains. (**D**) The number of datasets where the tools correctly predicted the strain number.

We also studied the predicted strain abundance on the simulated datasets (Figure 2 and Supplementary Tables S3, S4, and S5). We calculated the MAE of the abundance of the predicted strains and the abundance of their corresponding known strains (Figure 2). MixtureS and EVORhA predictions were four times closer to the true abundances compared to either BHap and StrainFinder, and 8-10 times closer when compared to MetaSNV and StrainPhlAn abundance predictions. While the other tools could predict multiple strains, MetaSNV and StrainPhlAn could predict only one strain no matter how many strains there were in the datasets.

We studied the predicted strain numbers by different tools as well (Supplementary Tables S3 and S4). As we mentioned above, we input the actual strain numbers to run StrainFinder since it cannot predict the strain number. We thus did not include StrainFinder in this comparison. BHap, EVORhA, MetaSNV, mixtureS, and StrainPhlAn predicted the correct strain numbers in 16, 7, 0, 33, and 0 of the 36 unshared datasets, respectively. Similarly, the five tools predicted the correct strain numbers in 22, 10, 0, 45, and 0 of the 72 shared datasets, respectively. It is thus clear that BHap and mixtureS predicted the correct strain numbers in much more datasets.

Tool comparison on experimental datasets

We evaluated the six tools on the MTB datasets and the Staph dataset. Since we had the approximate strain number and abundance in each MTB dataset from a previous study [51], we

could evaluate how well the tools predicted the correct strain number and abundance on the MTB datasets. On the Staph dataset, two S. epidermis strains were assembled more completely [43]. We thus could evaluate the abundance and SNPs on the Staph dataset.

We studied the predicted strain numbers on the MTB datasets. BHap, EVORhA and mixtureS predicted the correct strain number in 22, 0, and 77 of the 195 datasets, respectively. As mentioned above, StrainPhlAn and MetaSNV could not predict the strain number and always output one strain in every dataset. For StrainFinder, we input the actual strain number to test it, as it could not predict the correct strain number. It was thus evident that BHap and mixtureS performed better in predicting the strain numbers (Table 2).

Table 2: Tool performance on the 195 MTB experimental datasets.

We also studied the predicted strain abundance on the MTB datasets. For each known strain in a dataset, we identified its corresponding predicted strain, which had the most similar abundance among all predicted strains. Because a predicted strain may have abundance most similar to the abundance of multiple known strains, we assigned each known strain a different predicted strain. With each pair composed of a known strain and its corresponding predicted strain, we calculated MAE again as the largest difference of the abundance of the corresponding strains in a dataset (Table 2). We found mixtureS predicted the correct strain numbers in more datasets and had the smallest MAE on average across the 195 datasets. Although MetaSNV had the second smallest average MAE, its better performance was because there were two strains, with one more abundant than the other, in the vast majority of the 195 datasets here, which was in favor of predicting one strain with the relative abundance of 1 (Supplementary Table S6 and Figure S1).

We also studied the tools on the Staph dataset [43] (Table 3 and Supplementary Table S7 and Figure S2). Since MetaSNV and StrainPhlAn did not reconstruct strain genomes and Velvet used in the BHap package failed to construct the contigs, we only studied the predicted strains from EVORhA, mixtureS, and Strainfinder. We only considered SNPs that occurred in the three assembled strains and supported by the mapped reads to the reference genome for easy comparisons. Since there were three strains, we ran StrainFinder with the specified strain number as three. In contrast, we ran mixtureS and EVORhA without the specified strain number. We matched a known strain with a predicted strain that shared the most SNPs, and a predicted strain was assigned to at most one known strain. EVORhA and mixtureS predicted 9 and 3 strains, respectively. The three predicted strains by mixtureS shared much more SNPs with the corresponding known strains. For instance, the three predicted strains by mixtureS had 91.06%, 79.72% and 90.0% of SNPs in the corresponding known strains, while the known strains had 38.86%, 64.53%, and 83.33% of SNPs in the corresponding predicted strains by mixtureS. None of the three methods predicted the strain abundance well due to the challenging task of distinguishing three strains with similar abundance here.

Table 3: Comparison on the Staph dataset.

Discussion and future directions

We surveyed twenty tools on bacterial strain analysis. We classified these tools into three categories: known-strain-based, local novel-strain-targeting, and genome-scale novel-strain-targeting. The tools in the first category are useful for studying the presence of known strains and their abundance, where StrainSifter is a good choice for determining the source of the strains, and PanPhlAn can study strain accessory genomes. Note that because these tools rely on the known strain information, unknown strains in datasets could be left undetected or unreliably assigned. With bacterial genomes evolving, one may thus wish to investigate novel strains in samples. In this case, the tools in the last two categories are the right choice. The tools from the second category are faster and provide a global understanding of different species and strains in datasets, with StrainPhlAn as a popular choice. The tools from the third category provide more detailed analyses and understanding of strains than those in the second category. In this category, mixtureS has better accuracy on individual samples, while StrainFinder works well on multiple samples when provided with the correct strain number as input.

We considered known-strain-based tools as those relying on known strain information to predict strains. This consideration did not prevent the known-strain-based tools from predicting novel strains. Indeed, certain tools in the first category claimed that they can identify novel strains. For instance, StrainSeeker can detect the existence of new strains based on the difference between the observed and expected k-mer counts in an evolutionary tree [77]. PanPhlAn can actually define a new strain with the gene presence/absence profiles [80]. Because these tools depend on their known strain databases, we distinguished them from the novel-strain-targeting tools, which require only the species genome or marker genes instead of strain information.

Although dozens of tools are developed for bacterial strain analysis, no method and tool can work without a reference. In other words, we still do not have an effective way to study the strains of unknown species in metagenomics. Towards this goal, DESMAN tries to assemble reads into MAGs and uses MAGs as the alternative reference [84]. DESMAN indeed shows the hope to address this problem while clearly indicating the limitation of using MAGs. It has to use multiple MAGs to represent the E. coli genome, which cannot be done without prior knowledge of the reference E. coli genome. Some methods claim to be capable of constructing strain genomes or strain genes, whose efficiency is not widely tested or accepted [29].

In addition, the current tools for novel strain analyses focus on SNPs and essentially neglect the accessory genes in different strains. The accessory genes commonly exist in different strains of the same species, with certain genes only occurring in one strain. Current methods consider mainly the core genome or core genes of the reference species. PanPhlAn and MIDAS consider the accessory genes. They can only work with species and strains with known accessory genes [29, 84]. DESMAN mentions the accessory genes while not demonstrating the possibility of inferring accessory genes. Moreover, the quality of the MAGs used in DESMAN will likely affect the quality of the inference [84]. In the future, one has to consider the SNPs in the core genomes together with the de novo assembly of reads and more advanced computational methods to address this issue [117-119].

Almost all tools assume that different strains of the same species have different abundance in a sample. This assumption may be unmet. When a large number of strains of a species exist, different strains of the same species may have the same or similar abundance in individual

samples. To reliably distinguish these strains, we must consider multiple samples with the assumption that there are a limited number of strains and different subsets of them occur in different samples under consideration. In other words, methods that can work with multiple samples may distinguish strains with similar abundance in individual samples and provide more accurate reconstruction of bacterial strains in the future.

Finally, although it is the trend to consider multiple samples in the analyses, we need to be aware of the sample-specific variations. In other words, when analyzing strains in multiple samples, we expect that the same strains may have slightly different SNPs in different samples. Such a difference of the same strain in different samples may seem strange. But it is still an open question to define what a strain is and which two sequences should be considered two different strains instead of one strain. We hope we can define strains better in the future.

Key points:

- Twenty tools on bacterial strain genome analysis are surveyed.
- Six tools, including all tools on de novo bacterial strain genome reconstruction, are assessed on simulated and experimental datasets.
- Although some tools can predict the strain number and abundance quite well, none can completely predict the actual strain variations in bacterial strains.
- Future reconstruction of bacterial strain genomes may consider accessory genes and bacterial strains with similar abundance in multiple metagenomic samples.

Biographical Notes

- **Minerva Fatimae Ventolero** is a graduate student from Burnett School of Biomedical Science, University of Central Florida. She mainly works on metagenomics.
- **Saidi Wang** is a graduate student from the Department of Computer Science, University of Central Florida. He mainly works on chromatin interactions and metagenomics.
- **Haiyan Hu** is an Associate Professor from the Department of Computer Science, University of Central Florida. She works on miRNAs, epigenomics, and gene transcriptional regulation.
- **Xiaoman Li** is an Associate Professor from Burnett School of Biomedical Science, University of Central Florida. He works on chromatin interactions and metagenomics.

Supplementary Data

Supplementary data are available online at xxxx

Authors' contributions

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

Conflict of Interest

We declare that there is no conflict of interest regarding the publication of this article.

Funding

This work was supported by the National Science Foundation (1661414, 2015838, 2120907).

References:

1. Tyson GW, Chapman J, Hugenholtz P et al. Community structure and metabolism through reconstruction of microbial genomes from the environment, Nature 2004;428:37-43.
2. Venter JC, Remington K, Heidelberg JF et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea, Science 2004;304:66.
3. Eckburg PB, Bik EM, Bernstein CN et al. Diversity of the Human Intestinal Microbial Flora, Science 2005;308:1635.
4. Margulies M, Egholm M, Altman WE et al. Genome sequencing in microfabricated high-density picolitre reactors, Nature 2005;437:376-380.
5. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples, Nature Reviews Genetics 2005;6:805-814.
6. Tringe SG, von Mering C, Kobayashi A et al. Comparative Metagenomics of Microbial Communities, Science 2005;308:554.
7. Martín HG, Ivanova N, Kunin V et al. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities, Nature Biotechnology 2006;24:1263-1269.
8. Gill SR, Pop M, DeBoy RT et al. Metagenomic Analysis of the Human Distal Gut Microbiome, Science 2006;312:1355.
9. Poinar HN, Schwarz C, Qi J et al. Metagenomics to Paleogenomics: Large-Scale Sequencing of Mammoth DNA, Science 2006;311:392.
10. Strous M, Pelletier E, Mangenot S et al. Deciphering the evolution and metabolism of an anammox bacterium from a community genome, Nature 2006;440:790-794.
11. Woyke T, Teeling H, Ivanova NN et al. Symbiosis insights through metagenomic analysis of a microbial consortium, Nature 2006;443:950-955.
12. Warnecke F, Luginbühl P, Ivanova N et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite, Nature 2007;450:560-565.
13. Eisen JA. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes, PLoS Biol 2007;5:e82.
14. Quince C, Walker AW, Simpson JT et al. Shotgun metagenomics, from sampling to analysis, Nature Biotechnology 2017;35:833-844.
15. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly, Briefings in Bioinformatics 2019;20:1125-1136.
16. Sczyrba A, Hofmann P, Belmann P et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software, Nature Methods 2017;14:1063-1071.
17. Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads, Briefings in Bioinformatics 2020;21:584-594.
18. Bankevich A, Nurk S, Antipov D et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing, Journal of Computational Biology 2012;19:455-477.
19. Boisvert S, Raymond F, Godzaridis É et al. Ray Meta: scalable de novo metagenome assembly and profiling, Genome Biology 2012;13:R122.
20. Chapman JA, Ho I, Sunkara S et al. Meraculous: De Novo Genome Assembly with Short Paired-End Reads, PLOS ONE 2011;6:e23501.
21. Gao S, Sung W-K, Nagarajan N. Opera: Reconstructing Optimal Genomic Scaffolds with High-Throughput Paired-End Sequences, Journal of Computational Biology 2011;18:1681-1691.
22. Ghurye JS, Cepeda-Espinoza V, Pop M. Metagenomic Assembly: Overview, Challenges and Applications, The Yale journal of biology and medicine 2016;89:353-362.
23. Li D, Liu C-M, Luo R et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, Bioinformatics 2015;31:1674-1676.

24. Namiki T, Hachiya T, Tanaka H et al. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads, Nucleic Acids Research 2012;40:e155-e155.
25. Peng Y, Leung HCM, Yiu SM et al. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, Bioinformatics 2012;28:1420-1428.
26. van der Walt AJ, van Goethem MW, Ramond J-B et al. Assembling metagenomes, one community at a time, BMC Genomics 2017;18:521.
27. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs, Genome Research 2008;18:821-829.
28. Li D, Luo R, Liu C-M et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices, Methods 2016;102:3-11.
29. Cleary B, Brito IL, Huang K et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning, Nature Biotechnology 2015;33:1053-1060.
30. Yue Y, Huang H, Qi Z et al. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets, BMC Bioinformatics 2020;21:334.
31. Wang Y, Hu H, Li X. MBBC: an efficient approach for metagenomic binning based on clustering, BMC Bioinformatics 2015;16:36.
32. Wang Y, Hu H, Li X. MBMC: An Effective Markov Chain Approach for Binning Metagenomic Reads from Environmental Shotgun Sequencing Projects, OMICS: A Journal of Integrative Biology 2016;20:470-479.
33. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets, Bioinformatics 2016;32:605-607.
34. Wu Y-W, Ye Y. A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-tuples, Journal of Computational Biology 2011;18:523-534.
35. Albertsen M, Hugenholtz P, Skarshewski A et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes, Nature Biotechnology 2013;31:533-538.
36. Alneberg J, Bjarnason BS, de Bruijn I et al. Binning metagenomic contigs by coverage and composition, Nature Methods 2014;11:1144-1146.
37. Graham ED, Heidelberg JF, Tully BJ. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation, PeerJ 2017;5:e3035-e3035.
38. Imelfort M, Parks D, Woodcroft BJ et al. GroopM: an automated tool for the recovery of population genomes from related metagenomes, PeerJ 2014;2:e603-e603.
39. Kang DD, Li F, Kirton E et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies, PeerJ 2019;7:e7359-e7359.
40. Lin H-H, Liao Y-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes, Scientific Reports 2016;6:24175.
41. Lu YY, Chen T, Fuhrman JA et al. COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge, Bioinformatics 2017;33:791-798.
42. Nielsen HB, Almeida M, Juncker AS et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes, Nature Biotechnology 2014;32:822-828.
43. Sharon I, Morowitz MJ, Thomas BC et al. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization, Genome Research 2013;23:111-120.
44. Yu G, Jiang Y, Wang J et al. BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage, Bioinformatics 2018;34:4172-4179.

45. Wang Z, Wang Z, Lu YY et al. SolidBin: improving metagenome binning with semi-supervised normalized cut, Bioinformatics 2019;35:4229-4238.
46. Anyansi C, Straub TJ, Manson AL et al. Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data, Frontiers in microbiology 2020;11:1925-1925.
47. Luo C, Knight R, Siljander H et al. ConStrains identifies microbial strains in metagenomic datasets, Nature Biotechnology 2015;33:1045-1052.
48. Van Rossum T, Ferretti P, Maistrenko OM et al. Diversity within species: interpreting strains in microbiomes, Nature Reviews Microbiology 2020;18:491-506.
49. Truong DT, Tett A, Pasolli E et al. Microbial strain-level population structure and genetic diversity from metagenomes, Genome Research 2017;27:626-638.
50. Yassour M, Vatanen T, Siljander H et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability, Science Translational Medicine 2016;8:343ra381.
51. Sobkowiak B, Glynn JR, Houben RMGJ et al. Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data, BMC Genomics 2018;19:613.
52. Marx V. Microbiology: the road to strain-level identification, Nature Methods 2016;13:401-404.
53. Ali N, Dashti N, Khanafer M et al. Bioremediation of soils saturated with spilled crude oil, Scientific Reports 2020;10:1116.
54. Hou D, O'Connor D, Igalavithana AD et al. Metal contamination and bioremediation of agricultural soils for food safety and sustainability, Nature Reviews Earth & Environment 2020;1:366-381.
55. Abraham BS, Caglayan D, Carrillo NV et al. Shotgun metagenomic analysis of microbial communities from the Loxahatchee nature preserve in the Florida Everglades, Environmental Microbiome 2020;15:2.
56. Zommiti M, Feuilloley MGJ, Connil N. Update of Probiotics in Human World: A Nonstop Source of Benefactions till the End of Time, Microorganisms 2020;8:1907.
57. Ma Z, Li L, Gotelli NJ. Diversity-disease relationships and shared species analyses for human microbiome-associated diseases, The ISME Journal 2019;13:1911-1919.
58. Lloyd-Price J, Mahurkar A, Rahnavard G et al. Strains, functions and dynamics in the expanded Human Microbiome Project, Nature 2017;550:61-66.
59. Viesser JA, Sugai-Guerios MH, Malucelli LC et al. Petroleum-Tolerant Rhizospheric Bacteria: Isolation, Characterization and Bioremediation Potential, Scientific Reports 2020;10:2060.
60. Hoque E, Fritscher J. Multimetal bioremediation and biomining by a combination of new aquatic strains of Mucor hiemalis, Scientific Reports 2019;9:10318.
61. Ameen FA, Hamdan AM, El-Naggar MY. Assessment of the heavy metal bioremediation efficiency of the novel marine lactic acid bacterium, Lactobacillus plantarum MF042018, Scientific Reports 2020;10:314.
62. Colombo M, Castilho NPA, Todorov SD et al. Beneficial properties of lactic acid bacteria naturally present in dairy production, BMC Microbiology 2018;18:219.
63. Kim J-A, Bayo J, Cha J et al. Investigating the probiotic characteristics of four microbial strains with potential application in feed industry, PLOS ONE 2019;14:e0218922.
64. Nakatsuji T, Hata TR, Tong Y et al. Development of a human skin commensal microbe for bacteriotherapy of atopic dermatitis and use in a phase 1 randomized clinical trial, Nature Medicine 2021;27:700-709.
65. Lee S-H, Cho S-Y, Yoon Y et al. Bifidobacterium bifidum strains synergize with immune checkpoint inhibitors to reduce tumour burden in mice, Nature Microbiology 2021;6:277-288.

66. Su L, Zhang Y, Zhang X et al. Combination immunotherapy with two attenuated Listeria strains carrying shuffled HPV-16 E6E7 protein causes tumor regression in a mouse tumor model, Scientific Reports 2021;11:13404.
67. Pfeiffer F, Gröber C, Blank M et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing, Scientific Reports 2018;8:10950.
68. Ma X, Shao Y, Tian L et al. Analysis of error profiles in deep next-generation sequencing data, Genome Biology 2019;20:50.
69. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing, Nature Reviews Genetics 2017;18:473-484.
70. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis, Briefings in Bioinformatics 2021;22:178-193.
71. Westra ER, Sünderhauf D, Landsberger M et al. Mechanisms and consequences of diversity-generating immune strategies, Nature Reviews Immunology 2017;17:719-728.
72. Li X, Saadat S, Hu H et al. BHap: a novel approach for bacterial haplotype reconstruction, Bioinformatics 2019;35:4624-4631.
73. Pulido-Tamayo S, Sánchez-Rodríguez A, Swings T et al. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations, Nucleic Acids Research 2015;43:e105-e105.
74. Hong C, Manimaran S, Shen Y et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples, Microbiome 2014;2:33.
75. Ahn T-H, Chai J, Pan C. Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance, Bioinformatics 2014;31:170-177.
76. Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing, Nature Communications 2017;8:2260.
77. Roosaare M, Vaher M, Kaplinski L et al. StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees, PeerJ 2017;5:e3353.
78. Tamburini FB, Andermann TM, Tkachenko E et al. Precision identification of diverse bloodstream pathogens in the gut microbiome, Nature Medicine 2018;24:1809-1814.
79. Anyansi C, Keo A, Walker BJ et al. QuantTB – a method to classify mixed Mycobacterium tuberculosis infections within whole genome sequencing data, BMC Genomics 2020;21:80.
80. Scholz M, Ward DV, Pasolli E et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics, Nature Methods 2016;13:435-438.
81. Sankar A, Malone B, Bayliss SC et al. Bayesian identification of bacterial strains from sequencing data, Microbial Genomics 2016;2.
82. Zolfo M, Tett A, Jousson O et al. MetaMLST: multi-locus strain-level bacterial typing from metagenomic samples, Nucleic Acids Research 2017;45:e7-e7.
83. Nayfach S, Rodriguez-Mueller B, Garud N et al. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography, Genome Research 2016;26:1612-1625.
84. Quince C, Delmont TO, Raguideau S et al. DESMAN: a new tool for de novo extraction of strains from metagenomes, Genome Biology 2017;18:181.
85. Tan C, Cui W, Cui X et al. Strain-GeMS: optimized subspecies identification from microbiome data based on accurate variant modeling, Bioinformatics 2018;35:1789-1791.
86. Wang S, Jiang Y, Li S. PStrain: an iterative microbial strains profiling algorithm for shotgun metagenomic sequencing data, Bioinformatics 2020;36:5499-5506.
87. Costea PI, Munch R, Coelho LP et al. metaSNV: A tool for metagenomic strain level analysis, PLOS ONE 2017;12:e0182392.

88. Smillie CS, Sauk J, Gevers D et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation, Cell Host & Microbe 2018;23:229-240.e225.
89. Li X, Hu H, Li X. mixtureS: a novel tool for bacterial strain genome reconstruction from reads, Bioinformatics 2020.
90. Pasolli E, De Filippis F, Mauriello IE et al. Large-scale genome-wide analysis links lactic acid bacteria from food with the gut microbiome, Nature Communications 2020;11:2610.
91. Ghensi P, Manghi P, Zolfo M et al. Strong oral plaque microbiome signatures for dental implant diseases identified by strain-resolution metagenomics, npj Biofilms and Microbiomes 2020;6:47.
92. Goltsman DSA, Sun CL, Proctor DM et al. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome, Genome Research 2018;28:1467-1480.
93. Zolfo M, Asnicar F, Manghi P et al. Profiling microbial strains in urban environments using metagenomic sequencing data, Biology Direct 2018;13:9.
94. Brooks B, Olm MR, Firek BA et al. Strain-resolved analysis of hospital rooms and infants reveals overlap between the human and room microbiome, Nature Communications 2017;8:1814.
95. Tett A, Pasolli E, Farina S et al. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis, npj Biofilms and Microbiomes 2017;3:14.
96. Petersen TN, Lukjancenko O, Thomsen MCF et al. MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads, PLOS ONE 2017;12:e0176469.
97. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors, Proceedings of the National Academy of Sciences of the United States of America 1977;74:5463-5467.
98. Nimmo C, Shaw LP, Doyle R et al. Whole genome sequencing Mycobacterium tuberculosis directly from sputum identifies more genetic diversity than sequencing from culture, BMC Genomics 2019;20:389.
99. Shockey AC, Dabney J, Pepperell CS. Effects of Host, Sample, and in vitro Culture on Genomic Diversity of Pathogenic Mycobacteria, Frontiers in genetics 2019;10:477-477.
100. Kyrgyzov O, Prost V, Gazut S et al. Binning unassembled short reads based on k-mer abundance covariance using sparse coding, GigaScience 2020;9.
101. Truong DT, Franzosa EA, Tickle TL et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling (vol 12, pg 902, 2015), Nat Methods 2016;13:101-101.
102. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2, Nat Methods 2012;9:357-359.
103. Asnicar F, Thomas AM, Beghini F et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0, Nature Communications 2020;11:2500.
104. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools, Bioinformatics 2009;25:2078-2079.
105. Ciccarelli FD, Doerks T, von Mering C et al. Toward Automatic Reconstruction of a Highly Resolved Tree of Life, Science 2006;311:1283.
106. Lan Y, Rosen G, Hershberg R. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains, Microbiome 2016;4:18.
107. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference, Genome Biology 2008;9:R151.
108. Cai X, Hu H, Li XS. Tree Gibbs Sampler: identifying conserved motifs without aligning orthologous sequences, Bioinformatics 2007;23:2013-2014.
109. Geman S, Geman D. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, IEEE Transactions on Pattern Analysis and Machine Intelligence 1984;PAMI-6:721-741.

110. Murillo GH, You N, Su X et al. MultiGeMS: detection of SNVs from multiple samples using model selection on high-throughput sequencing data, Bioinformatics 2016;32:1486-1492.
111. McKenna A, Hanna M, Banks E et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Research 2010;20:1297-1303.
112. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 1977;39:1-22.
113. Li L, Cheng ASL, Jin VX et al. A mixture model-based discriminate analysis for identifying ordered transcription factor binding site pairs in gene promoters directly regulated by estrogen receptor-α, Bioinformatics 2006;22:2210-2216.
114. Homer N. DWGSIM: Whole Genome Simulator for Next-Generation Sequencing. https://github.com/nh13/DWGSIM.
115. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 2014;30:2114-2120.
116. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools, Bioinformatics 2009;25:2078-2079.
117. Talukder A, Barham C, Li X et al. Interpretation of deep learning in genomics and epigenomics, Briefings in Bioinformatics 2021;22:bbaa177.
118. Talukder A, Saadat S, Li X et al. EPIP: a novel approach for condition-specific enhancer–promoter interaction prediction, Bioinformatics 2019;35:3877-3883.
119. Sarker IH. Machine Learning: Algorithms, Real-World Applications and Research Directions, SN Computer Science 2021;2:160.