Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data

Anyansi, Christine; Straub, Timothy J.; Manson, Abigail L.; Earl, Ashlee M.; Abeel, Thomas

DOI
10.3389/fmicb.2020.01925

Publication date
2020

Document Version
Final published version

Published in
Frontiers in Microbiology

Citation (APA)
Anyansi, C., Straub, T. J., Manson, A. L., Earl, A. M., & Abeel, T. (2020). Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data. *Frontiers in Microbiology*, *11*, 1-17. Article 1925. https://doi.org/10.3389/fmicb.2020.01925

frontiers in Microbiology

REVIEW
published: 18 August 2020
doi: 10.3389/fmicb.2020.01925

Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data

*Christine Anyansi[1,2], Timothy J. Straub[2,3], Abigail L. Manson[2], Ashlee M. Earl[2] and Thomas Abeel[1,2]**

[1] Delft Bioinformatics Lab, Delft University of Technology, Delft, Netherlands, [2] Infectious Disease and Microbiome Program, Broad Institute of MIT and Harvard, Cambridge, MA, United States, [3] Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, United States

Metagenomic sequencing is a powerful tool for examining the diversity and complexity of microbial communities. Most widely used tools for taxonomic profiling of metagenomic sequence data allow for a species-level overview of the composition. However, individual strains within a species can differ greatly in key genotypic and phenotypic characteristics, such as drug resistance, virulence and growth rate. Therefore, the ability to resolve microbial communities down to the level of individual strains within a species is critical to interpreting metagenomic data for clinical and environmental applications, where identifying a particular strain, or tracking a particular strain across a set of samples, can help aid in clinical diagnosis and treatment, or in characterizing yet unstudied strains across novel environmental locations. Recently published approaches have begun to tackle the problem of resolving strains within a particular species in metagenomic samples. In this review, we present an overview of these new algorithms and their uses, including methods based on assembly reconstruction and methods operating with or without a reference database. While existing metagenomic analysis methods show reasonable performance at the species and higher taxonomic levels, identifying closely related strains within a species presents a bigger challenge, due to the diversity of databases, genetic relatedness, and goals when conducting these analyses. Selection of which metagenomic tool to employ for a specific application should be performed on a case-by case basis as these tools have strengths and weaknesses that affect their performance on specific tasks. A comprehensive benchmark across different use case scenarios is vital to validate performance of these tools on microbial samples. Because strain-level metagenomic analysis is still in its infancy, development of more fine-grained, high-resolution algorithms will continue to be in demand for the future.

Keywords: metagenomics, microbial detection, strain-level classification, methods review, whole genome sequencing, bioinformatics

OPEN ACCESS

Edited by:
Fumito Maruyama, Hiroshima University, Japan

Reviewed by:
Lu Fan, Southern University of Science and Technology, China
So Nakagawa, Tokai University, Japan

***Correspondence:**
Thomas Abeel
t.abeel@tudelft.nl;
thomas@abeel.be

Specialty section:
This article was submitted to Infectious Diseases, a section of the journal Frontiers in Microbiology

Received: *27 February 2020*
Accepted: *22 July 2020*
Published: *18 August 2020*

Citation:
Anyansi C, Straub TJ, Manson AL, Earl AM and Abeel T (2020) Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data. Front. Microbiol. 11:1925. doi: 10.3389/fmicb.2020.01925

INTRODUCTION

Within a species, bacteria can be highly diverse in terms of their virulence, resistance to antibiotics, geographical transmission patterns, and other phenotypic characteristics (Fournier et al., 2014; Maxson and Mitchell, 2016). Individual strains can vary greatly with respect to pathogenicity, treatment options, transmissibility, and growth rate (Balmer and Tanner, 2011; Alizon et al., 2013).

In order to effectively treat patients, study bacterial population dynamics, conduct epidemiological surveillance, and stem outbreaks, it is critical to identify which specific strains of a species present in a sample (Fournier et al., 2014; Deurenberg et al., 2017). Tracking and comparing individual strains shared across sets of samples would allow for the assessment of the evolution of population diversity in longitudinal samples within a patient or other host system. The ability to identify specific strains in a noisy background of other organisms present in a metagenomic sample could allow for improved tracking of strains involved in an outbreak across a population.

Accurately identifying specific pathogenic strains would aid in patient diagnosis, allowing for personalized treatment regimens, improved treatment outcomes, and a reduction in the spread of antibiotic resistance. Mixed infections, defined as infections caused by multiple strains of a single pathogen species (Marshall, 2002; Cohen et al., 2012), represent an underappreciated challenge to understanding infections and have been described for at least 22 bacterial species (Balmer and Tanner, 2011), including *M. tuberculosis* (Cohen et al., 2012; Plazzotta et al., 2015), *C. difficile* (Eyre et al., 2013, 2012), and *Streptococcus pneumoniae* (Esposito et al., 2002; Minagawa et al., 2008). It is estimated that 10–20% of *M. tuberculosis* patients in high risk areas (Huang et al., 2010; Navarro et al., 2011; Plazzotta et al., 2015) and 10% of *Staphylococcus aureus* (Lessing et al., 1995; Cespedes et al., 2005) patients are infected with multiple pathogenic strains. Mixed infections put patients at a higher risk of treatment failure (Balmer and Tanner, 2011; Cohen et al., 2012; Plazzotta et al., 2015), as strains with different drug susceptibility and antibiotic resistance profiles (Falagas et al., 2008; El-Halfawy and Valvano, 2015) can complicate diagnosis and identification of the optimal treatment regimen (Balmer and Tanner, 2011). In addition to poor treatment outcomes, mixed strain infections can increase pathogen virulence due to selective pressure within the host (Frank, 1996). Accurate classification of individual strains is critical for identifying mixed infections and will help determine proper treatment options for patients with complex infections, track transmission of pathogenic strains in a population, and differentiate between reinfection and intra-host pathogen evolution.

While there is clearly substantial value in being able to pinpoint individual strains within metagenomic samples, most current widely used tools for metagenomic analysis only allow for an assessment of composition at the genus or species level, not the strain level. For example, the current most popular metagenomics taxonomic classification programs, including Kraken (Wood et al., 2014) MetaPhlAn2 (Truong et al., 2015) and GTDB-Tk (Chaumeil et al., 2019), are capable of identifying mixed populations only at the species or genus level–not at the individual strain level within a species. Tools capable of conducting classification of metagenomic samples for higher taxonomic levels such as the family, genus, or species have been previously reviewed (Hunter et al., 2012; Mande et al., 2012; Teeling and Glöckner, 2012; Goldman and Domschke, 2014). In contrast, tools to detect taxonomy at a finer-grained taxonomic levels within metagenomic samples – targeting specific strains within a species – are still in their infancy (Marx, 2016; Segata, 2018), with most tools only published within the past 5 years.

To date, there have been no reviews focused on strategies to computationally classify heterogeneous bacterial populations using WGS data at the level of specific strains within a species. This literature review gives an overview of recent methods for classification at the intra-species, or strain level, including methods based on WGS data to identify both specific strains, as well as mixes of strains. These tools are divided into assembly based, alignment based, and reference free methods. We have included both secondary sources (reviews or methods papers) and original research, where the main objective is developing a novel methodology for detecting heterogeneous bacterial communities, e.g., mixed infections or within host evolution. The majority of these tools operate using short-read sequencing data, due to the abundance and affordability of the Illumina platform. However, the advent of both long-read sequencing and single-cell sequencing holds great promise in enabling effective strain-level identification. We also cover the few presently existing metagenomic tools specifically made for these sequencing platforms in this review. Although we focus on clinical applications here, the methods discussed are applicable to a broad range of biological ecosystems typically analyzed using metagenomics, including soil, wastewater or other environments. We discuss appropriate applications of each strategy, evaluation of these strategies in literature, as well as the applicability of these algorithms to health and disease.

APPROACHES FOR DETECTING INDIVIDUAL STRAINS OF BACTERIA WITHIN A SPECIES

Currently available approaches to classifying genetically distinct populations from a sequencing read set can be binned into three categories (see **Table 1**): (i) methods using (metagenomic) assembly or *de novo* reconstruction of genomes within the sample (assembly based), (ii) aligning genomes to a reference database (including full genome alignment based and pattern based), and (iii) reference database free approaches that rely on applying statistics directly to allele (variant) frequencies.

Assembly Based Approaches for *de novo* Strain Level Reconstruction

Assembly based approaches attempt to identify individual strains in a mixture by performing (whole) genome assembly, drawing on tools developed for haplotype (single clone or strain) reconstruction in diploid species. To obtain an accurate reconstruction there must be a sufficient number of sites that differ between the component strains in order to separate or cluster variants into distinct strains (Yuan et al., 2012; Votintseva et al., 2017). Therefore, accurate reconstruction of distinct strains requires sufficient read length to capture overlap between reads, enough discriminating sites to separate populations, and the presence of at least one variant site in most reads. **Figure 1** gives

TABLE 1 | Tool benchmark and technical details.

Author	Method name	Type[1]	Technical details[2]	Sample benchmarks[3]	Test metrics[4]	Required coverage level per strain[5]
Pulido-Tamayo et al., 2015	EVORha	assembly based	– java	– *E. coli* time series (lab grown) – *C. difficile* mixed infection samples	reliability score, mean absolute error, rmse	50× coverage
Quince et al., 2017	DESMAN	assembly based	– git/python – linear runtime – 5 strains in 117 min	– fecal metagenome samples – community of 100 species and 210 strains with 96 samples (synthetic)	accuracy	–
Ahn et al., 2015	Sigma	alignment based	– C++ – scaled for supercomputers (alignment with 10,000 cores takes 10 min) – sample with 5 strains takes 20 h and 62GB RAM on a computer with 64CPU	– fecal metagenome dataset – numerous spike ins of fecal set to simulate outbreaks	accuracy, TP/FP	0.02× coverage
Sankar et al., 2015	BIB	alignment based	– 1 million reads in 10 min on single CPU – git/python	– mixtures of 2–6 staphylococcus strains (synthetic) – *S. aureas* sample data	absolute error	
Francis et al., 2013; Byrd et al., 2014; Hong et al., 2014	Pathoscope	alignment based	– git/BioConda – 1 sample using 16 CPU and 256GB RAM took 17 min	– European *E. coli* outbreak 2011 (O104:H4) – mixed read datasets of 3 strains	TP/FP	20% genome coverage
Fischer et al., 2017	DiTASiC	alignment based	– git/conda – requires R and python	– 3 simulated set groups – low, medium, and high complexity metagenomic benchmark datasets (synthetic) – lacks real world testing	sum of squared errors, TP/FP/FN/FP –	–
Huson et al., 2007	MEGAN	alignment based	– gui/java – took 180 h using 64CPU for 300 k reads	– Sargasso sea dataset – mammoth bone – simulation studies – mostly species level testing	FP	–
Dilthey et al., 2019	MetaMaps	alignment based	– git/Perl – takes 16–210 h using 262GB RAM – cannot make own DB	– simulated data – human microbiome project data (PacBio, species) – Zymo synthetic community (Oxford Nanopore Technology)	Precision, recall	–
Smillie et al., 2018	StrainFinder	pattern based[6]	– git/python – 100 samples across 649 reference genomes using 100–200cores takes 48 + hours – needs alignment file with some preprocessing as input	– 2–32 strains across 2–64 samples (synthetic) – recurrent *C. difficile* infection over time	Unifrac distance	25×
Gan et al., 2016		pattern based[6]	– not available	– TB datasets	–	1× coverage

(Continued)

TABLE 1 | Continued

Author	Method name	Type[1]	Technical details[2]	Sample benchmarks[3]	Test metrics[4]	Required coverage level per strain[5]
Luo et al., 2015	ConStrains	pattern based[6]	– git/python – took 8.5 h and 2 GB ram on infant gut dataset – custom DB not possible	– *E. coli* admixtures 2–7 strains (synthetic) – gut microbiome time series – microbiome time series (synthetic) – cystic fibrosis patient infection data	Jenson-Shannon divergence	10× coverage
Freitas et al., 2015	GOTTCHA	pattern based[6]	– git/Perl – used 16cores and 132GB RAM while being 2–5× slower than other tools – custom DB not possible	– human microbiome project mixtures of 22 genomes – spiked air filter metagenome spiked – spiked human stool – synthetic communities of 25–300 genomes	precision, recall, F-score, false discovery rate and accuracy	
Sahl et al., 2015	WG-FAST	pattern based[6]	– conda – uses phylogeny	– fecal specimens *E. coli* O104:H4 outbreak	accuracy	1×
Roosaare et al., 2016	StrainSeeker	pattern based[6]	– online web tool – Perl/R – needs 300GB space to build DB – uses 1 cpu, 512GB RAM and took 1.1 min for classification	– *E. coli*, *K. pneumoniae*, *E. faceilius*, *S. enterica* isolate identification (synthetic)	accuracy	<1× coverage
Albanese and Donati, 2017	StrainEst	pattern based[6]	– git/docker/python – takes 12–25 min for a 10× –100× coverage sample using 129–591MB RAM and 4 cores	– paired strains from 4 species (synthetic) – 2 HMP mock communities (21 organisms) – specific strain in skin microbiome – cross sectional *E. coli* strains in stool samples – gut microbiome time series	Matthew Correlation Coefficient, Jensen-Shannon divergence	10× coverage
Truong et al., 2017	StrainPhlAn	pattern based[6]	– git/conda	– human microbiome	accuracy	2×
Nayfach et al., 2016	MIDAS	pattern based[6]	– git/docker/python – on 1CPU process 5,000 reads per second using 3 GB RAM – 1.5–2 h for typical gut metagenome	– stool metagenomes time series – marine metagenomes	(only of genes) accuracy, TP/FP	1 × coverage
Costea et al., 2017	metaSVN	pattern based[6]	– git/conda – 676 samples in 223 min using 2,488 GB RAM and 32 cores	– oral metagenome	–	5 × coverage
Tu et al., 2014	GSMer	pattern based[6]	– git/Perl scripts	– diabetes patients gut microbiome – obesity associated microbiome	TP	<0.25 × (100 GSMs) >0.25 × (50 GSMs)
Scholz et al., 2016	PanPhlAn	pattern based[6]	– git/python	– *E. coli* outbreak O104:H4 – gut microbiomes – skin microbiome – oral microbiome – marine metagenomes	F1 score	1 × coverage

(Continued)

TABLE 1 | Continued

Author	Method name	Type[1]	Technical details[2]	Sample benchmarks[3]	Test metrics[4]	Required coverage level per strain[5]
Koslicki and Falush, 2016	MetaPalette	pattern based[6]	– git/docker/python	– spiked HMP community (22 organisms) – soil metagenome	Divergence, FP	22 × coverage
Anyansi et al., 2020	QuantTB	pattern based[6]	– git/python – <10 min for single sample using single core and pre-build database	– TB datasets	precision, recall, F-score, FP/TP	–
Eyre et al., 2013		reference db free	– R script in supplements	– *C. difficile* infected patients	RMSE	–
O'Brien et al., 2016	pfmix	reference db free	– R – for a 5 strain sample takes 10 min on single core	– blood from malaria patients	Mean squared error	25 reads
Assefa et al., 2014	*estMOI*	reference db free	– git/Perl – little documentation	– clinical isolates of *P. falciparum*	accuracy	30 × coverage
Zhu et al., 2017	DEploid	reference db free	– R package – 1–6 h	– clinical isolates *of P. falciparum*	accuracy	1% abundance
Sobkowiak et al., 2018	MixInfect	reference db free	– R script/git – no documentation	– tested on TB samples	accuracy	10 × coverage

[1]Category of algorithm. [2]Details about the computational parameters of the tool in terms of code base/runtime/memory usage/availability. [3]Example datasets tool was tested on in paper. [4]Metrics by which each method was evaluated. [5]The required coverage for the tool per stain to perform. If no value is indicated, this indicates the particular value could not be determined from the article where the method was published. [6]Pattern based methods use a database of predefined markers to classify genetic diversity within a sample.

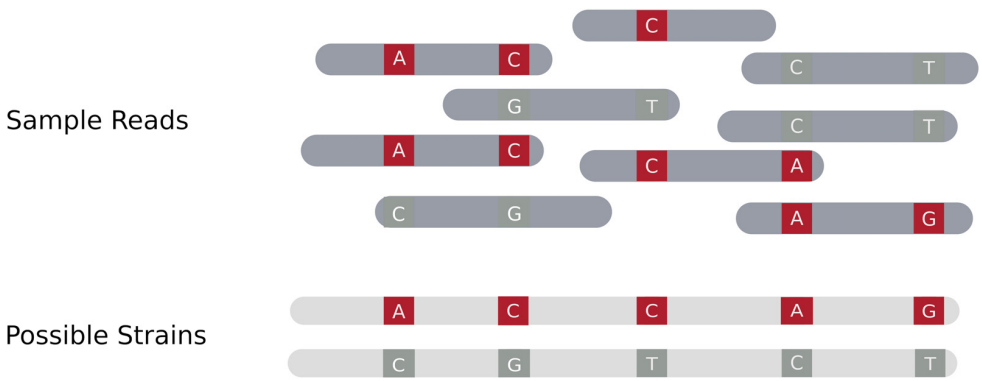


FIGURE 1 | Assembly of multiple distinct strains from a read set. The blue areas in the sample reads represent regions where the strains have identical sequence. Variant locations in the reads are denoted as red or dark gray stripes. Red variants originate from one haplotype, whereas dark gray variants originate from the other. The goal of an assembly based method is to resolve distinct strains based on the coverage and distribution of the read data, drawing on methods previously developed for resolving haplotypes.

an overview of how a read set can be resolved into a set of distinct individual strains using an assembly based procedure.

EVORhA, one of the few assembly based methods designed for reconstructing complete bacterial genomes from bulk metagenomic sequencing data, identifies strains via local haplotype assembly (**Table 1**; Pulido-Tamayo et al., 2015). For each genomic region containing a sufficient amount of genetic variation, candidate strains are first defined as individual genetically distinct combinations of polymorphisms. To filter out candidate strains that are actually sequencing errors, a minimum number of reads must support an initial candidate strain. In an extension step, candidates are merged with nearby locally constructed candidate strains, based on read frequency and overlap of polymorphism combinations. Ultimately, a mixture model is used to group extended candidate strains occurring at similar frequencies and match these together on a genome-wide level, making the read frequency ratios of observed candidate strains crucial to this method. However, this read frequency criteria for merging strains can produce chimeric strains due to the presence of subpopulations with similar frequencies, similar to a key problem encountered in phasing with whole genome assembly. Given very high coverage, sufficient frequency

diversity and sufficient segregating sites, assembly based methods such as EVORhA can resolve the full genomes of genetically distinct subpopulations and yield the most accurate strain identification results when compared to other categories of strain-level identification tools.

Knowing the full sequences of organisms within a sample then allows for comparison and tracking of strains at the highest resolution possible. As such, these methods would be suitable for observing a strain's evolutionary trajectory as well as detecting mixed infections composed of strains that are highly similar to each other. In order to estimate frequencies, a method would need to account for relative abundance of reads specific to each strain. DESMAN (Quince et al., 2017) does this by exploiting differences in read coverage between genes conserved within a species and other parts of the genome. DESMAN requires a group of metagenome assembled genomes (MAGS) to do estimate relative abundances.

A major drawback of assembly based methods is that a large amount of coverage, 50–100 × for each strain, is required to achieve an accurate reconstruction, demanding extremely high depth sequencing for strains at a low abundance within a sample (Zagordi et al., 2011). High levels of coverage are required to account for errors introduced by sequencing: each distinct strain must be sequenced with sufficient coverage in order to differentiate spurious variation from true distinct strains. Such high coverages can be achieved in studies where sample complexity is low, with typically less than 5 strains present.

Reference Database Approaches

In order to relate strains observed within a sample to previously studied genomes or species, it is necessary to use a reference database. Reference databases can vary greatly in different dimensions, such as genome quantity or species diversity. Methods employing a reference database can be broken down into two major categories: (i) approaches that have full genomes within their database, and (ii) approaches that only use subsets of these genomes within their database. Here we cover these two overarching approaches and show the pros and cons of each.

FULL GENOME ALIGNMENT BASED APPROACHES

Full genome alignment based methods (alignment methods for short) classify strains by aligning reads to a predefined set of reference genomes and applying probabilistic models to calculate a statistical measure representing the likelihood a specific read is associated with a given reference (**Figure 2** and **Table 1**; Li and Homer, 2010). These methods are often considerably faster than assembly based methods and require less coverage, some methods claim to work with less than 1× coverage. These methods can achieve such low coverages compared to assembly based methods due to their use of a reference database – where the most likely candidate is selected based on the available data using the probabilistic model. Alignment based methods share the same similarities and limitations, such as reference database composition, alignment method, and strain abundance quantification. We will discuss these similarities and limitations on the whole toward the end of this section.

Pathoscope, (Hong et al., 2014) one of the most commonly used classification pipelines for metagenomic analysis, uses different aligners three aligners [GNUMAP (Clement et al., 2009), Bowtie 2 (Langmead and Salzberg, 2012) and BLAST (Altschul et al., 1997)] to align reads to reference genomes. Scores for each alignment are converted to posterior probabilities that represent the likelihood that an alignment is the source of the read. Non-unique reads are reassigned to their nearest reference using a Bayesian mixture model which uses both the mapping scores and the proportions of non-unique reads. Another alignment based method, Sigma, allows users to choose their own short-read alignment algorithm, using Bowtie2 as a default (Ahn et al., 2015). Instead of using scores given by an aligner, Sigma computes its own probability scores for each read to originate from an alignment by examining the number of matches and mismatches between the two.

Calculation of strain abundance in alignment based approaches leverages the number of reads mapping to each reference genome. For Sigma the relative abundance of a genome is simply the proportion of aligned reads out of the total number of reads, whereas Pathoscope calculates relative abundance from the sum of the probability of reads mapped to different genomes in the reference database. BIB exploits the similarities between alignment based strain identification and the more well-established field of RNA-seq data analysis (Kim and Salzberg, 2011; Glaus et al., 2012; Langmead and Salzberg, 2012; Trapnell et al., 2012) for calculating relative abundances, by implementing the RNA-seq algorithm BitSeq (Glaus et al., 2012) within its identification pipeline to calculate relative abundances, after aligning reads to a reference database with Bowtie 2. Unlike other alignment methods, StrainFinder (Smillie et al., 2018) calculates abundances for all the genomes in the reference database using SNP frequencies after aligning reads with BWA. Because StrainFinder uses the Expectation Maximization algorithm to estimate strain frequencies, the user needs to input the expected number of strains expected to be in the sample, to ensure the best likelihood. This not only makes StrainFinder exceptionally computationally intensive, but also makes it less suitable for broad metagenomic studies with unknown number of strains.

While alignment based detection methods work well for species with clear and well-separated sub-lineages, the selection of genomes and choice of size for the reference database is critical for applications to more closely related strains. Some tools aim to draw on large and comprehensive databases in order to gain higher resolution. Sigma offers users the opportunity to define their own reference databases and claims support for up to tens of thousands of genomes. The entirety of RefSeq (2266 genomes at time of publication) has been used as the reference database for Sigma. PathoScope generates a reference database from all genome sequences in NCBI for a given query taxID. The resulting redundancy from using a taxID which could potentially include very closely related strains, instead of a database of filtered genomes such as RefSeq, ensures coverage at all genomic levels, but can result in non-specific strain

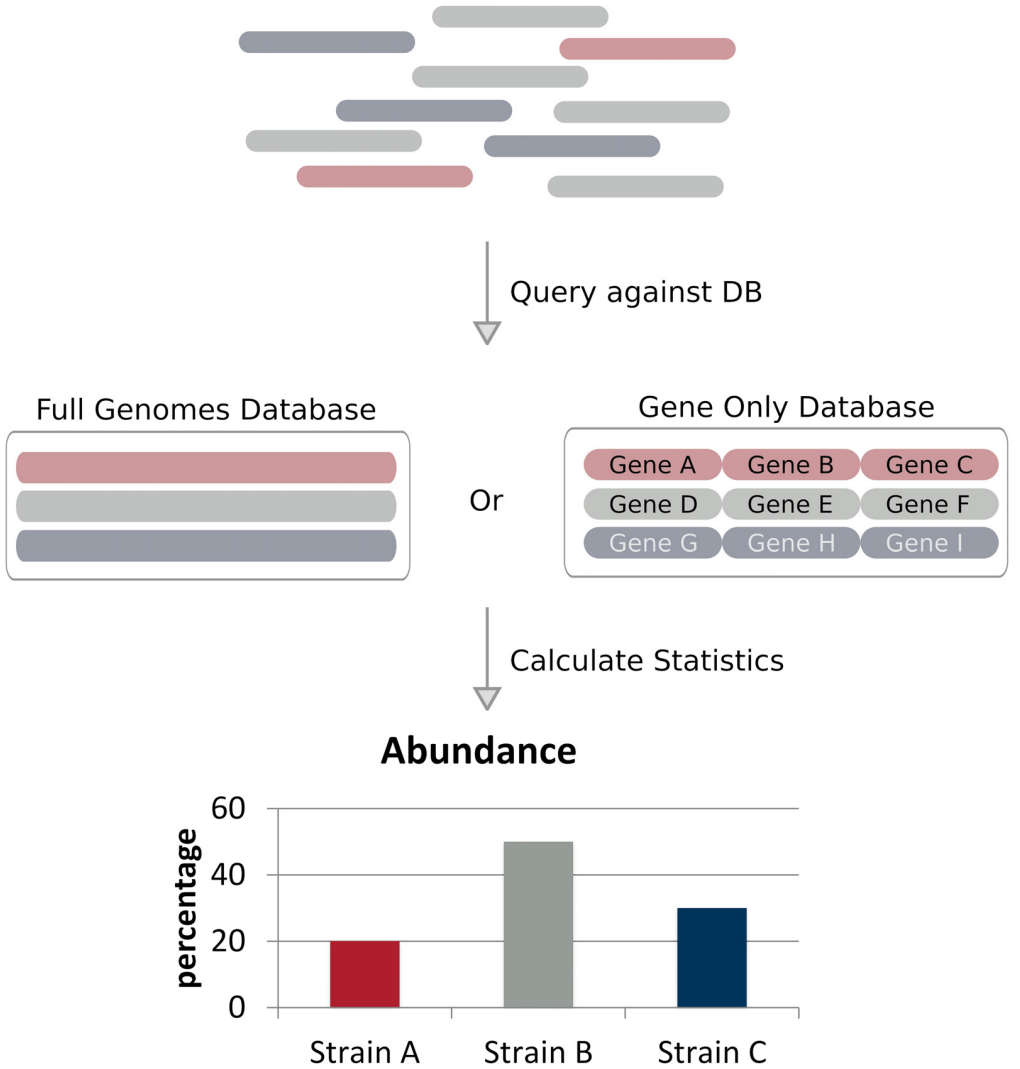


FIGURE 2 | Alignment based approaches. Reads of a sequencing dataset – where different colors denote genetically distinct strains – are aligned to a reference database of full genomes or taxonomic markers (in this case genes). Strain abundances can be estimated by the relative number of reads aligning to each reference genome.

identification calls. Even if similar sequences are excluded, it is often not practical to have a reference genome for every genetically distinct, closely related strain in a species. While a large reference database can increase coverage of intra-species diversity, it also requires a larger computational search space for matching reads. In addition, differentiating between closely related strains in a highly comprehensive reference database is nearly impossible and can result in an inflated number of false positive predictions. Removal of closely related reference genomes when using BIB improved accuracy and reduced non-specific predictions to multiple unrelated strains. Therefore, proper pruning of representative reference sequences to an appropriate level of resolution is essential.

A major drawback of alignment based methods is that they are dependent on details of the underlying alignment tool and its parameters. Different alignment methodologies can result in discordant results between methods and impacts our ability to perform comparisons between tools. For example, most alignment based methods use a short-read aligner (Hong et al., 2014; Ahn et al., 2015; Sankar et al., 2015), while DiTASiC (Fischer et al., 2017) uses the pseudo alignment approach found in Kallisto (Bray et al., 2016) used for aligning RNA seq reads. Some strain identifiers [Pathoscope, and MEGAN (Huson et al., 2007)] make predictions using the quality score of the alignment of each read. Sigma and BIB use Bowtie2 as an aligner by default which reports all reads that map in multiple locations while Pathoscope and DiTASiC (Fischer et al., 2017) post-process multi-mapping reads within their algorithm, and StrainFinder uses BWA which randomly assigns multi-mapping reads to a specific location. Sigma additionally allows users to select their own aligner. The differences between alignment methods and their impact on results have been

reported before in literature (Canzar and Salzberg, 2017). Because these strain classification methods depend on the information given via the alignment, variation at the alignment stage may have consequences throughout the entire method. Each approach can limit the ability to correctly identify strains in a sequencing set in different circumstances. The impact of these variations has not yet been characterized, but will ultimately depend on the species under examination and the parameters of the alignment method and how the classification methods employ the alignment information.

PATTERN BASED METHODS BASED ON ALIGNMENT TO GENETIC MARKERS

Methods where alignments are done to a set of genetic marker, rather than complete genomes were developed to offer decreased compute time and memory requirements. We will refer to these as pattern based methods. These methods classify genetic diversity within a sample using a database of predefined markers, such as unique genes, SNPs, genome-specific k-mers, or fluctuations in GC content. The choice of marker type can vary based on the species, data type, and classification goals. Similar to alignment based methods, pattern based identification methods require a reference database with which to "learn" parameters for their statistical models. However, pattern based methods first preprocess the reference database, extract useful features, and apply these features for a new classifier algorithm, resulting in decreased run times. New sequencing reads can then be classified based off the constructed model.

An example of a method that uses a database of universal single-copy gene families as the predefined marker set is MIDAS, which aims to provide both species and strain-level taxonomic identification. MIDAS first determines species content by aligning reads to a single-copy gene database containing a single representative genome per species (Nayfach et al., 2016). In order to determine strain-level information, reads are mapped to a pan-genome database containing genes from the species found in the first alignment step. Abundance estimation per strain is calculated by normalizing by the coverage of universal single copy gene families. However, this sort of strain level inference using variation in genes alone is not practical for discrimination purposes, because universal single-copy genes represent a smaller portion of the genome and are, by definition, conserved between strains of species (Jordan et al., 2002; Martín et al., 2003). MIDAS requires at least 1 × coverage per strain to determine the presence or absence of a gene.

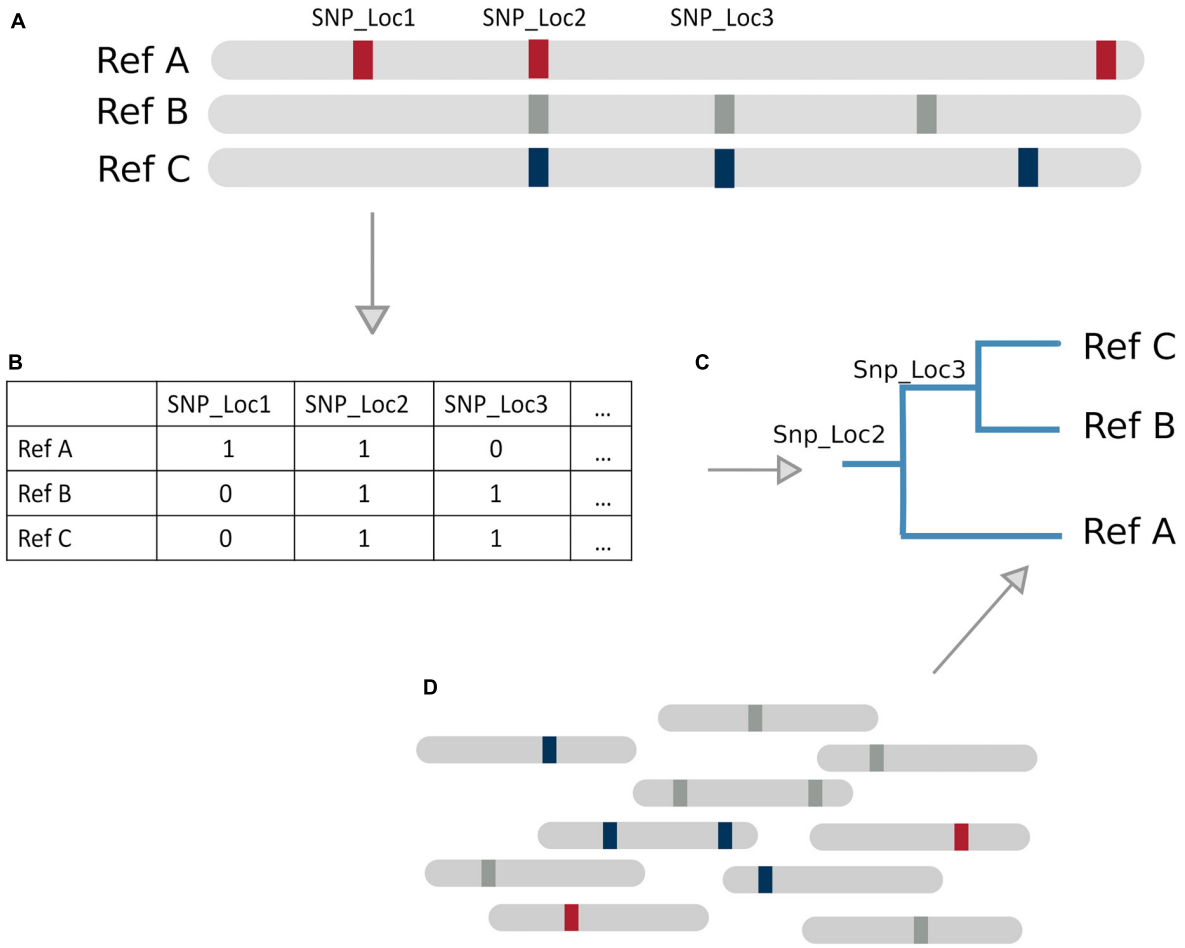
K-mers are often used in pattern based methods because unlike genes, they are sampled across the whole genome, including regions that are not especially conserved. In order to gain greater resolution than can be obtained by using only genes, GSMer identifies strains by capitalizing upon a strain-specific database of strain-specific k-mers, or GSMs (genome specific markers) (Tu et al., 2014). Each strain in the database is represented by a set of at least 50 GSMs (optimized for k-mer size and number). If a strain has fewer than 50 unique GSMs, it is not included in the database. A strain is only identified in a read set if a perfect match for all 50 GSMs of that strain is identified within the read set, resulting in a high false negative rate and an inability to identify strains not similar to those in the database. This may work well for slow evolving and well conserved organisms that will not change and can be expected to always include the set of 50 GSMers required to be identified. But not in settings where strains are diverse and quickly changing as there is a higher chance for the set 50 GSMers required to be present to have been mutated or changed due to evolutionary drift.

Phylogenetic trees complement pattern based methods by offering a more informative database structure where paths can be indexed with a series of markers leading to a presence of a particular strain. Trees also provide an intuitive visualization of the phylogenetic placement of a strain. Given the tree, these tools map k-mers or SNPs from unknown samples onto nodes within the tree to determine phylogenetic "paths," sequences of nodes, which represent presence of a particular strain in the sample. Strain abundances are calculated based on the SNP or k-mer coverage.

SNP based tree methods differ in their SNP calling, variant filtering, tree construction, and path determination techniques. Relying solely on SNPs limits the inclusion of other types of genomic variation such as indels, which could be picked up in a k-mer based method. SNP/phylogenetic hybrid methods are particularly suitable for species with low genomic divergence like *Mycobacterium tuberculosis*, because it is a clonal organism with strains differing by very few SNPs. Gan et al. (2016) and Sahl et al. (2015) (WG-FAST) have both developed tree based classification methods constructed using SNP variations between reference genomes (**Figure 3**). Another SNP based method, StrainEST (Albanese and Donati, 2017), is not based on a phylogenetic tree model but uses SNP frequencies within each genome of a reference database to predict strains based on co-occurring SNPs within a sample. This is done by modeling the SNP profile of a sample as a linear combination of the SNPs in a reference database using LASSO regression.

In contrast, k-mer based tree approaches can be more suitable for species that have larger degree of genetic variation or bigger structural variations that are not detectable by only considering SNPs. They would be less efficient at differentiating strains which are only a few SNPs apart as the impacts of a genetic sequencing error are more pronounced in the tree construction and classification process when working with k-mers. Roosaare et al. (2016) (StrainSeeker) have developed guide-tree based classification methods based on k-mers. A phylogenetic tree detailing the relationship between reference genomes must first be provided by the user.

Another kind of approach, GOTTCHA, generates a database of unique signatures for each genome at different taxonomic levels (Freitas et al., 2015). The unique signatures of a strain are the collection of all subsequences not found in any other available sequences at the desired taxonomic level. The unique signature of an unknown query sample can then be mapped against this database to determine coverage statistics for the query's unique signature. The abundance of predicted strains is obtained through a statistic comparing the total number of mapped bases to the signature for the reference, and the number



	SNP_Loc1	SNP_Loc2	SNP_Loc3	...
Ref A	1	1	0	...
Ref B	0	1	1	...
Ref C	0	1	1	...

FIGURE 3 | Tree Based Method Overview. **(A)** Example database of genomes with SNPs present as markers. **(B)** Representation of genome database, where 1 denotes a SNP and 0 absence of a SNP **(C)** SNP tree constructed based on SNPs from the database. **(D)** SNPs present in new reads can be matched against the tree to infer likely reference genome of origin by identifying sequences of successfully matching nodes (a path).

of unique bases mapped. StrainPhlAn (Truong et al., 2017) also uses species specific marker sets to classify strains, but only identifies the most abundant strain for each detected species in a metagenomic sample. The presence of other strains is assessed by calculating the number of polymorphic positions per species.

Other pattern based methods employ clustering to help delineate strains and augment pattern based detection techniques. For example, ConStrains assimilates elements of *de novo* assembly to detect genetically distinct strains (Luo et al., 2015). Reads for each species are first mapped against species-specific marker genes using MetaPhlAn2 (Segata et al., 2013) to generate a multiple alignment, and SNPs are determined using *Samtools* (Li, 2011) based on sufficient coverage criteria. The resulting SNP profiles are clustered into groups representing genetically distinct strains, with abundances calculated using a Monte-Carlo algorithm. In order to delineate strains, ConStrains requires a relatively high coverage (10×).

The major drawback of reference database methods (both pattern and alignment) is that detection of totally novel pathogens is not possible. In contrast, assembly based methods, which reconstruct genetically distinct genotypes without need for a reference, can detect and reconstruct novel strains. When confronted with a novel strain that is not represented in the reference database, a good reference database based detection method should output the nearest possible strain as well as the uncertainty of the match. Ultimately, meaningful results are limited to the identification of strains with reasonably close matches within the database.

Reference Database Free Approaches

The methods described above all depend on either the presence of genome sequences in a reference database, or the reconstruction of a genome from reads. However, an additional subgroup of methods exist that do not use a reference database, but rather models within-sample diversity using a statistical model in order to delineate genetically distinct strains. These reference database free approaches apply statistics directly from elements acquired from the sequencing read set such as SNPs or k-mers.

For example, Eyre et al. (2013) applied a probabilistic model to allele frequencies at specific variable sites with the underlying

assumption that the sample was a mixture of two haplotypes. Variable sites were defined across the whole genome as locations with ambiguous calls. As this approach is limited to modeling a maximum of two strains in the data, other methods have extended this approach to allow for the presence of multiple strains in the sample data, including *estMOI,* DEploid, and pfmix (Assefa et al., 2014; O'Brien et al., 2016; Zhu et al., 2017). Both DEploid and *estMOI* use variant calls to infer the number of haplotypes in the dataset first locally (short regions), then globally. DEploid goes further by using a reference panel of known genomes to create a prior in their Bayesian approach to estimate the relative abundance, number of haplotypes, and their allelic states. Pfmix similarly uses a Bayesian model but does not estimate haplotypes, instead uses a single reference to provide variants and allele frequencies to directly infer the number and proportions of strains from allele frequencies.

Reference database-free approaches do not attempt to identify the presence of a specific, previously sequenced strain; rather, they utilize allele (variant) discrepancies within a WGS read set to quantify the number and proportion of unique strains present in a sample. These methods are therefore unable to offer insight on the relationship of strains in the sample compared to previously documented strains, since there is no mapping of the sample to a database of previously seen strains. However, they are especially effective in determining strain number of species within cultured WGS samples.

COMPARATIVE DISCUSSION OF DIFFERENT METHODOLOGIES

The methods mentioned in this review all aim to utilize the discriminative capability of WGS data to taxonomically classify samples at the level of individual strains within a species. These algorithms differ in required coverage, the number of strains that can be detected, the ability to detect higher level taxa (**Table 2**), and other criteria. To help guide tool selection we have made a flow chart (**Figure 4**) showing which types of tools would work well with different use cases.

Reference database methods (alignment and pattern based) are the most broadly applicable group of methods. They can be used on samples with lower coverage levels of the species of interest ($<1\times$) making them faster and more robust than assembly based approaches. In addition, they can be used to taxonomically classify or examine intra-species heterogeneity within an isolate culture expected to contain a single, well-studied species (such as *E. coli)*, as these methods require prior knowledge of a species. This is not possible for reference database free approaches. Also, some methods, such as GSMer and Sigma, are able to classify at both the species and strain level, which is useful when exploring strain level variety in metagenomic samples containing multiple species.

Biological uses of reference database methods can be quite broad. A common goal is to detect strains from only a particular pathogenic species. Pathoscope, SIGMA, WG-FAST, and PanPhlAN were all used to identify samples containing a particular toxic strain of *E. coli* from fecal metagenomic data obtained during a 2011 outbreak. In this case, although Sigma and Pathoscope are able to remove DNA from extraneous species, possibly providing a slight boost in computational efficiency, these methods are still both computationally intensive programs. Database methods can also be used to track transmission of strains between hosts. MIDAS was used to track strain transmission between mothers and their infants from stool metagenomes for a number of different microbial species. In a similar vein, StrainFinder was developed to track microbial strain transfer in fecal transplant cases. Phylogenetic-based methods such as those of Gan et al. (2016) and StrainSeeker can also track evolutionary divergence of the same strain within longitudinal metagenomic samples. These methods have the advantage of including a visual representation of the underlying decision process which can be easier to explain and understand. The phylogenetic framework also offers users the ability to sanity check results. For example, multiple closely related strains can be detected when the "true" strain is not present in the database.

If the single species present in the isolate sample is not as well-studied, then assembly methods are suitable, as they are not as dependent on prior knowledge encapsulated in a reference database. Assembly methods can also be useful in tracking progression of a single genome. For example, EVORhA was used to examine an evolving clonal population of *E. coli* strain. Because assembly methods require sufficient coverage ($50\times$ for EVORhA) to resolve haplotypes, these methods are not suitable for communities of samples with low coverages.

Certain methods quantify the number of strains or the relative abundance of strains within a sample using allelic variations within the dataset and do not require a database of known genomes. These reference database free tools are useful when the relationships between strains in a single-species sample are of interest, rather than the exact strain identities or their relationships to previously studied strains. This would be suitable for testing multiplicities of strains in uncultured soil samples or other extreme environments which are still under sampled. Reference database free approaches can also be applied for well characterized species, however, since pattern and alignment based tools can also offer strain identity – these might be preferred due to the extra information given.

Ease of use and speed of analysis are both important concerns when considering a metagenomic tool. **Table 1** details the different machine requirements and speed tests given by the methods reported in this paper. Though versatile and adaptable to different scenarios, tools requiring extensive mapping to a reference database can be extremely computationally intensive. Sigma required nearly 20 h resolving a single 5 strain community (20 million reads) against a database of 2,266 reference genomes with 62GB of memory and 64 cores. StrainFinder, another alignment method, took more than 48 h with 100–200 cores for 100 samples. Some methods were tested in high performance computing environments (i.e., Pathoscope, MEGAN, GOTTCHA, all >100GB memory) which may not always be available for clinicians. Additionally, tools requiring a database typically only report times/requirements to process a sample, but rarely include the time required to generate a custom database. We were only able to find both values for

TABLE 2 | Tool use cases and detection details.

Method name	Taxonomic level[1]	A[2]	Sample setting[3]	Use cases[4]
EVORhA	strain	Y	– high coverage data	– reconstruct evolutionary trajectories – clonal populations – resolve genomes in metagenomic communities
DESMAN	strain	Y	– better with low complexity (<20 strains) communities	– environmental populations – metagenomic communities
Sigma	strain, species	Y	– made specifically to provide useful information for outbreaks	– metagenomic bio surveillance for outbreaks
BIB	strain	Y	– species with clear population structure and well-separated lineages – unsuitable for species with frequent recombination (maybe the case for many alignment methods)	– clinical use, mixed samples – flagging contaminated/problematic samples
Pathoscope	multiple levels	Y	– designed to be complete framework to analyze metagenomic data	– environmental samples – clinical samples
DiTASiC	strain	Y	– comparing abundances across samples	– general strain identification and abundance – allows for differential abundance testing across samples
MEGAN	strain, species	Y	– broad taxonomic classification	– environmental populations
MetaMaps	strain, species	Y	– long read data	– medium complexity environmental communities – medium complexity
StrainFinder	strain, species	Y	– track strain genotypes over time – specifically made to understand real world clinical problem – requires prior knowledge for number of strains	– clinical/pathogen identification – human microbiome
Gan, Mingyu	strain	Y	– specifically for TB	– clinical TB samples – mixed infections of few strains
ConStrains	strain, species	Y	– only needs one genome per species – robust against unknown strains	– clinical microbiome sets – time series data – finding specific strains within population at low abundance
GOTTCHA	user defined	Y	– designed to find low abundance populations	– clinical diagnosis – bio surveillance – community profiling
WG-FAST	strain	N	– isolate identification (single isolate and complex samples – designed for low coverage strains	– disease outbreaks – pathogen identification
StrainSeeker	strain, species	Y	– phylogeny based – identifying clade of novel strain – unable to differentiate strains with few SNV	– pathogen identification
StrainEst	strain	Y	– identifying strains of particular species – best at lower than species level – limited for poorly characterized species	– ecological/environmental samples – human/skin microbiome – molecular epidemiology
StrainPhlAn	strain, species	N	– identifies most abundant strain of particular species within metagenomes not all strains – reconstruction of stain level phylogenies of species	– human microbiome
MIDAS	strain, species	N	– cannot quantify novel species	– transmission gut microbiome
metaSNV	strain, species	N	– strain level variation within metagenomes	– environmental samples
GSMer	strain, species	Y	– identify species/strain specific for well-studied organisms – possible false negatives if not all GSMs covered – false positives due to overlapping GSMs with incorrect strains	– human microbiome
PanPhlAn	strain, species	Y	– characterization of strain level gene elements – useful for population genomics where few reference genomes exist – culture free	– outbreak epidemiology – human microbiome
MetaPalette	strain, speices	Y	– metagenomic profiling	– environmental samples – human microbiome

(Continued)

TABLE 2 | Continued

Method name	Taxonomic level[1]	A[2]	Sample setting[3]	Use cases[4]
QuantTB	strain	Y	– specifically for TB	– mixed infections of few strains – clinical TB pathogen identification
Eyre, David W.	strain	Y	– mixed infection detection – assumes only mixes of 2 strains	– mixed infection screening in outbreak surveillance
pfmix	strain	Y	– mixed infection detection – specifically for *P. falciparum*	– pathogen identification
estMOI	strain	N	– specifically made for *P. falciparum* – estimates multiplicity of infection – might not be possible for highly related genomes	– pathogen identification – transmission intensity
DEploid	strain	Y	– estimating mixed infections – originally developed for *P. falciparum* – can be used for any mixture of strains within species	– pathogen identification
MixInfect	strain	Y	– detecting mixed infections in TB – not suitable for non-TB species	– pathogen identification

[1]Taxonomy levels the method claims to be able to accurate identify. [2]Denotes whether a method gives the abundance of a strain. [3]Specifics about which context the tool was originally demonstrated for. [4]Different use case scenarios that the tool can be used for or has been tested for.

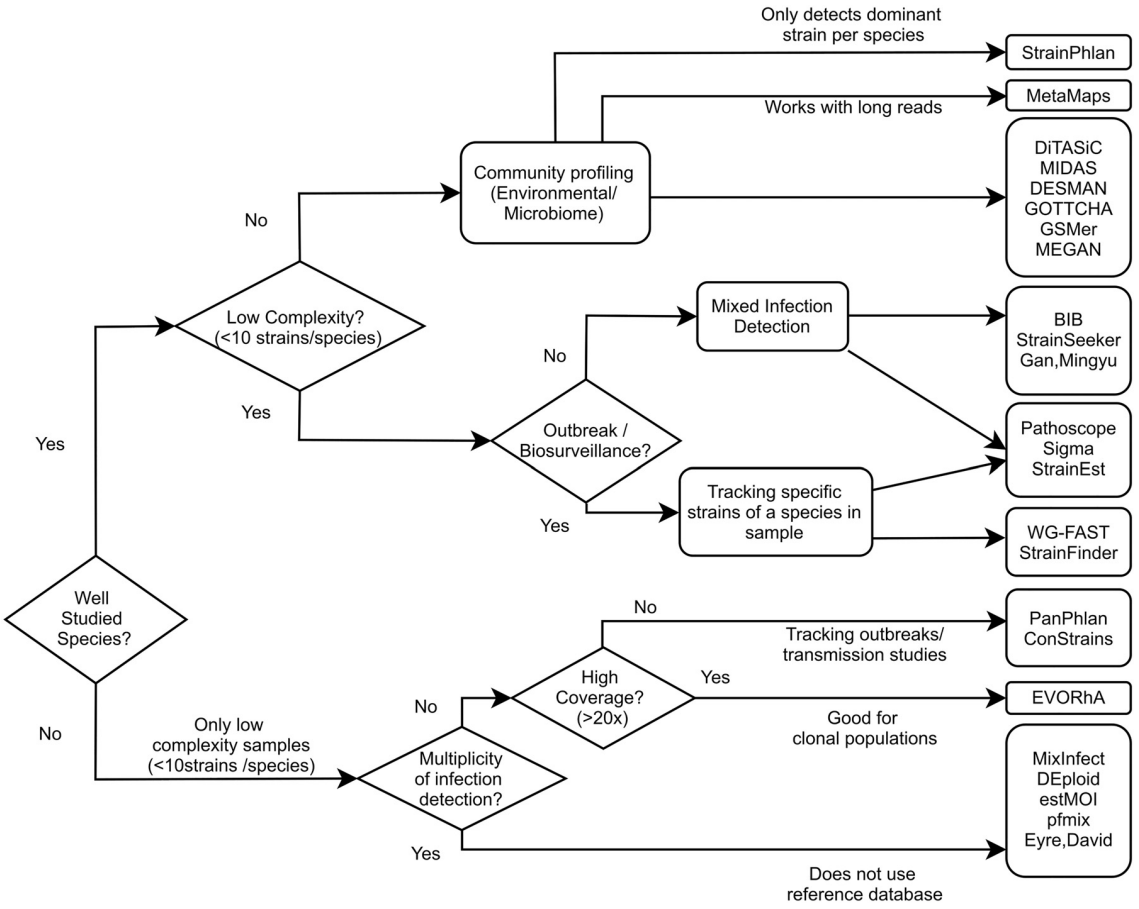


FIGURE 4 | Flow chart of tool selection depending on scenario. Guide chart showing which tools can be used in which use case. Presence of a tool under one use case doesn't necessarily exclude it from being applicable to another use case.

StrainSeeker, which process samples relatively quickly (<2 min) but suggests 300GB of space and 512GB of ram available to generate a database. In terms of usability, almost all of the tools were made to run in a Linux environment, therefore requiring some level of computational expertise in order to navigate requirements and installation setups. Few tools offer an

online accessible functionality (MEGAN and StrainSeeker). That being said, certain tools are bundled in easy to install package managers like Conda and R (i.e., DEploid, pfmix, StrainPhlAn), while others only offer a collection of scripts (i.e., MixInfect, and Eyre et al). Due to the requirements for installation and use (Bash/Linux), using most of these methods would require some bioinformatics knowledge. Further work would need to go into making these tools accessible and open for general use, such as online web tools, or a easy to use/install gui.

Most of the methods described in this review have not been benchmarked across all possible use case scenarios in a systematic or independent manner; therefore, a researcher using these tools will need to carefully determine whether a particular tool would work for their data type of interest. We discuss more about benchmarking in the next section.

METHOD EVALUATION, BENCHMARKING AND SIMULATION

Thorough and robust benchmarking of algorithms for a particular application and data type is critical. As this field is relatively new, there has yet to be a proper comparative study benchmarking the efficiency, accuracy, and specificity of these methods in a diversity of application domains: clinical pathogens (Cassir et al., 2016; Ward et al., 2016), microbiomes (Fang et al., 2018; Goltsman et al., 2018) and industrial biotechnology (Capece et al., 2016; Walsh et al., 2017; De Filippis et al., 2019) as examples.

The types of validation that have been performed for each method are indicated in **Table 1**. For all tools, an initial validation of model performance was performed using *in silico* simulated reads of known composition, generated from genomes of known host strains using tools such as *MetaSim, Grinder,* and *Art* (Richter et al., 2011; Angly et al., 2012; Huang et al., 2012). Alternatively, sequencing reads from presumed pure strains can be used. Testing applicability to strain mixes involves constructing a more complex synthetic dataset containing a mixture of varying quantities of individual strain read sets. Factors that must be considered in the construction of synthetic validation datasets include: (1) Determining the actual sequencing depth necessary to be able to identify a particular strain in a read set and number of reads to use. (2) The diversity in strain composition in terms of taxonomic levels that should be represented or background non-target species. (3) The level of complexity that needs to be introduced in the reads (in terms of SNVs and genomic distance between strains) and (4) the scalability of the method to fluctuation in sample size (e.g., low abundance strains in large sample sets). Validation on synthetic datasets addresses performance of the algorithms in the best-case scenario. Subsequent to these validation experiments, performance needs to be examined on test-case "real" samples, as this is often presents a much greater challenge than testing on *in silico*-generated datasets.

In order to compare the results of benchmarking different tools, metrics for comparing results across different types of outputs from various tools must be carefully chosen. The published benchmarking methods for the tools described in **Table 1** use a variety of different metrics. The most common method employed for the published tools involves testing the specific algorithm on a dataset of known diversity and abundance, and comparing accuracy metrics. For alignment- and pattern based methods, a true and false positive would be defined as whether the algorithm was able to detect the correct strain within the sample, or whether it detected the wrong strain, respectively. A false negative would be defined if the algorithm failed to detect a strain present in the sample, and a true negative would be called if the algorithm did not output any strains not present. An important consideration in the assessment of true negatives is whether the algorithm informs the user of the uncertainty of the match and outputs the nearest strain. Most methods mentioned in this paper quantified the reliability of their method by either calculating the true positive rate/false discovery rate or by checking manually whether the results were correct.

In addition to simply identifying which strains are present or absent in a sample, additional metrics must assess the accuracy in estimating strain abundances. One method to do this, used by the assembly based detection method, EVORhA, uses the mean absolute error (MAE) metric between the true abundances and estimated abundances. In addition, they also calculated the root mean squared error (RMSE), which was also used by Eyre et al. Another method to assess accuracy in strain abundance is the Jenson-Shannon divergence, which was used in ConStrains to measure their prediction accuracy.

A comprehensive comparison and benchmarking of these tools is needed to provide further insight into the efficiency of these tools at performing strain-level identification on a wide range of sample types, be it metagenomic, clinical, or cultures. This benchmarking strategy would need to deal with the nuances between tools, as they have different goals, different use-case scenarios, and different criteria for success. It might be possible to conduct these comprehensive benchmarks in categories such that similar tools could be evaluated together on novel datasets with a common evaluation metric.

CONCLUSION AND FUTURE DIRECTIONS

Whole genome sequencing of microbial populations has the capability to offer a view into genetic diversity at varying taxonomic levels. Current widely used taxonomic classifiers allow for the identification of species within WGS sets. However, algorithms for finer-grained classification, at the individual strain level within a species, are still relatively new. Such techniques have the capacity to greatly impact healthcare and other fields by precise tracking of disease outbreaks, differentiation of commensal and pathogenic strains, and linking strain level genotypic traits with phenotypic characteristics of clinical and industrial importance (Capece et al., 2016; Cassir et al., 2016; Ward et al., 2016; Walsh et al., 2017; Fang et al., 2018; Goltsman et al., 2018; De Filippis et al., 2019). One assumption almost universally made within taxonomic tools is that a direct relationship exists between strain read coverage and

strain abundance in the sample. As such, calculations of strain abundance levels take into account the variations of coverage across variant sites or reads. Though intuitive, none of the tools presented here presented analysis to prove this assumption. Conducting such verification steps is particularly important for tools focusing on clinical use and pathogen identification, where it is typical for a culturing step to be conducted before sequencing. In actuality, there could be many reasons why read abundance does not directly reflect the composition of the sample: isolation technique (culture sweep vs. single colony isolation), cell lysis efficiency, contamination skewing read depth, or the sequencing process itself (Morgan et al., 2010; Pereira et al., 2018).

There are numerous ways in which current strain identification methods can improve their benchmarking. Firstly, very few algorithms tested the performance of their tools on multiple (>2) low abundance strains (<1–2×). Detecting low abundance strains would be preferred for microbial communities such as the gut, where specific strains exhibit differing pathogenicity. Secondly, no methods quantified or benchmarked how genetically distant a strain needs to be in order to properly delineate it. Third, there are no tools that allow a user to compare strains within and across samples, which would be useful for transmission studies. Lastly, delineating extremely closely related strains remains a difficult problem for the metagenomic tools. Many tools requiring a reference database remove genomes from the database that are extremely close together or self-report that they would not work well with highly related genomes (Assefa et al., 2014; Sankar et al., 2015; Albanese and Donati, 2017). Such analysis remains difficult due to the problems that arise when considering closely related strains such as an increase in false positives due to both strains being reported when only one is actually present or problems within the model itself driven by high levels of collinearity. The difficulty with detecting extremely close strains is further compounded due to the ambiguous definition of a strain.

The methods detailed in this literature review are almost all directed toward sequencing technologies that produce reads from mixtures of cells. Direct sequencing of individual cells would bypass this need to computationally subdivide reads produced from current NGS technologies into those originating from different strains. Single-cell sequencing strategies such as Drop-Seq (Macosko et al., 2015) and 10× Genomics (Zheng et al., 2017) are rapidly improving to provide a systematic and comprehensive view of the genetic diversity of complex communities. Having sequencing data originating from individual cells would greatly simplify studies of heterogeneous populations of strains. However, there are still technical difficulties to overcome before single-cell sequencing becomes widely adopted. It is probable that the next iteration of strain-level identification algorithms will be focused on such technologies. One pioneering example is *MetaSort,* which combines the advantages of both WGS and single cell sequencing data (Ji et al., 2017). This method assembles genomes from both WGS reads and single cell sequencing reads and integrates the two using a machine-learning algorithm, resulting in genomes present in the sample. The increased resolution from single cell sequencing based detection is likely to uncover novel forms of genetic heterogeneity. In addition, advances in long read sequencing continue to change the scope and direction of strain-level detection in metagenomic samples.

Longer read lengths could make it easier and more practical to phase haplotypes, as well as identify strains with fewer reads. A number of studies have applied long read sequencing data from third generation sequencing platforms such as Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) to assemble individual strains within metagenomic communities (Tsai et al., 2016; Bertrand et al., 2019). For example, Somerville et al. (2019) used the long-read assembler, Flye (Kolmogorov et al., 2019), to reconstruct individual contigs from a long read metagenomic sample, followed by a phylogenetic analysis using NCBI RefSeq to determine strain identity. Long-reads can also be beneficial for alignment based strain identification approaches. For example, MetaMaps developed its own mapping algorithm to align long reads to genomes in a database. Challenges for strain-level identification using long-read sequencing can vary based on the tools. In the case of MetaMaps, a minimum read-length is required for a read to be considered, resulting in numerous unassigned reads. Overall, the use of longer reads can mitigate some of the limitations of short-reads, allowing for the resolution of difficult to sequence regions and longer contigs. However, this comes at the expense of increased errors, lower coverage and higher cost. We still expect many more tools will be released for long-read platforms as it continues to gain in popularity.

The ability to quantify and detect bacterial strains within heterogeneous environments has applications in numerous fields including diagnostics (Dekkera, 2018), clinical studies for the microbiome (Wang et al., 2015), bio surveillance (Ahn et al., 2015), tracking transmission of infectious strains in an outbreak (Hong et al., 2014; Ahn et al., 2015; Nayfach et al., 2016), providing insight into the spread of antibiotic resistance (Sukhum et al., 2019), tracking progression of within-host bacterial evolution (Pulido-Tamayo et al., 2015) and exploring diverse environments (Tringe and Rubin, 2005). We look forward to the wide range of applications and effects these tools will have in shaping and progressing sequencing based research.

AUTHOR CONTRIBUTIONS

CA and TA conceived, designed, and wrote the manuscript. AM, TS, and AE edited and proofread the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research is supported by the TU Delft | Global Initiative, a program of the Delft University of Technology to boost Science and Technology for Global Development. This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under grant number U19AI110818 to the Broad Institute.

REFERENCES

Ahn, T. H., Chai, J., and Pan, C. (2015). Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. *Bioinformatics* 31, 170–177. doi: 10.1093/bioinformatics/btu641

Albanese, D., and Donati, C. (2017). Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.* 8:2260. doi: 10.1038/s41467-017-02209-5

Alizon, S., de Roode, J. C., and Michalakis, Y. (2013). Multiple infections and the evolution of virulence. *Ecol. Lett.* 16, 556–567. doi: 10.1111/ele.12076

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., and Tyson, G. W. (2012). Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.* 40:e94. doi: 10.1093/nar/gks251

Anyansi, C., Keo, A., Walker, B. J., Straub, T. J., Manson, A. L., Earl, A. M., et al. (2020). QuantTB – a method to classify mixed *Mycobacterium tuberculosis* infections within whole genome sequencing data. *BMC Genomics* 21:80. doi: 10.1186/s12864-020-6486-3

Assefa, S. A., Preston, M. D., Campino, S., Ocholla, H., Sutherland, C. J., and Clark, T. G. (2014). EstMOI: estimating multiplicity of infection using parasite deep sequencing data. *Bioinformatics* 30, 1292–1294. doi: 10.1093/bioinformatics/btu005

Balmer, O., and Tanner, M. (2011). Prevalence and implications of multiple-strain infections. *Lancet Infect. Dis.* 11, 868–878. doi: 10.1016/S1473-3099(11)70241-9

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A. H. Q., Kumar, M. S., Li, C., et al. (2019). Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* 37, 937–944. doi: 10.1038/s41587-019-0191-2

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 34, 525–527. doi: 10.1038/nbt.3519

Byrd, A. L., Perez-Rogers, J. F., Manimaran, S., Castro-Nallar, E., Toma, I., McCaffrey, T., et al. (2014). Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC Bioinformatics* 15:262. doi: 10.1186/1471-2105-15-262

Canzar, S., and Salzberg, S. L. (2017). "Short read mapping: an algorithmic tour," in *Proceedings of the IEEE*, (Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc.), 436–458. doi: 10.1109/JPROC.2015.2455551

Capece, A., Granchi, L., Guerrini, S., Mangani, S., Romaniello, R., Vincenzini, M., et al. (2016). Diversity of *Saccharomyces cerevisiae* strains isolated from two Italian wine-producing regions. *Front. Microbiol.* 7:1018. doi: 10.3389/fmicb.2016.01018

Cassir, N., Benamar, S., and La Scola, B. (2016). *Clostridium butyricum*: from beneficial to a new emerging pathogen. *Clin. Microbiol. Infect.* 22, 37–45. doi: 10.1016/J.CMI.2015.10.014

Cespedes, C., Said-Salim, B., Miller, M., Lo, S. H., Kreiswirth, B. N., Gordon, R. J., et al. (2005). The clonality of *Staphylococcus aureus* nasal carriage. *J. Infect. Dis.* 191, 444–452. doi: 10.1086/427240

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848

Clement, N. L., Snell, Q., Clement, M. J., Hollenhorst, P. C., Purwar, J., Graves, B. J., et al. (2009). The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics* 26, 38–45. doi: 10.1093/bioinformatics/btp614

Cohen, T., van Helden, P. D., Wilson, D., Colijn, C., McLaughlin, M. M., Abubakar, I., et al. (2012). Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. *Clin. Microbiol. Rev.* 25, 708–719. doi: 10.1128/CMR.00021-12

Costea, P. I., Munch, R., Coelho, L. P., Paoli, L., Sunagawa, S., and Bork, P. (2017). metaSNV: A tool for metagenomic strain level analysis. *PLoS One* 12:e0182392. doi: 10.1371/journal.pone.0182392

De Filippis, F., La Storia, A., Villani, F., and Ercolini, D. (2019). Strain-level diversity analysis of *Pseudomonas fragi* after In Situ pangenome reconstruction shows distinctive spoilage-associated metabolic traits clearly selected by different storage conditions. *Appl. Environ. Microbiol.* 85:e02212-18. doi: 10.1128/AEM.02212-18

Dekkera, J. P. (2018). Metagenomics for clinical infectious disease diagnostics steps closer to reality. *J. Clin. Microbiol.* 56, e850–e818. doi: 10.1128/JCM.00850-18

Deurenberg, R. H., Bathoorn, E., Chlebowicz, M. A., Couto, N., Ferdous, M., García-Cobos, S., et al. (2017). Application of next generation sequencing in clinical microbiology and infection prevention. *J. Biotechnol.* 243, 16–24. doi: 10.1016/j.jbiotec.2016.12.022

Dilthey, A. T., Jain, C., Koren, S., and Phillippy, A. M. (2019). Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nat. Commun.* 10:3066. doi: 10.1038/s41467-019-10934-2

El-Halfawy, O. M., and Valvano, M. A. (2015). Antimicrobial heteroresistance: an emerging field in need of clarity. *Clin. Microbiol. Rev.* 28, 191–207. doi: 10.1128/CMR.00058-14

Esposito, S., Bosis, S., Cavagna, R., Faelli, N., Begliatti, E., Marchisio, P., et al. (2002). Characteristics of *Streptococcus pneumoniae* and atypical bacterial infections in children 2-5 years of age with community-acquired pneumonia. *Clin. Infect. Dis.* 35, 1345–1352. doi: 10.1086/344191

Eyre, D. W., Cule, M. L., Griffiths, D., Crook, D. W., Peto, T. E. A., Walker, A. S., et al. (2013). Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in clostridium difficile transmission. *PLoS Comput. Biol.* 9:e1003059. doi: 10.1371/journal.pcbi.1003059

Eyre, D. W., Walker, A. S., Griffiths, D., Wilcox, M. H., Wyllie, D. H., Dingle, K. E., et al. (2012). Clostridium difficile mixed infection and reinfection. *J. Clin. Microbiol.* 50, 142–144. doi: 10.1128/JCM.05177-11

Falagas, M. E., Makris, G. C., Dimopoulos, G., and Matthaiou, D. K. (2008). Heteroresistance: a concern of increasing clinical significance? *Clin. Microbiol. Infect.* 14, 101–104. doi: 10.1111/j.1469-0691.2007.01912.x

Fang, X., Monk, J. M., Nurk, S., Akseshina, M., Zhu, Q., Gemmell, C., et al. (2018). Metagenomics-based, strain-level analysis of *Escherichia coli* from a time-series of microbiome samples from a Crohn's disease patient. *Front. Microbiol.* 9:2559. doi: 10.3389/fmicb.2018.02559

Fischer, M., Strauch, B., and Renard, B. Y. (2017). Abundance estimation and differential testing on strain level in metagenomics data. *Bioinformatics* 33, i124–i132. doi: 10.1093/bioinformatics/btx237

Fournier, P.-E., Dubourg, G., and Raoult, D. (2014). Clinical detection and characterization of bacterial pathogens in the genomics era. *Genome Med.* 6:114. doi: 10.1186/s13073-014-0114-2

Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-nallar, E., et al. (2013). Pathoscope: species identification and strain attribution with unassembled sequencing data Pathoscope: species identification and strain attribution with unassembled sequencing data. *Genome Res.* 23, 1721–1729. doi: 10.1101/gr.150151.112

Frank, S. A. (1996). Models of parasite virulence. *Q. Rev. Biol.* 71, 37–78. doi: 10.1086/419267

Freitas, T. A. K., Li, P.-E., Scholz, M. B., and Chain, P. S. G. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* 43:e69. doi: 10.1093/nar/gkv180

Gan, M., Liu, Q., Yang, C., Gao, Q., and Luo, T. (2016). Deep whole-genome sequencing to detect mixed infection of mycobacterium tuberculosis. *PLoS One* 11:e0159029. doi: 10.1371/journal.pone.0159029

Glaus, P., Honkela, A., and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28, 1721–1728. doi: 10.1093/bioinformatics/bts260

Goldman, D., and Domschke, K. (2014). Making sense of deep sequencing. *Int. J. Neuropsychopharmacol.* 17, 1717–1725. doi: 10.1017/S1461145714000789

Goltsman, D. S. A., Sun, C. L., Proctor, D. M., DiGiulio, D. B., Robaczewska, A., Thomas, B. C., et al. (2018). Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. *Genome Res.* 28, 1467–1480. doi: 10.1101/gr.236000.118

Hong, C., Manimaran, S., Shen, Y., Perez-Rogers, J. F., Byrd, A. L., Castro-Nallar, E., et al. (2014). PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. *Microbiome* 2:33. doi: 10.1186/2049-2618-2-33

Huang, H. Y., Tsai, Y. S., Lee, J. J., Chiang, M. C., Chen, Y. H., Chiang, C. Y., et al. (2010). Mixed infection with Beijing and non-Beijing strains and drug resistance pattern of *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* 48, 4474–4480. doi: 10.1128/JCM.00930-10

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: A next-generation sequencing read simulator. *Bioinformatics* 28, 593–594. doi: 10.1093/bioinformatics/btr708

Hunter, C. I., Mitchell, A., Jones, P., Mcanulla, C., Pesseat, S., Scheremetjew, M., et al. (2012). Metagenomic analysis: the challenge of the data bonanza. *Brief. Bioinform.* 13, 743–746. doi: 10.1093/bib/bbs020

Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res.* 17, 377–386. doi: 10.1101/gr.5969107

Ji, P., Zhang, Y., Wang, J., and Zhao, F. (2017). MetaSort untangles metagenome assembly by reducing microbial community complexity. *Nat. Commun.* 8:14306. doi: 10.1038/ncomms14306

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* 12, 962–968. doi: 10.1101/gr.87702

Kim, D., and Salzberg, S. L. (2011). TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* 12:15. doi: 10.1186/gb-2011-12-8-r72

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* 37, 540–546. doi: 10.1038/s41587-019-0072-8

Koslicki, D., and Falush, D. (2016). MetaPalette: a k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. *mSystems* 1:e00020-16. doi: 10.1128/msystems.00020-16

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

Lessing, M. P., Jordens, J. Z., and Bowler, I. C. (1995). Molecular epidemiology of a multiple strain outbreak of methicillin-resistant *Staphylococcus aureus* amongst patients and staff. *J. Hosp. Infect.* 31, 253–260. doi: 10.1016/0195-6701(95)90204-x

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.* 11, 473–483. doi: 10.1093/bib/bbq015

Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J., and Gevers, D. (2015). ConStrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.* 33, 1045–1052. doi: 10.1038/nbt.3319

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 161, 1202–1214. doi: 10.1016/j.cell.2015.05.002

Mande, S. S., Mohammed, M. H., and Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.* 13, 669–681. doi: 10.1093/bib/bbs054

Marshall, J. A. (2002). "Mixed infections of intestinal viruses and bacteria in humans," in *Polymicrobial Diseases*, eds K. Brogden and J. Guthmiller (Washington, DC: ASM Press).

Martín, M. J., Herrero, J., Mateos, A., and Dopazo, J. (2003). Comparing bacterial genomes through conservation profiles. *Genome Res.* 13, 991–998. doi: 10.1101/gr.678303

Marx, V. (2016). Microbiology: the road to strain-level identification. *Nat. Methods* 13, 401–404. doi: 10.1038/nmeth.3837

Maxson, T., and Mitchell, D. A. (2016). Targeted treatment for bacterial infections: prospects for pathogen-specific antibiotics coupled with rapid diagnostics. *Tetrahedron* 72, 3609–3624. doi: 10.1016/j.tet.2015.09.069

Minagawa, S., Takayanagi, N., Hara, K., Takaku, Y., Tsutiya, Y., Hijikata, N., et al. (2008). [Clinical features of mixed infections in patients with *Streptococcus pneumoniae* pneumonia]. *Nihon Kokyuki Gakkai Zasshi* 46, 278–284.

Morgan, J. L., Darling, A. E., and Eisen, J. A. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One* 5:e10209. doi: 10.1371/journal.pone.0010209

Navarro, Y., Herranz, M., Pérez-Lago, L., Lirola, M. M., Ruiz-Serrano, M. J., Bouza, E., et al. (2011). Systematic survey of clonal complexity in tuberculosis at a populational level and detailed characterization of the isolates involved. *J. Clin. Microbiol.* 49, 4131–4137. doi: 10.1128/JCM.05203-11

Nayfach, S., Rodriguez-Mueller, B., Garud, N., and Pollard, K. S. (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* 26, 1612–1625. doi: 10.1101/gr.201863.115

O'Brien, J. D., Iqbal, Z., Wendler, J., and Amenga-Etego, L. (2016). Inferring strain mixture within clinical *Plasmodium falciparum* isolates from genomic sequence data. *PLoS Comput. Biol.* 12:e1004824. doi: 10.1371/journal.pcbi.1004824

Pereira, M. B., Wallroth, M., Jonsson, V., and Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19:274. doi: 10.1186/s12864-018-4637-6

Plazzotta, G., Cohen, T., and Colijn, C. (2015). Magnitude and sources of bias in the detection of mixed strain *M. tuberculosis* infection. *J. Theor. Biol.* 368, 67–73. doi: 10.1016/j.jtbi.2014.12.009

Pulido-Tamayo, S., Sánchez-Rodríguez, A., Swings, T., Van Den Bergh, B., Dubey, A., Steenackers, H., et al. (2015). Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. *Nucleic Acids Res.* 43:e105. doi: 10.1093/nar/gkv478

Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G., et al. (2017). DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.* 18:181. doi: 10.1186/s13059-017-1309-9

Richter, D. C., Ott, F., Auch, A. F., Schmid, R., and Huson, D. H. (2011). "MetaSim: a sequencing simulator for genomics and metagenomics," in *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*, ed. F. J. de Bruijnin (Hoboken, NJ: John Wiley & Sons, Inc.), 417–421. doi: 10.1002/9781118010518.ch48

Roosaare, M., Vaher, M., Kaplinski, L., Möls, M., Andreson, R., and Lepamets, M. (2016). StrainSeeker: fast identification of bacterial strains from unassembled sequencing reads using user-provided guide trees. *bioRxiv* [Preprint]. doi: 10.1101/040261

Sahl, J. W., Schupp, J. M., Rasko, D. A., Colman, R. E., Foster, J. T., and Keim, P. (2015). Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data. *Genome Med.* 7:52. doi: 10.1186/s13073-015-0176-9

Sankar, A., Malone, B., Bayliss, S., Pascoe, B., Méric, G., Hitchings, M. D., et al. (2015). Bayesian identification of bacterial strains from sequencing data. *bioRxiv* [Preprint]. doi: 10.1099/mgen.0.000075

Scholz, M., Ward, D. V., Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., et al. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat. Methods* 13, 435–438. doi: 10.1038/nmeth.3802

Segata, N. (2018). On the road to strain-resolved comparative metagenomics. *mSystems* 3:e00190-17. doi: 10.1128/msystems.00190-17

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2013). Metagenomic microbial community profiling using unique clade- specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066.Metagenomic

Smillie, C. S., Sauk, J., Gevers, D., Friedman, J., Sung, J., Youngster, I., et al. (2018). Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe* 23, 229–240.e5. doi: 10.1016/J.CHOM.2018.01.003

Sobkowiak, B., Glynn, J. R., Houben, R. M. G. J., Mallard, K., Phelan, J. E., Guerra-Assunção, J. A., et al. (2018). Identifying mixed Mycobacterium tuberculosis infections from whole genome sequence data. *BMC Genomics* 19:613. doi: 10.1186/s12864-018-4988-z

Somerville, V., Lutz, S., Schmid, M., Frei, D., Moser, A., Irmler, S., et al. (2019). Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol.* 19:143. doi: 10.1186/s12866-019-1500-0

Sukhum, K. V., Diorio-Toth, L., and Dantas, G. (2019). Genomic and metagenomic approaches for predictive surveillance of emerging pathogens and antibiotic resistance. *Clin. Pharmacol. Ther.* 106, 512–524. doi: 10.1002/cpt.1535

Teeling, H., and Glöckner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis-A bioinformatic perspective. *Brief. Bioinform.* 13, 728–742. doi: 10.1093/bib/bbs039

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Tringe, S. G., and Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet.* 6, 805–814. doi: 10.1038/nrg1709

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

Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Res.* 27, 626–638. doi: 10.1101/gr.216242.116

Tsai, Y. C., Conlan, S., Deming, C., Nisc Comparative Sequencing Program, Segre, J. A., Kong, H. H., et al. (2016). Resolving the complexity of human skin metagenomes using single-molecule sequencing. *mBio* 7:e01948-15. doi: 10.1128/mBio.01948-15

Tu, Q., He, Z., and Zhou, J. (2014). Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Res.* 42, 1–12. doi: 10.1093/nar/gku138

Votintseva, A. A., Bradley, P., Pankhurst, L., Del Ojo Elias, C., Loose, M., Nilgiriwala, K., et al. (2017). Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. *J. Clin. Microbiol.* 55, 1285–1298. doi: 10.1128/JCM.02483-16

Walsh, A. M., Crispie, F., Daari, K., O'Sullivan, O., Martin, J. C., Arthur, C. T., et al. (2017). Strain-level metagenomic analysis of the fermented dairy beverage nunu highlights potential food safety risks. *Appl. Environ. Microbiol.* 83:e01144-17. doi: 10.1128/AEM.01144-17

Wang, W. L., Xu, S. Y., Ren, Z. G., Tao, L., Jiang, J. W., and Zheng, S. S. (2015). Application of metagenomics in the human gut microbiome. *World J. Gastroenterol.* 21, 803–814. doi: 10.3748/wjg.v21.i3.803

Ward, D. V., Scholz, M., Zolfo, M., Taft, D. H., Schibler, K. R., Tett, A., et al. (2016). Metagenomic sequencing with strain-level resolution implicates uropathogenic *E. coli* in necrotizing enterocolitis and mortality in preterm infants. *Cell Rep.* 14, 2912–2924. doi: 10.1016/J.CELREP.2016.03.015

Wood, D. E., Salzberg, S. L. S., Venter, C., Remington, K., Heidelberg, J., Halpern, A., et al. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46

Yuan, S., Cohen, D. B., Ravel, J., Abdo, Z., and Forney, L. J. (2012). Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PLoS One* 7:e33865. doi: 10.1371/journal.pone.0033865

Zagordi, O., Bhattacharya, A., Eriksson, N., and Beerenwinkel, N. (2011). ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12:119. doi: 10.1186/1471-2105-12-119

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* 8:14049. doi: 10.1038/ncomms14049

Zhu, S. J., Almagro-garcia, J., and Mcvean, G. (2017). Deconvoluting multiple infections in *Plasmodium falciparum* from high throughput sequencing data. *Bioinformatics* 34, 9–15. doi: 10.1093/bioinformatics/btx530

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