Bioinformatics Advance Access published February 1, 2016

Optimization of alignment-based methods for taxonomic binning of metagenomics reads

Magali Jaillard [1,2,†,*], Maud Tournoud [1,†], Faustine Meynier [3], and Jean-Baptiste Veyrieras [1]

[1]Bioinformatics Research Department, bioMerieux, 69280 Marcy L'Etoile, France
[2]LBBE, UMR CNRS 5558 Univ. Lyon 1, F-69622 Villeurbanne, France
[3]Biomathematics System Development, bioMerieux, 69280 Marcy L'Etoile, France

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: Alignment-based taxonomic binning for metagenome characterization proceeds in two steps: reads mapping against a reference database (RDB) and taxonomic assignment according to the best hits. Beyond the sequencing technology and the completeness of the RDB, selecting the optimal configuration of the workflow, in particular the mapper parameters and the best hit selection threshold, to get the highest binning performance remains quite empirical.

Results: We developed a statistical framework to perform such optimization at a minimal computational cost. Using an optimization experimental design and simulated datasets for three sequencing technologies, we built accurate prediction models for five performance indicators and then derived the parameter configuration providing the optimal performance. Whatever the mapper and the dataset, we observed that the optimal configuration yielded better performance than the default configuration and that the best hit selection threshold had a large impact on performance. Finally, on a reference dataset from the Human Microbiome Project, we confirmed that the optimized configuration increased the performance compared to the default configuration.

Availability and implementation: Not applicable.

Contact: magali.dancette@biomerieux.com

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Metagenomics is the study of the DNA content of a sample particularly powerful to recover complex mix of organisms, including those difficult to isolate by standard techniques (Padmanabhan *et al.*, 2013; Hugenholtz *et al.*, 2002). In the last decade, the advent of Next Generation Sequencing (NGS) technologies has quickly shifted metagenomics approaches from low-scale studies to large-scale investigations, thereby creating a need for improvement and development of specific bioinformatic methods. Taxonomic profiling and taxonomic binning are popular methods to assess the taxonomic diversity of a sequenced sample (Dröge *et al.*, 2015). Taxonomic profiling methods aim to estimate the relative abundance of each taxon in the sample by classifying reads using a set of marker-genes specific to each taxonomic clade (Sunagawa *et al.*, 2013; Segata *et al.*, 2012; Liu *et al.*, 2010) or based on their k-mer composition (Koslicki *et al.*, 2014). Taxonomic binning methods aim to assign each individual read to a given taxon within the microbial taxonomy. They are not as fast as taxonomic profiling methods to generate a taxonomic profile. Nevertheless reads binning is a mandatory step for subsequent analysis that require draft genome reconstruction (Dröge *et al.*, 2015).

Taxonomic binning algorithms are based either on read alignment strategies or on compositional approaches (such as nucleotide k-mer frequencies), or possibly on a mixture of these two approaches (Mande *et al.*, 2012). In this article, we focused on alignment-based taxonomic binning methods, whose principle was introduced by Huson *et al.* (2007). Such methods proceeds generally in two steps. First, reads are aligned against a reference database (RDB). Then, for each read, the Lowest Common Ancestor (LCA) among the best hits is retrieved. To this end, a threshold is applied to select only hits whose scores lie within a percentage of the best hit score.

Several evaluation studies have been published that compare mapper performance for resequencing applications (Ruffalo *et al.*, 2011; Hatem *et al.*, 2013; Caboche *et al.*, 2014; Schbath *et al.*, 2012; Holtgrewe *et al.*, 2011; Břinda *et al.*, 2015). However, in the context of taxonomic binning, there is still no consensus on which mapper and configuration to use. The most comprehensive study to date was performed in the context of the Human Microbiome Project (HMP), where Martin *et al.* (2012) compared six mappers and optimized the parameters for the best mapper, CLC[1], by testing six combinations of two parameter values.

Beyond the NGS technology (*e.g.* read length, sequencing error profile) and the completeness of the RDB, taxonomic binning performance depends on the mapper configuration and the taxonomic assignment strategy. Mappers can have many parameters (e.g. more than 15 alignment and score parameters for Bowtie2), some being numerical and other categorical (often with many levels). It is nearly impossible to evaluate the performance of all parameter configurations in order to identify the best one. Here,

*to whom correspondence should be addressed

†M Jaillard and M Tournoud contributed equally to this work.

[1] CLCbio website: `http://www.clcbio.com`

we present a statistical framework to efficiently select the optimal configuration (mapper parameters and best hit threshold for LCA calculation) which maximizes taxonomic binning performance for a given NGS technology and level of RDB completeness. We relied on the Design Of Experiments[2] (DOE) methodology to reduce the number of configurations to run while still efficiently analyzing the parameter space. This optimization strategy could be adapted to any NGS technology, applied to different mappers and adjusted according to the targeted microbial complexity and distinct levels of the RDB completeness.

2 METHODS

As previously stated, our objective was to find the optimal parameter configuration for a set of sequencing technologies and mappers. To fulfill this objective, we implemented the two following steps: first, we screened the parameters expected to have the largest effect on the pipeline performance, and then we optimized the values of the selected parameters to select the pipeline configuration leading to the best performance. The first screening step was mandatory because optimization studies can only be performed on a limited number of factors (or parameters). In this step, we relied on previously published studies to select the most influential parameters and reduce the number of parameters to optimize up to five. Then, we used an optimization experimental design to find the optimal parameter configuration. As detailed below, the advantage of the experimental design is to identify the optimal configuration at a minimal computation cost.

2.1 Screening pipeline parameters

2.1.1 Mappers Four mappers were included in this study: BWA-backtrack (Li and Durbin, 2009) (version 0.7.4), BWA-MEM (Li, 2012) (version 0.7.4), Bowtie2 (Langmead and Salzberg, 2012) (version 2.1.0), and TMAP[3] (version 3.0.1). Mappers from the BWA suite include around twenty parameters, mostly integer and Boolean. In addition to >15 alignment and scoring parameters, Bowtie2 offers eight preset modes for fast, sensitive, local, or end-to-end mapping. TMAP, which embeds four mappers, is fully customizable with >50 numerical parameters and >20 categorical parameters.

In order to reduce the parameter space, we selected parameters expected to have the largest effect on the taxonomic binning performance. This selection was based on mapper documentation and previously published benchmarks (Hatem *et al.*, 2013). We targeted parameters typically controlling the seed characteristics, the clipping or the alignment effort. Briefly, the parameters evaluated for BWA-MEM were: minimum seed length ($s.length$), off-diagonal X-dropoff ($zdropoff$), trigger re-seeding ($rseed.length$), the clipping penalty ($clipping$), for TMAP: score threshold ($score$), soft-clipping ($clipping$), mapping sequence ($mode$), for Bowtie2: seed length ($s.length$), number of errors in the seed ($s.error$), number of consecutive seed extensions ($effort$), local or end-to-end mode ($mode$), and for BWA-backtrack: iterative search ($iterative$), fraction of missing alignments ($align.error$), maximal edit distance in the seed ($s.error$), and seed length ($s.length$). For each quantitative parameter, we selected lower and upper values based on the mapper documentation, but default mapper parameters were also evaluated. For each mapper, the parameters, their corresponding option in the software, the tested values (lower, default, and upper values for quantitative parameters and levels for qualitative parameters) are presented in Supplementary Table 2.

2.1.2 Best hit selection threshold For each hit j obtained by mapping read i against the RDB, an alignment quality score AS_{ij} was defined as the alignment length minus the edit distance. This score was calculated for each read independently and was not used to compare hits from different reads. Both values could be found in the SAM output (Li and Durbin, 2009) irrespective to the mapper.

Then for each read i, a hit j was retained as a candidate for LCA calculation if $AS_{ij} \geq (1 - bestHit) * \max_j AS_{ij}$, with $bestHit$ a threshold parameter in $[0; 1]$. When $bestHit = 0$ only the best hits (including ties) were retained, and when $bestHit = 1$, all hits were considered as best hits and retained for the LCA calculation. $bestHit$ was included in our study as the model parameter controlling for the best hit selection; the three tested values for the $bestHit$ threshold parameter were: $bestHit = 0, 0.25, 0.5$.

2.2 Optimization plan

The objective of this study was to find the optimal pipeline configuration for each mapper. However, running the pipeline for all the combinations of the selected mapper parameters and the $bestHit$ values was not feasible, motivating the use of an experimental plan. Indeed, an optimal experimental design ensures to find the best configuration in a minimum number of experiments. In this application, an experiment was a run of the pipeline with a set of parameters and the outcome to optimize was the pipeline performance. The set of experiments to perform are given by the experimental plan. There exists several optimization plans, with different assumptions for the nature of the parameter effects (linear or quadratic) and interactions, and thus leading to different number of experiments. Given that we wanted to evaluate the optimal configurations for four mappers (BWA-backtrack, BWA-MEM, TMAP, and Bowtie2) and three sequencing technologies (Roche 454, IT PGM, Illumina HiSeq), we had to make two reasonable assumptions to limit the number of experiments to perform: linear effect of parameters and negligible interactions. Given these assumptions, we chose a Hadamard design. This design relies on a Hadamard matrix, which is an orthogonal matrix with all elements equal either to -1 or $+1$, such that $HH^T = nI_n$, where I_n is the $n \times n$ identity matrix. This design ensures that the variance of the estimators for the parameter effects are minimal. Plackett and Burman (1946) proposed an algorithm to generate the Hadamard design for a given number of factors. In practice, we used the Nemrodw software [4] to generate automatically all the Hadamard designs. The $-1/+1$ values in the Hadarmard matrix were replaced by the lower and upper bounds of the interval considered for each factor. Table 1 presents the Hadamard design used for BWA-MEM; each line corresponds to an experiment, *i.e.* to a configuration of the pipeline with the values of each parameter (BWA-MEM parameters and $bestHit$ parameter), the original $-1/+1$ values are mentioned into brackets. It should be noted that Hadamard designs are also frequently used as screening designs when the number of parameters to screen is very large.

The Hadamard experimental plans for BWA-backtrack, TMAP, and Bowtie2 are given in Supplementary Tables 3 to 5. Without the experimental plan, testing all the parameter combinations would have led to 162, 243, 216 and 81 experiments while 18, 9, 40 and 45 experiments were tested for respectively BWA-backtrack, BWA-MEM, Bowtie2 and TMAP, leading thus to a total of 112 experiments. In practice, these 112 experiments were run on each of the 3 simulated *small* studies (Roche 454, IT PGM, Illumina HiSeq, see section 3). As detailed in section 2.3, models were built from the performance observed in these experiments, then these models were used to predict the best pipeline configuration for each mapper and each dataset.

2.3 Performance prediction

2.3.1 Read categories and performance indicators The completeness of the RDB depends on the application. For instance, pathogenic bacteria present in bronchoalveolar lavage sample from patients in intensive care unit

[2] NIST/SEMATECH e-Handbook of Statistical Methods, `http://www.itl.nist.gov/div898/handbook/`, 2015-07-14.

[3] The Ion Torrent flow sequence mapping, see `https://github.com/iontorrent/TMAP`

[4] `http://www.nemrodw.com`

Table 1. Experimental plan for BWA-MEM. Figures within brackets correspond to the $-1/+1$ values in the corresponding Hadarmard matrix.

Run	s.length	zdropoff	reseed.length	clipping	bestHit
1	19 (-1)	80 (-1)	0.5 (-1)	7 (+1)	0 (-1)
2	27 (+1)	80 (-1)	0.5 (-1)	3 (-1)	0.5 (+1)
3	27 (+1)	120 (+1)	0.5 (-1)	3 (-1)	0 (-1)
4	19 (-1)	120 (+1)	1.5 (+1)	3 (-1)	0 (-1)
5	27 (+1)	80 (-1)	1.5 (+1)	7 (+1)	0 (-1)
6	19 (-1)	120 (+1)	0.5 (-1)	7 (+1)	0.5 (+1)
7	19 (-1)	80 (-1)	1.5 (+1)	3 (-1)	0.5 (+1)
8	27 (+1)	120 (+1)	1.5 (+1)	7 (+1)	0.5 (+1)
9	23 (0)	100 (0)	1 (0)	5 (0)	0.25 (0)

are very likely to have at least one genome of their species available in the RDB. On the other hand, species found in environmental samples such as those found in marine ecosystems, may have no RDB representation (Yang *et al.*, 2015; Magasin and Gerloff, 2014).

To cover different levels of RDB completeness, we defined three read categories. *i) Non-stringent*, where the genome from which the read derives was present in the RDB, *ii) Reachable*, where the genome was absent from the RDB, however the RDB included at least two genomes of the same species. *iii) Unreachable*, where the genome was absent from the RDB and the RDB did not include any genome of the same species. Therefore the best possible prediction for an unreachable read was the LCA among the closest genomes present in the RDB (see Supplementary Figure 1).

Considering these three read categories, we introduced five performance indicators: *i)* the proportion of mapped reads, *ii)* the proportion of non-stringent mapped reads mapped at the correct position along the correct genome sequence, *iii)* the pipeline running time, *iv)* the proportion of mapped reachable reads correctly predicted at the species rank, abbreviated "reachable" thereafter, *v)* the proportion of mapped unreachable reads correctly predicted at the expected rank or at a higher rank (*i.e.* the predicted taxon should be the best possible prediction, or any taxon among its parents), abbreviated "unreachable" thereafter. The first three indicators referred to the intrinsic quality of the mapper, while the other two indicators evaluated the performance of the whole taxonomic binning pipeline.

2.3.2 Models For each mapper, the taxonomic binning pipeline was run for all the configurations given by the experimental plan of the mapper. For example, BWA-MEM pipeline was run for the 9 configurations described in Table 1. The five individual performance indicators described above were computed from the outputs of each pipeline run. Then the performance indicators were modeled using logistic and Weibull regression models. The logistic model (McCullagh and Nelder, 1989) was used for proportion performance indicators (lying in $[0;1]$), and the Weibull model (Kalbfleisch and Prentice, 2011) was used for the running time. A linear effect for all the continuous covariates was assumed without interaction. For instance, the following logistic model was used to study the effect of parameters and to predict the proportion of reads mapped with BWA-MEM: $logit(p_i) = \beta_0 + \beta_1 \times s.length_i + \beta_2 \times zdropoff_i + \beta_3 \times reseed.length_i + \beta_4 \times clipping_i + \beta_5 \times bestHit_i$, with $i = 1 \ldots 9$, with p_i the observed proportion of mapped reads, $s.length_i$ the seed length value, $zdropoff_i$ the off-diagonal X-dropoff value, $reseed.length_i$ the trigger re-seeding, $clipping_i$ the clipping penalty, and $bestHit_i$ the best hits threshold value in the $i^{th}, i = (1 \ldots 9)$ row of the BWA-MEM experimental plan (see Table 1).

2.3.3 Optimal mapper configuration Using the previous models, it was then possible to predict each performance indicator, for any combination of the values of the parameters. For example, using the BWA-MEM mapper, the proportion of mapped reads can be predicted for $s.length = 23$, $zdropoff = 80$, $reseed.length = 1$, $clipping = 5$, and $bestHit = 0$ (although this pipeline configuration was not included in the experimental plan): proportion mapped $= \exp(\hat{\mu}))/(1+\exp(\hat{\mu})))$, with $\hat{\mu} = \hat{\beta_0} + \hat{\beta_1} \times 23 + \hat{\beta_2} \times 80 + \hat{\beta_3} \times 1 + \hat{\beta_4} \times 23 + \hat{\beta_5} \times 5 + \hat{\beta_6} \times 0$, and $\hat{\beta_0}$ to $\hat{\beta_6}$ the parameters of the fitted logistic model for the proportion of mapped reads. In practice, the performance indicators were predicted for all the combinations of the parameter values presented in Supplementary Table 2.

Then, to have a single indicator which acts as a compromise between the five individual performance indicators, we defined a Composite Performance Indicator (CPI) as a weighted sum of the individual performance indicators. Since read indicators had values in $[0;1]$, to compute the CPI we standardized the running time with the minimum taken over all predicted running times for a given mapper: $\mathrm{std.time} = \min(\mathrm{time})/\mathrm{time}$.

Finally, we retained three parameter configurations: the "best", maximizing the CPI, the "worst", minimizing the CPI, and the "default", using default mapper parameters (see Supplementary Table 2) and $bestHit = 0.25$.

3 DATASETS

3.1 Simulated studies

We designed small and large studies, each containing a simulated metagenome sample and a RDB. The small studies were used for the taxonomic binning pipeline optimization, while the large one was used to evaluate how predicted parameter configurations scale to larger and more realistic datasets.

The simulated metagenome used for the small studies included reads from the three categories to cover a large range of applications: 26% of non-stringent reads, 27% of reachable reads, and 47% of unreachable reads. Among unreachable reads, the taxonomic rank of the best possible prediction was genus for 33%, family for 11%, order for 18%, class for 27%, and phylum for 11%. Based on this metagenome, three datasets of 500,000 reads each were simulated using Grinder 0.5.3 (Angly *et al.*, 2012) for three sequencing technologies: Roche 454 (average read length=450 bp, sd=50), Ion Torrent PGM (average read length=200 bp, sd=20), and Illumina HiSeq (read length=100 bp) (see Grinder profiles in Supplementary Data, section 2). The corresponding small RDB contained 356 sequences from 52 species. Thereafter, we refer to the *small454*, *smallPGM* and *smallHiSeq* studies while mentioning the simulated reads described above and the small RDB on which they were mapped.

The metagenome simulated for the large study included only reachable reads from 287 bacterial strains, 2 fungi and 7 archaeal bacteria. For this metagenome, we simulated an Illumina HiSeq dataset of 12.5 10^6 read pairs, using the same Grinder profile as for the smallHiSeq. The corresponding large RDB was the HMP RDB (see below). Thereafter, the *largeHiSeq* study refers to the large Illumina HiSeq dataset mapped against the large HMP RDB.

3.2 HMP mock community

We validated the optimized configuration on a spiked dataset, the HMP Microbial Mock Community dataset (Even, Low Concentration, 454 GS FLX Titanium, SRA accession SRX030841). This dataset contains 1,386,198 reads from a mock sample made of a genomic DNA mixture obtained from 20 bacterial plus 1 archaeal spiked strains. The detailed list of the spiked organisms can be found in Martin *et al.* (2012). Reads with quality score $<$ 20 were trimmed, and reads with length $<$ 25 bp were filtered out. A RDB was built from the HMP Reference Genome Database (Martin *et al.*, 2012). Sequences without reference to the NCBI taxonomy, redundant sequences, very short sequences ($<$ 100 bp), and sequences with a number of ambiguous bases $>$ 0.01% of the sequence length were filtered out. This RDB included 179,988 of the 188,039 sequences initially present in the HMP database, corresponding to 3941 species.

Since the HMP mock community only contained non-stringent and reachable reads, we favored reachable and non-stringent indicators in the

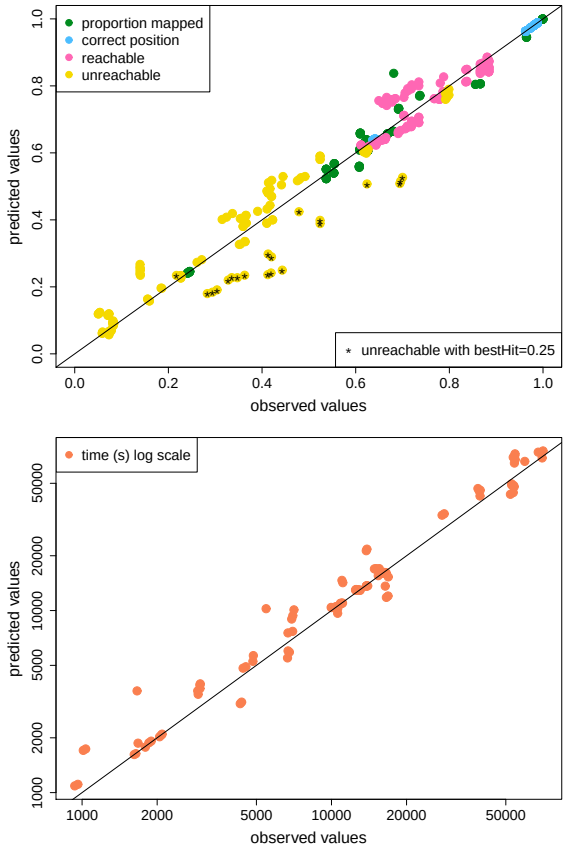

Fig. 1. Models goodness-of-fit for the 112 experiments of the experimental plan (small454 study, all mappers). The upper panel presents the predicted *vs.* observed values for $[0; 1]$ performance indicators: the proportion of mapped reads (pink dots), the proportion of reads mapped at the correct position (blue dots), the proportion of correct predictions among reachable reads (green dots), and the proportion of correct predictions among unreachable reads (yellow dots). The lower panel presents the predicted *vs.* observed running time, in seconds. Unreachable indicators for experiments with $bestHit = 0.25$ are identified with an asterisk.

CPI. So for this study we used weights 1, 2, 0.8, 4, and 0.3 for the five indicators (resp. proportion mapped, correct position, time, reachable and unreachable).

In the following, for the optimization study, we focused on results obtained on the small454 study as the HMP mock community was sequenced on this platform. (Results obtained on the smallPGM and smallHiSeq studies are available as Supplementary Data).

4 RESULTS

4.1 Models validation and impact of the parameters

Figure 1 presents the predicted *vs.* observed values for all the performance indicators. "Observed" values are the performance indicators obtained on the 112 experiments of the experimental plan (including all the mappers), while "predicted" values are the performance predicted with the logistic and Weibull models. The upper panel corresponds to the $[0; 1]$ performance indicators predicted with logistic models, while the lower panel corresponds to the running time predicted with a Weibull model. The fit of the models was pretty good, except for the unreachable indicator corresponding to the experiments with $bestHit = 0.25$, for which we observed a departure from the linearity assumption. Similar results were obtained from the smallPGM and smallHiSeq studies (see Supplementary Figures 2 and 3).

Figure 2 presents the odds ratios and the hazard ratios of the parameter effects (estimated respectively from the logistic and the Weibull models), obtained on the small454 study, for each mapper and performance indicator. An odds ratio above 1 corresponds to a positive effect on the proportion performance indicators and a hazard ratio above 1 corresponds to a decrease of the running time.

With BWA-MEM, the $s.length$ parameter had an effect on the proportion of mapped reads. As expected, longer seeds led to a smaller proportion of mapped reads. Furthermore, higher $clipping$ penalty increased the proportion of reads mapped at the correct position.

The TMAP $clipping$ parameter had an important effect on several indicators. As expected, limited clipping decreased the proportion of mapped reads, but increased the proportion of reads mapped at the correct position. Limited clipping also increased the proportion of correct predictions among unreachable reads.

The Bowtie2 $mode$ (local or end-to-end) was the most important parameter. Local modes ("L1" and "L2") had negative effects compared to end-to-end ("ETE1") mode on all the indicators, excepted on the proportion of mapped reads.

BWA-backtrack was less sensitive than other mappers to parameters modification, except for the $iterative$ parameter that decreased (*resp.* increased) the proportion of correct predictions among reachable (*resp.* unreachable) reads, and decreased the mapping time when the -N option was used. Decreasing the number of errors in the seed ($s.error$) also decreased the mapping time.

The $bestHit$ had no effect on the proportion of mapped reads nor the proportion of reads mapped at the correct position because this parameter was only used for LCA calculation after the mapping steps. This parameter strongly impacted the performance of the reachable and unreachable indicators. Larger $bestHit$ values increased the proportion of correct predictions among unreachable reads because more mapping hits were considered for the LCA calculation, and conversely decreased the proportion of correct predictions among reachable reads because prediction ranks were too high (not specific enough). Hence, no configuration maximizing all the indicators simultaneously could be found. That explained why we had to find a tradeoff between indicators using a weighted score.

For the smallPGM study, the impact of the mapper parameters on performance indicators was very similar to what was observed on the small454 study, while some specific effects were observed for the smallHiSeq study, such as the effect of the $iterative$ parameter on BWA-backtrack correct positions (see Supplementary Figures 4 and 5).

4.2 Configurations comparison

The best and the worst configurations as well as the default configuration were run for each mapper. Figure 3 presents the predicted *vs.* observed CPI values obtained for these three configurations. Observed and predicted values were close, highlighting the high generalization ability of the prediction models, except for configurations with $bestHit = 0.25$ (this is the case for

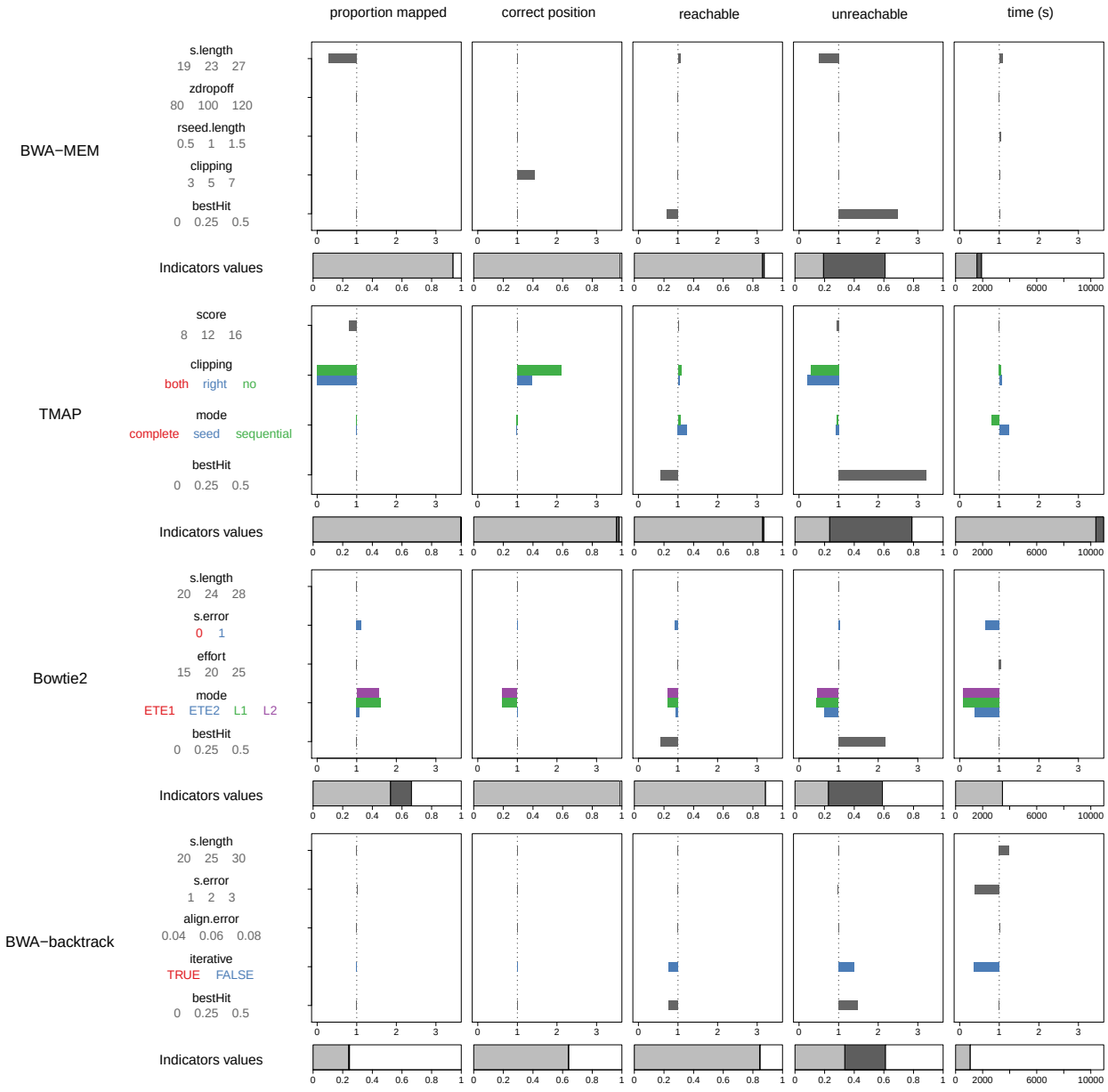

Fig. 2. Overview of the parameter effects for the small454 study. Each plot presents the effect of the parameters (in rows) on an indicator (in columns), for a given mapper. Columns 1 to 4 present the odds ratios of the parameter effects on the proportion of mapped reads, the proportion of reads mapped at the correct position, the proportion of correct predictions among reachable reads and the proportion of correct prediction among unreachable reads. An odds ratio above 1 corresponds to a positive effect on the performance indicator. Column 5 presents the hazard ratio of the parameter effect on the running time; an hazard ratio above 1 corresponds to a decreased running time. Categorical parameters are represented by colored bars: each level is compared to the reference level (written in red). Below each plot a barplot shows the indicator values for the best configuration (light grey bars for columns 1 to 4 and dark grey bars for column 5) as well as the maximum value of the indicator (dark grey bars for columns 1 to 4 and light grey bars for column 5). The dark grey segments thus represent on all barplots the cost on each indicator to build the best configuration.

all default scenarios), for which the predictions are globally above the observed values.

As expected, the CPI increased from the worst, then to the default, and then to the best pipeline configuration. This highlights the ability of our method to identify configurations that improve performance. The same results were observed in the smallPGM and the smallHiSeq studies (see Supplementary Figures 8 and 9). BWA-MEM presented a small difference of CPI between the default and the best configurations, meaning that the default parameters for BWA-MEM were already optimal in our context.

Table 2. Predicted performance indicators for the four mappers (BWA-MEM, TMAP, Bowtie2, and BWA-backtrack) and the three sequencers (small454, smallPGM, and smallHiSeq studies). "Best" values are obtained with the best configuration, and "Max" (resp. "Min") are the maximum values (resp. minimum value, for the running time) predicted for each indicator.

Sequencer	Mapper	proportion mapped		correct position		reachable		unreachable		time (s)	
		Best	Max	Best	Max	Best	Max	Best	Max	Best	Min
Roche 454 (450 bp)	BWA-MEM	0.95	0.95	0.99	0.99	0.86	0.88	0.19	0.61	1924	1583
	TMAP	1.00	1.00	0.96	0.98	0.87	0.87	0.23	0.79	10960	10376
	Bowtie2	0.52	0.66	0.99	0.99	0.89	0.89	0.23	0.59	3468	3468
	BWA-backtrack	0.24	0.25	0.64	0.64	0.85	0.85	0.34	0.61	1090	1089
IT PGM (200 bp)	BWA-MEM	0.87	0.87	0.99	0.99	0.82	0.84	0.17	0.45	633	517
	TMAP	1.00	1.00	0.96	0.98	0.83	0.84	0.30	0.80	3978	3266
	Bowtie2	0.53	0.64	0.99	0.99	0.85	0.85	0.26	0.55	1325	1307
	BWA-backtrack	0.41	0.42	0.56	0.56	0.83	0.83	0.37	0.58	376	358
Illumina Hiseq (100 bp)	BWA-MEM	0.86	0.86	0.57	0.57	0.89	0.91	0.37	0.41	252	197
	TMAP	0.67	1.00	0.90	0.94	0.78	0.81	0.11	0.81	1119	1119
	Bowtie2	0.54	0.62	0.97	0.97	0.81	0.81	0.29	0.49	796	748
	BWA-backtrack	0.54	0.54	0.96	0.99	0.79	0.79	0.40	0.54	98	98

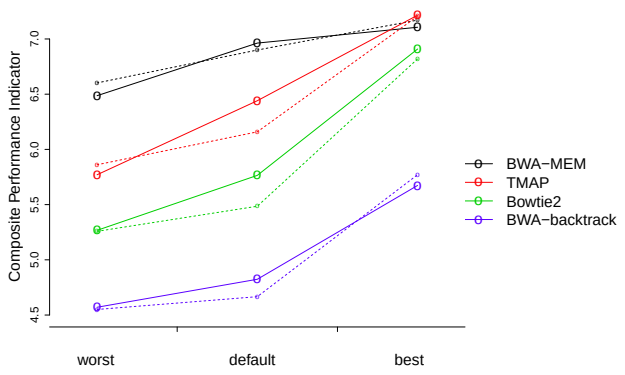

Fig. 3. Worst, default, and best configurations in the small454 study. Observed (dashed lines) and predicted (solid lines) Composite Performance Indicator for the worst, default, and best configurations, for each mapper.

The best configurations for the three datasets and the four mappers can be found in Supplementary Table 6.

4.3 Mappers comparison

Table 2 presents the predicted performance indicators for each mapper and sequencer, obtained with the best configuration, along with the maximum value of each performance indicator.

In the small454 study, TMAP mapped 100% of the reads, BWA-MEM mapped more than 95% of the reads, and Bowtie2 could map a maximum of 66% of reads but the best configuration only mapped 52%. BWA-backtrack did not manage to map more than a quarter of the reads. Indeed, BWA-backtrack was developed and optimized to align short reads (less than $<$100 bp) and failed to map the Roche 454 simulated reads which had an average length equal to 450bp. The proportion of reads mapped at the correct position among mapped non-stringent reads was close to 100% for all mappers, except for BWA-backtrack.

As the weight of the reachable indicator was very high in the CPI, all the mappers had a reachable indicator value close to their maximum value for the best configuration. However, the price to reach this maximum value was paid for at the expense of the unreachable indicator. This was true for all the mappers. The largest drop was observed for TMAP, which could theoretically reach 79% for the unreachable indicator but got 23% under the best configuration. Using more balanced weights for the reachable and unreachable indicators would reduce this gap.

The shortest running times were obtained by BWA-backtrack and BWA-MEM, while TMAP running time was particularly long.

Performance indicator values predicted in the smallPGM study were close to those predicted in the small454 study. BWA-backtrack presented better performance in the smallHiSeq study than in the two other studies, with a larger predicted proportion of mapped reads, and reads mapped at a correct position. As expected, the predicted running time decreased with the read length (see Supplementary Figure 6).

4.4 Generalization to a larger metagenomic sample and RDB with different composition

We evaluated how predictions generalized to both a larger RDB and to a larger metagenomic dataset with a different taxonomic composition. We evaluated the pipeline performance on the largeHiSeq study under the best, default and worst configurations obtained on the smallHiSeq study, for the four mappers.

First, Bowtie2 failed to scale to realistic-sized metagenomic samples and RDBs: worst and default configurations were still running after 30 days. Second, we observed similar profiles for all the performance indicators for the three remaining mappers (see Figure 4). Even though profiles were similar for the proportion of mapped reads, the values obtained on the largeHiSeq study were higher because it only included reachable reads while the smallHiSeq study also included unreachable reads, leading thus to a smaller proportion of mapped reads. We also observed a drop of the reachable indicator between the worst and default

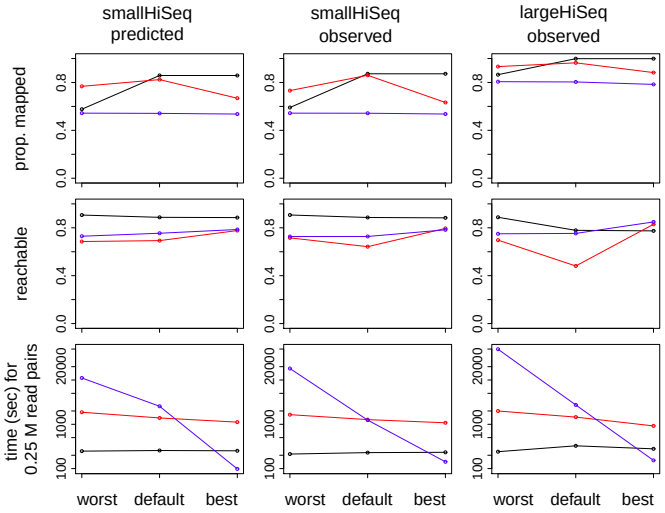

Fig. 4. Generalization of the performance indicators to the largeHiSeq study. The first and second column present the predicted and observed performance indicators for the smallHiSeq study, and the third column the observed performance indicators for the largeHiSeq study. The proportion of reads mapped at the correct position and the unreachable indicator are not presented as they cannot be computed on the large dataset which only contains reachable reads. Each plot presents a performance indicator values for the worst, default and best configurations, for BWA-MEM (black line), TMAP (red line), and BWA-backtrack (blue line).

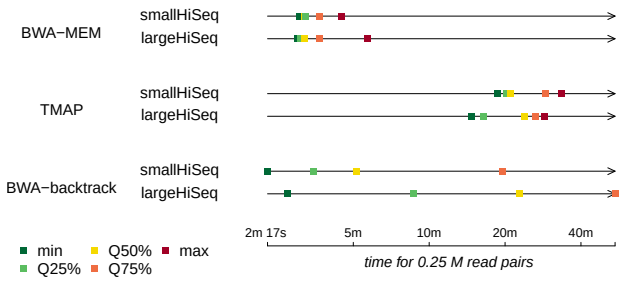

Fig. 5. Generalization of the running time to the large HiSeq study. Running times (scaled to 250,000 reads) observed for 3 mappers in the smallHiSeq and largeHiSeq studies for 5 configurations. The 5 configurations correspond to the minimum, maximum and the three quartiles of the predicted time distribution obtained in the smallHiSeq study.

configurations for TMAP. This drop was already observed between predicted and observed smallHiSeq but was greater on largeHiSeq. Although it was not testable with the Hadamard design, we suspect here an interaction between the $clipping$ and $bestHit$ parameters. Indeed, the default configuration allows right clipping with $bestHit$ =0.25, while the worst configuration does not allow clipping, with $bestHit$ = 0.5. Allowing clipping with a non-null $bestHit$ value tends to make more predictions at higher ranks that non allowing clipping with a very large $bestHit$ value. We also compared the observed running times in the smallHiSeq and largeHiSeq studies for 5 configurations, corresponding to the minimum, maximum and the three quartiles of the predicted time distribution obtained in the smallHiSeq study. Figure 5 presents the running times scaled for 250,000 pairs of reads for BWA-MEM, TMAP, and BWA-backtrack. All the configurations for Bowtie2 and the configuration corresponding to the maximum running time for BWA-backtrack are not presented, because running times exceeded 30 days. For BWA-MEM and TMAP the scaled running times were highly similar for the two studies. For BWA-backtrack we observed the same running time ordering between the two studies, but the observed running time values were not equal when scaled for the number of reads. Thus, contrary to BWA-MEM and TMAP, the BWA-backtrack running time did not increase linearly with the number of reads in the dataset but was also impacted by the size/composition of the RDB.

4.5 Validation on HMP mock community

We applied the taxonomic binning pipeline on the HMP mock community with the best, default and worst configurations obtained on the small454 study, for the three mappers suitable for Roche 454 reads (BWA-MEM, TMAP and Bowtie2). Figure 6 presents the predicted proportion of the 21 spiked organisms at two taxonomic ranks, species and family. At each rank, a barplot represents the proportions of *i)* unmapped reads, *ii)* reads assigned at a higher taxonomic rank, *iii)* reads assigned to a wrong taxon (*i.e.* not in the direct lineage of one of the 21 spiked strains) and *iv)* reads assigned to each of the 21 expected strain lineages. As a reference for the expected proportions of spiked strains, we used the relative genome depth obtained by Martin *et al.* (2012) with the CLC mapper on the same sample (sequenced on Illumina GAIIx in Martin *et al.* (2012) article).

Although results presented by Martin *et al.* (2012) were obtained with another sequencer, the taxon proportions attributed to the spiked organisms were highly consistent with the proportions we observed. Unfortunately, the proportion of unmapped reads and reads incorrectly assigned to other species could not be retrieved from Martin *et al.* (2012). Thus, we also presented the predicted proportion of the 21 spiked organisms at the species and family ranks, after removal of unmapped, incorrect, and reads assigned at higher ranks (see Supplementary Figure 10). We observed a lower consistency between HMP predictions and Bowtie2 and TMAP predictions for the default and worst configurations. Best configurations and HMP results were still highly consistent, whatever the mapper.

For each mapper, the best configuration retrieved all the spiked organisms, and the proportion of reads assigned at the species level was very high. Indeed, by definition, the reachable indicator contributed highly to the CPI, leading to an increased proportion of reads correctly predicted at the species rank under the best configuration, compared to the default and worst configurations. For all the mappers under the best configuration, 93% to 95% of the mapped reads were assigned at the species rank while this proportion dropped to 60% for TMAP and Bowtie2 when using the default configuration and to 49% for Bowtie2 when using the worst configuration. These results were consistent with the simulations.

Unfortunately, at the species rank, we also observed that TMAP default settings predicted more reads at a higher rank than the worst configuration. The same behaviour was observed in the optimization study on the small454 study: as can be seen in Supplementary Figure 6, the *predicted* reachable indicator presented a 10 points increase

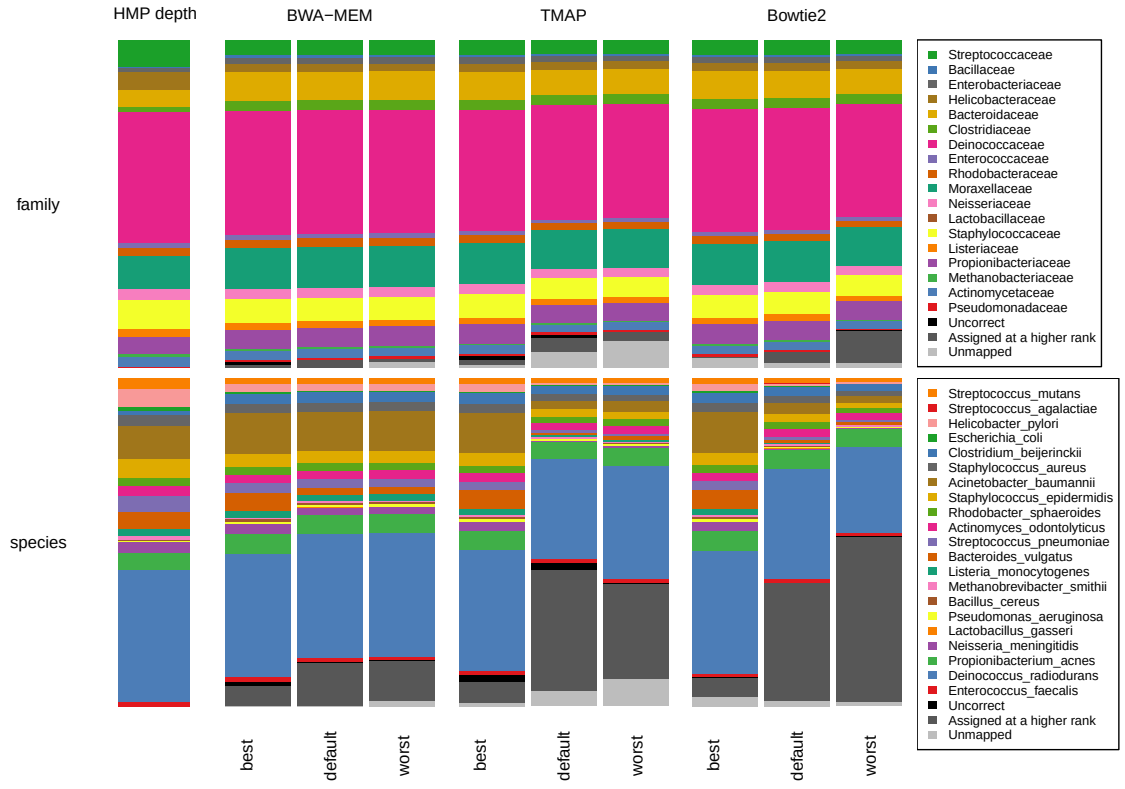

Fig. 6. HMP mock community metagenomic profiles Proportion of the 21 spiked organisms at species and family ranks obtained using BWA-MEM, TMAP, and Bowtie2 mappers with the three retained configurations (best, default and worst), with the proportion of unmapped reads shown in light grey, the proportion of reads assigned at a higher rank than the one considered in dark grey, and the proportion of reads assigned to a wrong taxon (taxon not in the lineage of the 21 spiked organisms) in black. The first column (HMP depth) presents the relative genome depth obtained by Martin *et al.* (2012).

between the worst and default configurations with TMAP, while the *observed* reachable indicator was stable between the worst and default configurations (see Supplementary Figure 7). As already mentioned, this drop was probably due to an interaction between the $clipping$ and the $bestHit$ parameter.

Finally, BWA-MEM under the best configuration was at least 3.8 times faster than both other mappers (see Supplementary Table 7).

5 DISCUSSION AND CONCLUSION

In this article, we proposed a strategy based on the DOE methodology to compare mappers, study the effect of parameters, and select the configuration that optimizes the taxonomic binning performance, for three main sequencing technologies in the field of metagenomics.

To this purpose, we designed an optimization experimental plan to find the best pipeline configuration while running a minimal number of configurations on simulated datasets. The DOE methodology made it possible to obtain precise parameter estimations at a minimal experimental cost. Based on previously published studies, we selected candidate parameters and then used a Hadamard design to find the pipeline configuration leading to the optimal performance. We could also have used in a first step, a screening experimental design to select the most influential parameters (instead of relying on published benchmarks) but given the large number of conditions we considered (4 mappers and 3 sequencing technologies), carrying out an experimental design for both the screening and the optimization steps was too cumbersome. For this large-scale optimization study, we assumed a linear effect of parameters and the absence of interactions, leading us to use a Hadamard design. For most of the configurations, we observed a good consistency between observed and predicted performance indicators (except for a possible interaction effect between the *clipping* and *bestHit* parameters on the reachable indicator for TMAP). A possible extension of this work would be to focus on a particular mapper and a read sequencing technology, and to use an alternative optimization experimental plan that allows us to study quadratic effect of parameters and interactions. For example, we could use classical surface response experimental plans (*e.g* central composite, Doehlert, Box-Behnken designs (Lundstedt *et al.*, 1998)) relaxing the previous assumptions at the cost of more experiments, but still reasonable for such a tailored application.

Contrary to previous studies, we defined performance indicators related to both the performance of the mapper itself (proportion of mapped reads, reads mapped at the correction position, and running time) and to the whole taxonomic binning strategy (reachable and unreachable indicators). The unreachable indicator was particularly important to evaluate the binning strategy in case

of incomplete RDB. However, this indicator favored high rank predictions (*e.g.* superkingdom) for unreachable reads because best possible prediction and correct prediction at a higher rank were both considered to be correct. To limit this effect, predictions made at a higher taxonomic rank than the rank of the best possible prediction could be down-weighted (*e.g.* using the distance between the prediction rank and the best possible rank).

Interestingly, our study showed that for each mapper, no more than two mapper parameters had a significant effect on the performance. It also highlighted the importance of the best hit threshold on the reachable/unreachable performance indicators.

Moreover, to select the best configuration, we defined a CPI as the weighted sum of the five performance indicators where the choice of the weights highly depends on the targeted application. The reachable indicator received a higher weight, because the HMP RDB was complete (in terms of reference sequences corresponding to the spiked species), and low-rank taxonomic predictions were preferred. In contrast, with incomplete RDB, the unreachable indicator should received a higher weight. For resequencing applications, the proportion of mapped reads, the proportion of reads mapped at the correct position, and the time would be the only three indicators with a non-zero weight. Hence, the advantage of the CPI is to cover many applications using the same prediction model, just by adjusting the weights of the performance indicators.

Finally, we evaluated the ability of the predicted performance to scale to both larger datasets and RDB with a different composition than the ones used in the optimization study. Although, our main indicators appear to well generalize, it would be safer for a targeted application to use the reference database envisioned for the final application when running the experiments for the pipeline performance optimization. As regards the simulated metagenomic dataset, it should have ideally the same taxonomic complexity than the real metagenomic samples, but with a limited number of reads so that the optimization study remains computationally feasible.

In conclusion, we hope this computationally tractable statistical framework will contribute to improve taxonomic binning performance by providing rational criteria to optimize any mapper configuration for a given application.

REFERENCES

Angly, F. E. *et al.* (2012). Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **40**(12), e94–e94.

Břinda, K. *et al.* (2015). RNF: a general framework to evaluate NGS read mappers. *arXiv preprint arXiv:1504.00556*.

Caboche, S. *et al.* (2014). Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC Genomics*, **15**(1), 264.

Dröge, J., Gregor, I., and McHardy, A. (2015). Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics*, **31**(6), 817–824.

Hatem, A. *et al.* (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, **14**(1), 184.

Holtgrewe, M. *et al.* (2011). A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinformatics*, **12**(1), 210.

Hugenholtz, P. *et al.* (2002). Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, **3**(2), 1–0003.

Huson, D. H. *et al.* (2007). Megan analysis of metagenomic data. *Genome Res.*, **17**(3), 377–386.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data, second edition*. John Wiley & Sons.

Koslicki, D., Foucart, S., and Rosen, G. (2014). Wgsquikr: fast whole-genome shotgun metagenomic classification. *PloS one*, **9**(3), 91784.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. methods*, **9**(4), 357–359.

Li, H. (2012). Exploring single-sample snp and indel calling with whole-genome de novo assembly. *Bioinformatics*, **28**(14), 1838–1844.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**(14), 1754–1760.

Liu, B., Gibbons, T., Ghodsi, M., and Pop, M. (2010). Metaphyler: Taxonomic profiling for metagenomic sequences. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on*, pages 95–100. IEEE.

Lundstedt, T., Seifert, E., Abramo, L., Thelin, B., Nyström, Å., Pettersen, J., and Bergman, R. (1998). Experimental design and optimization. *Chemometrics and intelligent laboratory systems*, **42**(1), 3–40.

Magasin, J. D. and Gerloff, D. L. (2014). Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism. *Bioinformatics*, page btu546.

Mande, S. S. *et al.* (2012). Classification of metagenomic sequences: methods and challenges. *Brief. Bioinformatics*, page bbs054.

Martin, J. *et al.* (2012). Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS ONE*, **7**(6), e36427.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, second edition*. Chapman and Hall London.

Padmanabhan, R. *et al.* (2013). Genomics and metagenomics in medical microbiology. *J. Microbiol. Methods*, **95**(3), 415–424.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, pages 305–325.

Ruffalo, M. *et al.* (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, **27**(20), 2790–2796.

Schbath, S. *et al.* (2012). Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *J. Comput. Biol.*, **19**(6), 796–813.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, **9**(8), 811–814.

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., *et al.* (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*, **10**(12), 1196–1199.

Yang, X.-J., Wang, Y.-B., Zhou, Z.-W., Wang, G.-W., Wang, X.-H., Liu, Q.-F., Zhou, S.-F., and Wang, Z.-H. (2015). High-throughput sequencing of 16s rdna amplicons characterizes bacterial composition in bronchoalveolar lavage fluid in patients with ventilator-associated pneumonia. *Drug design, development and therapy*, **9**, 4883.