

Sequence analysis

# Spaced seeds improve *k*-mer-based metagenomic classification

Karel Břinda<sup>†</sup>, Maciej Sykulski<sup>†</sup> and Gregory Kucherov\*

LIGM/CNRS, Université Paris-Est, 77454 Marne-la-Vallée, France

\*To whom correspondence should be addressed.

<sup>†</sup>The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Inanc Birol

Received on April 1, 2015; revised on June 11, 2015; accepted on July 14, 2015

## Abstract

**Motivation:** Metagenomics is a powerful approach to study genetic content of environmental samples, which has been strongly promoted by next-generation sequencing technologies. To cope with massive data involved in modern metagenomic projects, recent tools rely on the analysis of *k*-mers shared between the read to be classified and sampled reference genomes.

**Results:** Within this general framework, we show that *spaced seeds* provide a significant improvement of classification accuracy, as opposed to traditional *contiguous k*-mers. We support this thesis through a series of different computational experiments, including simulations of large-scale metagenomic projects.

**Availability and implementation, Supplementary information:** Scripts and programs used in this study, as well as supplementary material, are available from <http://github.com/gregorykucherov/spaced-seeds-for-metagenomics>.

**Contact:** [gregory.kucherov@univ-mlv.fr](mailto:gregory.kucherov@univ-mlv.fr)

## 1 Introduction

Metagenomics is a powerful approach to study genetic material contained in environmental samples. The advent of high-throughput sequencing technologies (known as *Next-Generation Sequencing*, NGS) revolutionized this approach by avoiding the need of cloning the DNA and thus greatly facilitating the obtention of metagenomic samples, at the same time drastically decreasing its price. Present-day metagenomic studies focus on various bacterial, fungal or viral populations, exemplified by the Human Microbiome project (Methé *et al.*, 2012) that investigates microbial communities at different sites of human body.

Modern metagenomics deals with vast sequence datasets. On the one hand, metagenomic samples (*metagenomes*) obtained through NGS are commonly measured by tens of billions of bp (Nath Meth Editorial, 2009). These sequences generally come from a number of different species, some of which either have a previously sequenced reference genome or have a related sequenced species sufficiently close phylogenetically to determine this relationship by sequence

comparison. Other sequences, however, may come from organisms that have no sufficiently close relatives with sequenced genomes or from DNA fragments that show no significant similarity with any available genomic sequence. The *metagenomic classification* problem is to assign each sequence of the metagenome to a corresponding taxonomic unit or to classify it as ‘novel’.

A way to improve the accuracy of metagenomic classification is to match the metagenome against as large set of known genomic sequences as possible. This set, stored in a working reference database, may contain thousands of full microbial genomes (Wood and Salzberg, 2014).

*Alignment-based classifiers* (Mandé *et al.*, 2012) proceed by aligning metagenome sequences to each of the genomes in the reference database. This can be done with generic BLAST-like alignment programs (Altschul *et al.*, 1997). While this approach can be envisaged for small datasets (both metagenome and database) and is actually used in such software tools as MEGAN (Huson *et al.*, 2011) or PHYMMBL (Brady and Salzberg, 2011) (see (Mandé *et al.*, 2012) for

more), it is unfeasible on the scale of modern metagenomic projects. On the other hand, there exists a multitude of specialized tools for aligning NGS reads—BWA (Li and Durbin, 2009), NOVOALIGN (<http://www.novocraft.com/>), GEM (Marco-Sola *et al.*, 2012), BOWTIE (Langmead *et al.*, 2009), just to mention a few popular ones—which perform alignment at a higher speed and are adjusted to specificities of NGS-produced sequences. Still, aligning multimillion read sets against thousands of genomes remains computationally difficult even with optimized tools. Furthermore, read alignment algorithms are usually designed to compute high-scoring alignments only and are often unable to report low-scoring alignments. As a result, a large fraction of reads may remain unmapped (Lindgreen *et al.*, 2015).

Several techniques exist to reduce the computational complexity of this approach. One direction is to pre-assemble reads into longer contigs, potentially improving the accuracy of assignment. Assembly of metagenomic reads has been a subject of many works (Teeling and Glockner, 2012) and remains a fragile approach, due to its error-proneness and high computational complexity. Overall, it appears feasible mostly for small-size projects with relatively high read coverage.

To cope with increasingly large metagenomic projects, alignment-free methods have recently come into use. Those methods do not compute read alignments, thus do not come with benefits of these, such as gene identification. Alignment-free sequence comparison is in itself an established research area, reviewed in a recent dedicated special issue (Vinga, 2014). Most of alignment-free comparison methods are based on the analysis of words, usually of fixed size (*k*-mers), occurring in input sequences. A popular approach is to compute the distance between *frequency vectors* of all *k*-mers in each of the sequences. In the context of metagenomics, however, when one of the sequences is short (NGS read), the analysis is based on the shared *k*-mers, without taking into account their multiplicities in reference genomes. This is also dictated by the prohibitive computational cost of computing and storing *k*-mer multiplicities for metagenomics-size data.

Two recently released tools—LMAT (Ames *et al.*, 2013) and KRAKEN (Wood and Salzberg, 2014)—perform metagenomic classification of NGS reads based on the analysis of shared *k*-mers between an input read and each reference genome. Given a taxonomic tree involving the species of the database, those tools ‘map’ each read to a node of the tree, thus reporting the most specific taxon or clade the read is associated with. Mapping is done by sliding through all *k*-mers occurring in the read and determining, for each of them, the genomes of the database containing the *k*-mer. Based on obtained counts and tree topology, LMAT and KRAKEN algorithms assign the read to the tree node ‘best explaining’ the counts. Further similar tools have been published during last months (Kawulok and Deorowicz, 2015; Ounit *et al.*, 2015).

The goal of this work is to show that the metagenomics classification based on the analysis of shared *k*-mers can be improved by using *spaced k-mers* rather than contiguous *k*-mers.

The idea of using spaced *k*-mers goes back to the concept of *spaced seeds* for *seed-and-extend* sequence comparison (Burkhardt and Kärkkäinen, 2003; Ma *et al.*, 2002). There, the idea is to use as a seed (i.e. local match triggering an alignment) a sequence of matches interleaved with ‘joker positions’ holding either matches or mismatches. The pattern specifying the sequence of matches and jokers is called spaced seed. Remarkably, replacing contiguous seeds by spaced seeds in sequence alignment significantly improves the sensibility-selectivity trade-off with almost no incurred computational overhead. This has been first observed by Ma *et al.* (2002)

and then extended and formally analyzed in a series of further works, see Brown (2008) and Kucherov *et al.* (2006) and references therein.

Recently, it has been reported in several works that spaced seeds bring an improvement in alignment-free comparison as well. Leimeister *et al.* (2014) show that comparing frequency vectors of *spaced k-mers* (*k*-mers obtained by sampling must-match positions of one or several spaced seeds), as opposed to contiguous *k*-mers, leads to a more accurate estimation of phylogenetic distances and, as a consequence, to a more accurate phylogenetic reconstruction. Morgenstern *et al.* (2015) study another measure—the number of pairs of matching (not necessarily aligned) spaced *k*-mers between the input sequences—and show that it provides an even better estimator of phylogenetic distance. In Noé and Martin (2014), it is shown that the number of hits of appropriately chosen spaced seeds (i.e. the number of pairs of matching *aligned* spaced *k*-mers between the two sequences) and their *coverage* (i.e. the total number of matches covered by all hits) provides a much better estimator for the alignment distance than the same measures made with contiguous seeds. From a machine learning perspective, Ghandi *et al.* (2014) and Onodera and Shibuya (2013) show that spaced seeds provide better string kernels for SVM-based classification, confirmed by experiments with protein classification (see also Noé and Martin, 2014).

In this work, we show that using spaced *k*-mers significantly improves the accuracy of metagenomic classification of NGS reads as well. To support this thesis, we consider different scenarios. As a test case, we first study the problem of discriminating a read between two genomes, i.e. determining which of the two genomes is ‘phylogenetically closer’ to the read. We then analyze the correlation between the score of an alignment of a read to a genome and the seed count for this read, defined either as the number of hits or as the coverage. This analysis provides an insight on how well one can estimate the similarity between a read and a genome out of *k*-mer occurrences. Finally, we make a series of large-scale metagenomic classification experiments with KRAKEN software (Wood and Salzberg, 2014) extended by the possibility of dealing with spaced seeds. These experiments demonstrate an improved classification accuracy at the genus and family levels caused by the use of spaced seeds instead of contiguous ones.

## 2 Methods

A spaced seed is a binary pattern over symbols # and - denoting match and joker, respectively. For example, seed #-## specifies a match followed by either match or mismatch followed by two consecutive matches. A seed acts as a mask for comparing short oligonucleotides, for example sequences gaat and gc at differ at the 2nd position, but they match via seed #-## as the 2nd position is masked out. The number of #'s in a seed, called *weight*, defines the number *k* of matching nucleotides. In the above example, *k* = 3 and the matching (spaced) *k*-mer is gat. In a slightly different terms, a seed is a pattern that specifies a small part of a gapless alignment seen as a binary sequence of matches and mismatches. For example, seed #-## occurs in (or *bits*, as usually said) any alignment containing a match followed by another two matches shifted by one position.

When spaced seeds are used for sequence alignment within the *seed-and-extend* paradigm (Brown, 2008), a pair of matching *k*-mers (or, sometimes, a matching sequence of several closely located *k*-mers) indicates a potential alignment of interest in which the two *k*-mers are aligned together. When spaced seeds are used for

alignment-free sequence comparison (Vinga, 2014), the goal is to estimate the similarity of two sequences based on the number of matching  $k$ -mers, with no or limited information about their positions in one or both sequences. This measure can be defined in several different ways.

In the context of metagenomic classification of NGS reads, the goal is to estimate the distance between a short read and a long sequence (genome), which can be modeled by the best alignment score of the read against the sequence. Because of large genome size and large number of reference genomes involved in computations, one of the constraints is to avoid keeping track of positions of  $k$ -mers in genomes. We can only afford constructing an index to quickly answer queries whether a  $k$ -mer occurs in a given genome, without information on its positions. On the other hand,  $k$ -mer positions in the query read can be included to the analysis. Therefore, we have to derive our estimation from the number of  $k$ -mers shared between the read and the genome, together with their positions in the read alone.

One simple estimator is the number of  $k$ -mers in the read that occur in the target genome, we call this measure the *hit number*. However, one may want to favor cases when matching  $k$ -mers cover a larger part of the read versus those with matching  $k$ -mers clumped together due to overlaps. This leads to the concept of *coverage* (Noé and Martin, 2014), earlier used in seed-based alignment algorithms as well (Benson and Mak, 2008; Noé and Kucherov, 2004). The coverage is the total number of positions covered by all matching  $k$ -mers. For example, consider seed #-## and read gaatcagat. Assume the seed hits at positions 1, 4 and 6, i.e.  $k$ -mers g-at, t-ag and a-at occur in the target genome (joker symbol is shown for the sake of clarity). Here, the hit number is 3, and the coverage is 7 as seven positions are covered by hits, namely positions 1, 3, 4, 6, 7, 8 and 9. Hit number and coverage are two main estimators studied in this work.

### 3 Results

#### 3.1 Binary classification

From the machine learning viewpoint, hit number and coverage can be viewed as instances of kernel functions for sequence data (Ben-Hur et al., 2008). Our first step is to compare their capacity with respect to a simple binary classification task. Assume the (impractical) case when our database contains only two genomes. Given a read, we have to decide which of the two genomes the read is closer to. How good can we be at that? Which kernel works better for this task? And are spaced seeds better than contiguous seeds here?

##### 3.1.1 Classifying aligned reads

Consider alignments  $A_s$ ,  $A_l$  of a random read to two genomes and assume they are gapless and have mismatch probabilities  $P_s$ ,  $P_l$  respectively, with  $P_s < P_l$ . Throughout this article, the read length is set to 100, a typical length of ILLUMINA read. Therefore, the alignments can be viewed as random binary strings of length 100 of matches and mismatches, with mismatch probabilities  $P_s$  and  $P_l$  respectively. Given a seed, we compute two counts  $C_s$  and  $C_l$  on alignments  $A_s$  and  $A_l$  respectively, where by ‘count’ we mean, unless otherwise stated, either the hit number or the coverage. For example, if the seed is #-## and the alignment 111101111 (1 stands for match and 0 for mismatch), then the hit number is 3 and the coverage is 7. Note that in this model, common (spaced)  $k$ -mers are assumed to occur necessarily at the same position in the read alignment, although in reality, a  $k$ -mer of the read may not be aligned

**Table 1.** Classification of alignments with different mismatch probabilities

Seed weight	Contiguous		Spaced	
	Hits	Cover	Hits	Cover
$P_s = 0.1, P_l = 0.2$				
16	86/09	87/09	93/05	94/05
18	81/08	81/08	91/05	92/05
20	74/07	74/07	89/05	90/05
22	65/06	65/06	85/05	86/05
24	55/04	56/04	79/04	81/05
$P_s = 0.2, P_l = 0.3$				
16	40/06	40/06	63/07	64/07
18	28/04	28/03	50/05	50/05
20	18/02	18/02	37/03	37/03
22	12/01	12/01	25/02	26/02
24	08/00	08/00	17/01	18/01

Each entry contains a pair ‘Fraction (in %) of correct classification/ Fraction of incorrect classification’. The remaining fraction estimates the probability of a tie. Spaced seeds used for hit number: ####-###-###-###-###-### (weight 16), #####-###-###-###-###-### (weight 18), #####-###-###-###-###-### (weight 20), #####-###-###-###-###-### (weight 22) and #####-###-###-###-###-### (weight 24). Spaced seeds used for coverage: ###-###-###-###-###-### (weight 16), ###-###-###-###-###-### (weight 18), ###-###-###-###-###-### (weight 20), ###-###-###-###-###-### (weight 22) and ###-###-###-###-###-### (weight 24).

with the same  $k$ -mer in the genome. However, here we abstract from this fact.

If  $C_s > C_l$  (respectively,  $C_s < C_l$ ), then we report a correct (respectively, incorrect) classification, otherwise a tie is reported. By iterating this computation, we estimate the probability of correct/incorrect classification for each parameter set.

Some results are presented in Table 1. In all cases, spaced seeds showed a better classification power. In some cases, the difference is striking: for example discriminating between alignments with mismatch probabilities 0.1 and 0.2 (Table 1) using seeds of weight 22 yields 86% of correct classifications with a spaced seed (coverage count) and only 65% with the contiguous seed, whereas the fraction of incorrect classifications is essentially the same. The results also show a slight edge of the coverage count over the hit number, which will be confirmed later on in other experiments as well.

Compared with recent works (Noé and Martin, 2014; Onodera and Shibuya, 2013) showing the advantage of using spaced seeds in SVM kernels, here we demonstrate the superiority of spaced seeds in a very simple classification setting where sequences have to be classified according to identity rate, without a training stage and without resorting to SVM machinery.

##### 3.1.2 Classifying unaligned reads

Let us now turn to a more practical setting, where we want to classify reads coming from a genome  $G$  between two other genomes  $G_1$  and  $G_2$  based on the phylogenetic closeness.

To study this, we implemented the following experimental setup. Using DWGSIM read simulator (<http://github.com/nh13/DWGSIM>), we generate single-end ILLUMINA-like reads from genome  $G$ . In all experiments, we assumed 1% of base mutations (substitutions only) and 2% of sequencing errors (DWGSIM options `-e 0.02 -r 0.01 -R 0`). Given a seed, (contiguous or spaced)  $k$ -mers of  $G_1$  and  $G_2$  are indexed to support existence queries only. For each read, all  $k$ -mers

**Table 2.** Classification of *M.vanbaalenii* reads against *M.smegmatis* and *M.gilvum* genomes

Measure	Seed weight				
	14	16	18	20	22
Contig hit nb	52/41	39/48	24/37	11/24	07/17
Contig cover	54/42	44/47	25/37	11/24	06/17
Contig Jaccard	30/70	35/61	23/42	11/26	06/18
Spaced hit nb	51/40	34/47	20/40	12/32	08/23
Spaced cover	53/42	39/51	21/42	12/32	08/25
Spaced Jaccard	28/72	32/66	20/50	11/33	08/27

Each entry contains a pair 'Fraction (in %) of reads classified closer to *M.smegmatis*/Fraction of reads classified closer to *M.gilvum*'.

**Table 3.** Classification of *B.thuringiensis* reads against *B.anthraxis* and *B.cereus* genomes

Measure	Seed weight				
	14	16	18	20	22
Contig hit nb	83/14	81/11	79/09	77/08	76/08
Contig cover	78/17	80/12	79/09	77/08	76/08
Contig Jaccard	87/13	87/11	85/09	83/08	82/08
Spaced hit nb	83/13	82/11	80/09	79/09	79/08
Spaced cover	80/15	81/11	80/09	79/09	79/08
Spaced Jaccard	88/12	88/11	85/09	84/08	84/08

Each entry contains a pair 'Fraction (in %) of reads classified closer to *B.anthraxis*/Fraction of reads classified closer to *B.cereus*'.

are queried against  $G_1$  and  $G_2$  and corresponding counts  $C_1$  and  $C_2$  are computed. If  $C_1 > C_2$  (respectively,  $C_1 < C_2$ ), the read is classified to be closer to  $G_1$  (respectively,  $G_2$ ), otherwise a tie is reported. Besides considering absolute counts (hit number and coverage), we also considered hit number normalized by the number of distinct  $k$ -mers in the corresponding genome (computed at the indexing stage). This measure approximates the Jaccard index (Bao *et al.*, 2011) and reflects the Bayesian probability of seeing a  $k$ -mer relative to the ' $k$ -mer-richness' of genomes.

Note that the counts are computed here relative to the whole genome, as it is done in the approach of Ames *et al.* (2013) and Wood and Salzberg (2014) (see Section 1). This makes the seed weight an important parameter, as seeds of low weight increase random  $k$ -mer hits that do not evidence any alignment of the read.

We experimented with bacterial genomes of *Mycobacterium* genus. Members of this genus present low interspecies genetic variability and their phylogeny remains uncertain (Tortoli, 2012).

If  $G$  coincides with one of  $G_1$ ,  $G_2$ , i.e. reads have to be discriminated between their source genome and another genome, then all estimators correctly classify nearly all reads as soon as  $G_1$  and  $G_2$  are genomes of distinct species. For example, classifying reads obtained from *Mycobacterium tuberculosis* (H37Rv, acc. NC\_018143) against *M.tuberculosis* itself and *Mycobacterium avium* (104, NC\_008595) led to more than 99% of correct classifications for all estimators.

The more interesting case when  $G$  is distinct from  $G_1$ ,  $G_2$  corresponds to the real-life situation when reads to be classified come from a genome that is not represented in the database. Here, we expect our procedure to determine whether  $G$  is phylogenetically closer to  $G_1$  or to  $G_2$ .

For example, we classified *Mycobacterium vanbaalenii* (Pyr-1, NC\_008726) reads against *Mycobacterium smegmatis* (MC2.155, NC\_018289) and *Mycobacterium gilvum* (Pyr-GCK, NC\_009338)

**Table 4.** Classification of *B.licheniformis* reads against *B.anthraxis* and *B.pumilus* genomes

Measure	Seed weight				
	14	16	18	20	22
Contig hit nb	47/40	24/25	04/07	0.9/3.8	0.3/2.4
Contig cover	49/46	24/25	04/07	0.9/3.8	0.3/2.4
Contig Jaccard	37/62	24/30	04/08	0.8/4.0	0.3/2.6
Spaced hit nb	49/33	22/27	04/10	1.0/5.4	0.4/3.3
Spaced cover	53/39	23/27	04/10	1.0/5.4	0.5/3.3
Spaced Jaccard	41/59	22/35	04/10	0.9/4.6	0.4/3.5

Each entry contains a pair 'Fraction (in %) of reads classified closer to *B.anthraxis*/Fraction of reads classified closer to *B.pumilus*'.

genomes. Alternative phylogenies given in Tortoli (2012) (Figs 1–4) imply different evolutionary relationships among these three species. Our results, shown in Table 2, suggest that *M.vanbaalenii* is closer to *M.gilvum* than to *M.smegmatis*. For non-normalized hit number and coverage estimators, this conclusion is supported by seeds of weight 16 or more, while weight 14 supports the opposite conclusion. This is due to spurious hits that become dominating when the weight drops to 14 and to the larger size of *M.smegmatis* genome (6.99 Mb) compared with *M.gilvum* (5.62 Mb). This effect is corrected by Jaccard index due to normalization by the number of distinct  $k$ -mers (6.09 M for *M.smegmatis* versus 4.96 M for *M.gilvum* for the spaced seed of weight 14). Overall, we observe a significantly sharper discrimination produced by spaced seeds compared with contiguous seeds.

We also performed a series of experiments with the large and genetically variable *Bacillus* genus. Table 3 shows a demonstrative experiment with members of *Bacillus cereus* group: *Bacillus thuringiensis* (serovar konkukian 97-27, NC\_005957), *Bacillus anthracis* (Ames, NC\_003997) and *B.cereus* (ATCC 14579, NC\_004722). The three bacteria are close to the point of being considered to be different lineages of a single *B.cereus* species (Helgason *et al.*, 2000). The results provide a strong support that the *B.thuringiensis* strain is closer to the *B.anthraxis* strain than to the *B.cereus* strain, which agrees with phylogenies reported in the literature (Alcaraz *et al.*, 2010). Indeed, the *B.thuringiensis* and *B.anthraxis* strains have a much higher pairwise identity rate than the former has with the *B.cereus* strain [estimated DNA–DNA hybridization distance 81% versus 45%, as computed by GGDC (Auch *et al.*, 2010)].

However, for species with low sequence similarity, a large majority of reads may have no hits to either genome, and only a small fraction of reads may reveal a significant difference in distance. This situation is illustrated in Table 4 showing the results of classification of *Bacillus licheniformis* (ATCC 14580, NC\_006270) reads against *B.anthraxis* (Ames, NC\_003997) and *Bacillus pumilus* genomes (SAFR-032, NC\_009848). The results support a higher similarity of *B.licheniformis* to *B.pumilus* than to *B.anthraxis*, but the difference is revealed on a very small fraction of reads. The conclusion, however, is significant as those reads represent the majority of reads having any hits to one of the genomes.

In all our experiments reported in this section, spaced seeds showed a better classification capacity. The difference is especially significant in 'nontrivial' cases involving relatively dissimilar genomes, such as those illustrated by Tables 2 and 4. While the difference between hit number and coverage estimators appeared insignificant (in agreement with results of Section 3.1.1), the Jaccard index generally provides a more distinct discrimination and, combined with a spaced seed, appears to be the best estimator.

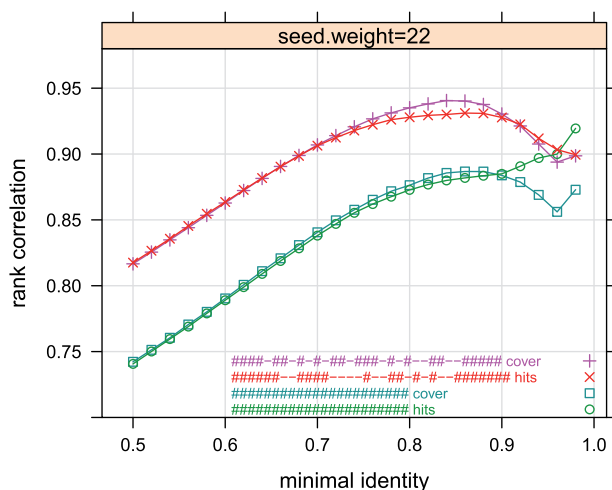


Fig. 1. Spearman's rank correlation between score and counts, depending on the minimal identity rate

### 3.2 Correlation of counts and alignment score

In metagenomic projects, reads to be classified may come from species whose genomes are not stored in the database. These species can be genetic variants of species of the database, such as different strains of the same bacteria, but can also come from organisms represented in the database only at the rank of genus or family or may have no representatives at low taxonomic ranks at all. Therefore, an accurate mapping of a read to a corresponding clade requires estimating its distances to each of the genomes to locate its position within the whole taxonomic tree.

With this motivation, we studied how well the measured counts correlate with the alignment score. For a fixed minimal identity rate  $P_{id}$ , we randomly sampled gapless alignments of length 100 with identity rate from interval  $[P_{id}, 1]$  and collected pairs (number of mismatches, count), where, as before, 'count' stands for either the number of hits or the coverage of a given seed. For these data, we computed Spearman's rank correlation.

A typical result, corresponding to seed weight 22, is shown in Figure 1. It implies that when the identity rate of alignments takes a large range of values (minimal id rate smaller than  $\approx 0.9$ ), spaced seeds yield a significantly higher correlation than contiguous seeds, for both hit-number and coverage counts. Furthermore, the coverage count slightly outperforms the hit-number count, especially for spaced seeds and larger weights.

For high-similarity alignments, however, the picture changes: the coverage count loses its performance, with its correlation value sharply decreasing. Furthermore, the correlation of hit number goes down for spaced seeds as well, while it continues to grow for contiguous seeds ending up by reaching and even slightly outperforming the one for spaced seed. This is due to a larger span of spaced seeds and to their combinatorial properties that cause the hit number values to be less sharply concentrated at certain values and therefore to be less well correlated with the number of mismatches.

In conclusion, while spaced seeds provide a much better estimator for alignments whose score ranges over a large interval, for high-scoring alignments ( $> 90\%$  of identity), the hit number of contiguous seed becomes a better estimator. The superiority of hit number over coverage for high-scoring alignments has also been reported in Noé and Martin (2014). Along with Spearman's correlation, we also made an analysis of mutual information computed on the same data (data not shown) that confirmed the above conclusions.

### 3.3 Correlation on real genomes

To validate the conclusions of the previous section in a real-life metagenomics framework and to analyze more closely how well different counts for a read correlate with the best alignment of the read to a real genomic sequence, we implemented another series of experiments.

Given a genome  $G$ , we generated a set of ILLUMINA-like single-end reads by selecting random substrings of  $G$  of length  $L$  ( $L = 100$ ) and introducing  $k$  mismatch errors, with  $k$  drawn randomly between 1 and 20 for every read. For each read, we computed the counts—hit number and coverage—with respect to genome  $G$  under a given seed, similar to Section 3.1.2. Collected data have then been analyzed.

Figure 2 shows a typical result of this experiment (for seed weight 20) applied to *M.tuberculosis* genome. Each plot shows the density of reads for each pair (number of mismatches, count), depending on the seed (contiguous or spaced) and count (hit number or coverage). Spearman's and Pearson's correlation coefficients are shown for each plotted dataset.

The plots clearly illustrate the advantage of spaced seeds over contiguous seeds for estimating the alignment quality. Plots for contiguous seeds are more blurred, whereas plots for spaced seeds demonstrate a better correlation between the two values. This is confirmed by the absolute values of Spearman's rank correlation coefficient that are significantly higher for spaced seeds, indicating a better statistical dependence.

Comparing hit-number and coverage estimators, we observe that coverage yields a slightly better Spearman correlation and a significantly better Pearson correlation, due to a convex shape of the dependence, compared with the more straight dependence for the coverage.

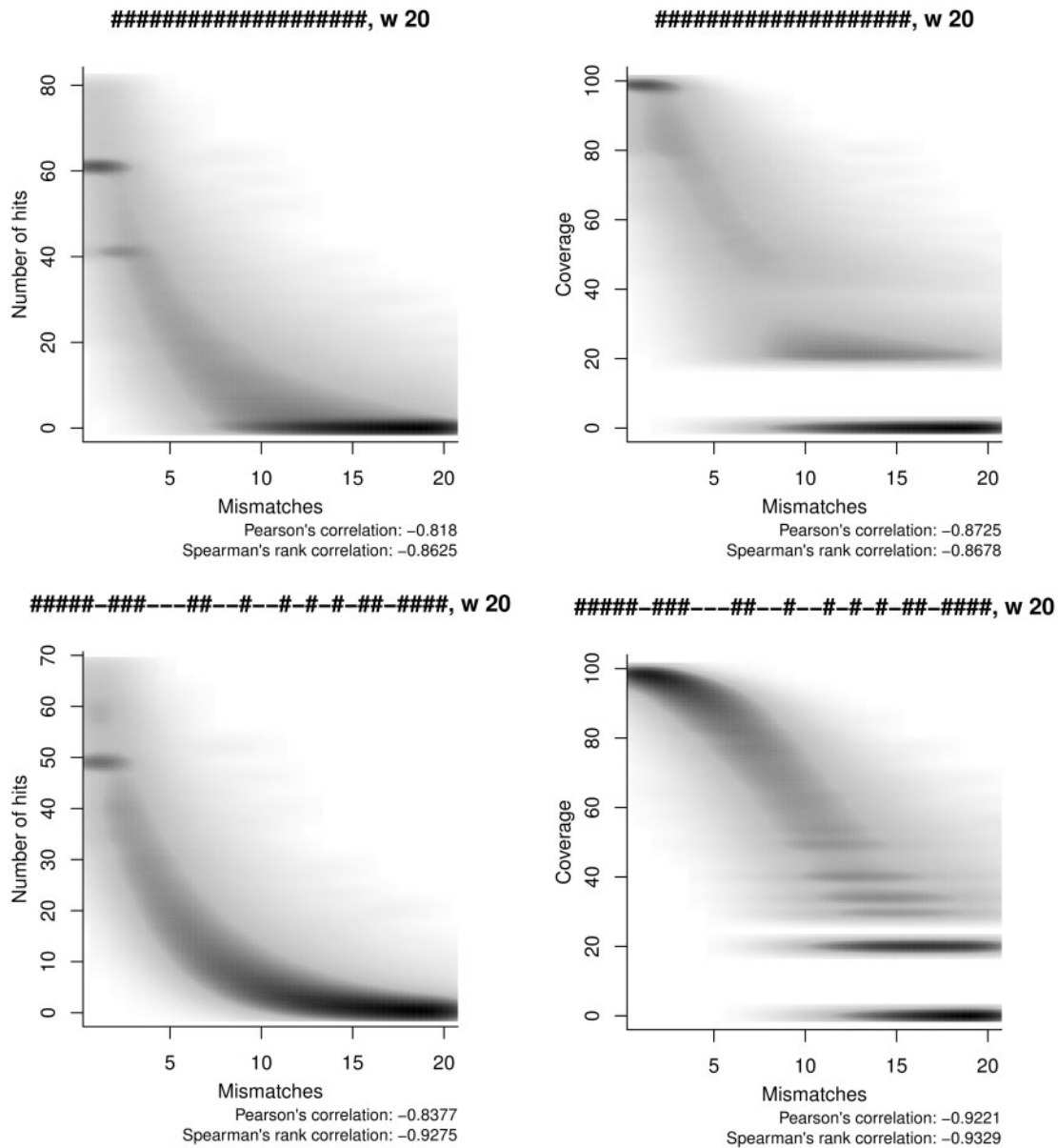
This analysis has been done for several other bacterial genomes, producing similar results. Plots for other genomes and other seed weights can be found in supplementary material.

### 3.4 Large-scale experiments

To validate the advantage of spaced seeds in large-scale metagenomic projects, we modified KRAKEN software (Wood and Salzberg, 2014) to make it work with spaced seeds rather than with contiguous seeds only. The limitation of this approach is that it allows estimating the effect of using spaced seeds only in combination with the KRAKEN algorithm and within its implementation. In return, this allows us to measure the effect of spaced seeds in an unbiased manner, by keeping unchanged all other factors that might influence the results.

Our extended implementation, that we call SEED-KRAKEN, allows the user to specify a spaced seed as a parameter. For a set of genomes, a database of spaced  $k$ -mers matching the seed is constructed, which is later used to classify reads with the original KRAKEN algorithm. Since KRAKEN uses  $k$ -mer counting JELLYFISH program (Marçais and Kingsford, 2011) as the  $k$ -mer indexing engine, we had also to modify JELLYFISH to allow it to deal with spaced  $k$ -mers.

Integration of spaced seed into KRAKEN required a minor modification of the way KRAKEN deals with complementary sequences. In KRAKEN, complementary  $k$ -mers have a single representative in the index, the lexicographically smallest of the two. With spaced seeds, dealing with complementary sequencing is more delicate, as the complement of a spaced  $k$ -mer does not match the same seed but its inverse. To cope with this, we modified KRAKEN to index each distinct  $k$ -mer. We then processed each read in direct and complementary directions separately and select the one which produced more hits. Compared with original KRAKEN, this procedure takes more index space (additional  $\sim 5\%$  in practice) and doubles  $k$ -mer query time.



**Fig. 2.** Hit number (left plots) and coverage (right plots) depending on the number of mismatches in randomly generated reads. Seed is shown above the plot, and Spearman's and Pearson's correlations are shown below. Grayscale shows the density of reads. Experiments made on *M.tuberculosis* genome

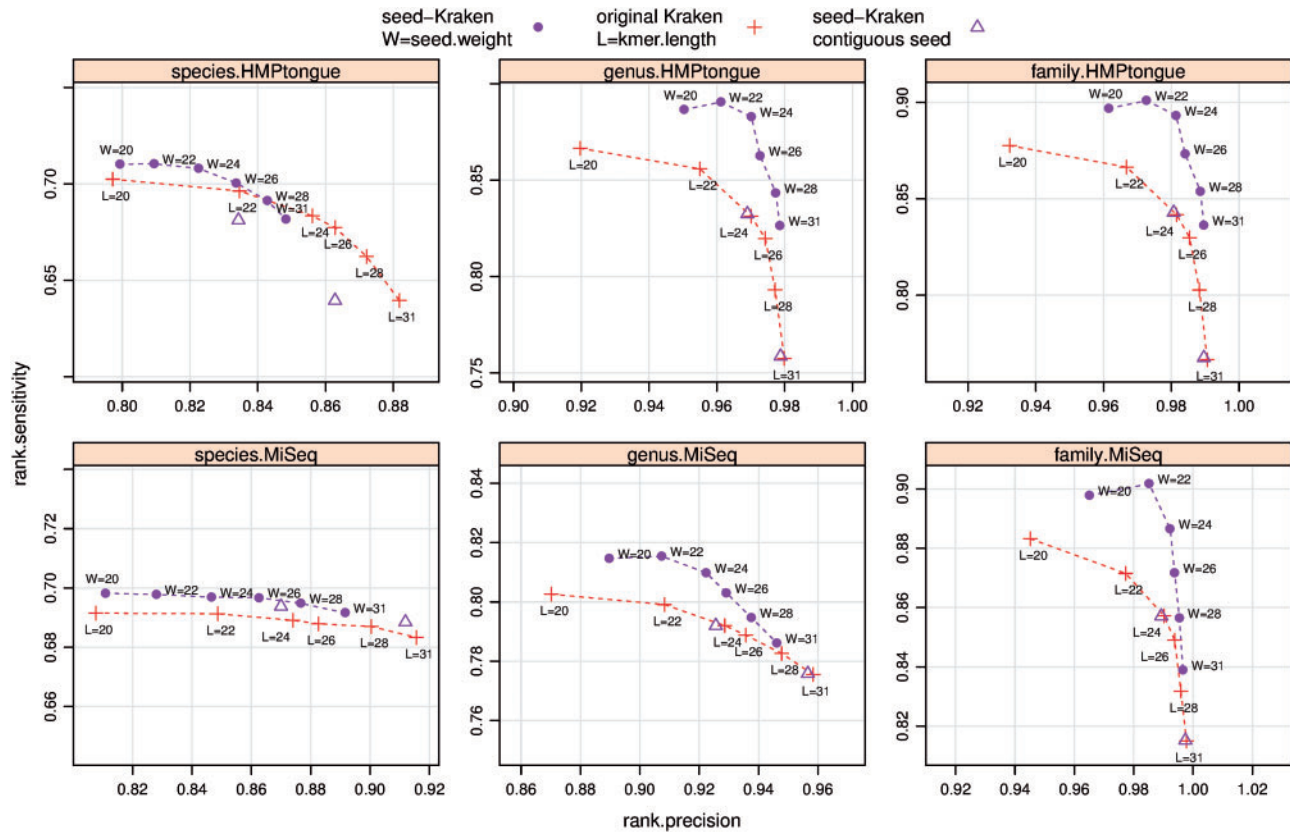
We compared the performance of KRAKEN and SEED-KRAKEN on several datasets. First, we performed experiments with three simulated metagenomes HiSeq, MiSeq and simBA-5 introduced in the original work (Wood and Salzberg, 2014), each containing 10 000 sequences. Furthermore, we created a dataset from Human Microbiome Project data by randomly selecting 50 000 sequences from SRS011086 Tongue dorsum metagenomic sample (<http://hmpdacc.org/HMSCP/>). Here we only report on results for MiSeq and HMPtongue datasets and refer to the supplementary material for a complete account including results for HiSeq and simBA-5. MiSeq is a merge of ILLUMINA reads of 10 bacterial genomes, and HMPtongue is a random sample of real ILLUMINA whole-metagenome sequences.

Because of resource limitations, the database we used in MiSeq experiments was half of the size of the KRAKEN's default database (which requires 70GB of RAM). Our database was obtained by choosing a single representative strain of each bacteria species, except for the species from HiSeq and MiSeq metagenomes for which

all strains were included. Overall, this represented 915 genomes of total size 3.3 GB. For HMPtongue dataset, this database was extended with a subset of HMP reference library, 0.8 GB in total, including references for the selected 50 000 sequences.

For each metagenomic dataset, we measured the sensitivity (percentage of correctly classified reads out of all reads) and precision (percentage of correct classifications out of all classifications) of KRAKEN and SEED-KRAKEN at three taxonomic levels: species, genus and family. In each case, this has been done with seeds of different weights between 20 and 31, and for each weight, SEED-KRAKEN has been run on a few different spaced seeds (see Section 4).

Figure 3 shows sensitivity-precision receiver operating characteristic curves for SEED-KRAKEN and for the unmodified KRAKEN. In the case of SEED-KRAKEN, the 'best performing' seed is charted for each weight. Furthermore, triangle points correspond to SEED-KRAKEN run on contiguous seeds, plotted to measure the effect of our modification in dealing with complementary sequences.



**Fig. 3.** Sensitivity/precision of SEED-KRACKEN (circle points) and original KRACKEN (cross points) for HMPtongue and MiSeq datasets and three taxonomic levels: species, genus and family. Triangle points correspond to SEED-KRACKEN run on contiguous seed of weight 24 and 31, plotted to highlight the effect of the change in the assignment algorithm

At the levels of genus and family, spaced seeds consistently show a better sensitivity-precision trade-off, with the sensitivity increase of about 2 percentage points for MiSeq and 3–5 points for HMPtongue, for a given precision rate. The results of SEED-KRACKEN with contiguous seeds (triangle points) confirm that this improvement is due to the use of spaced seeds and not to our slight modification of the assignment algorithm due to complementary sequences. For small weights (20–22), a spaced seed achieves simultaneously a better sensitivity and a better precision than the contiguous seed of the same weight. When the weight grows, the increment in precision disappears reaching the level of the contiguous counterpart or sometimes coming down below it. However, this is largely compensated by the increase in sensitivity.

For the species level, the picture turns out to be more involved. Here, we observe that due to the small modification of the assignment algorithm, SEED-KRACKEN run with contiguous seeds (triangle points) shows a modified behavior compared with KRACKEN. Specifically, we observe a drop in precision and a gain in sensitivity and those are different for MiSeq and HMPtongue datasets. The reason for this is that SEED-KRACKEN makes more species-level classifications than KRACKEN but at the same time makes more inaccurate assignments to a closely related organism (typically, different strain of the same bacteria), which eventually leads to a lower precision. This phenomenon has a bigger impact for rich databases (HMPtongue experiment) compared with ‘sparse’ databases where each species is represented by few organisms (MiSeq). As for the contribution of spaced seeds, we observe an improved sensitivity-precision trade-off here as well. Compared with SEED-KRACKEN applied to contiguous seeds, this improvement is small for MiSeq

but significant for HMPtongue, which shows a correction capacity of spaced seeds with respect to erroneous assignments to close strains. Compared with the original KRACKEN, we obtain a sensitivity increment of about 1% which becomes smaller (MiSeq) or completely disappears (HMPtongue) when the seed weight grows.

As mentioned earlier, the spaced seed corresponding to each plotted SEED-KRACKEN point has been selected out of a few (usually two to four) seeds tried. The full list of seeds applied in experiments and corresponding results can be found in the supplementary material. Here, we just mention that for large weights (24 and more), the span of the seed becomes an important factor, with seeds of large span showing a drop in sensitivity and best seeds being those with relatively few jokers.

Building a SEED-KRACKEN limited database takes approximately 1 hour on a server with 20 CPU cores, and the resulting size is 26 GB for seed of weight 24, which compares to the 25GB for original KRACKEN. The classification running times are longer than for original KRACKEN by a factor of 3–5.

## 4 Discussion

Through a series of computational experiments, we showed that spaced seeds significantly improve the accuracy of metagenomic classification of short NGS reads. The superiority of spaced seeds for different variants of alignment-free sequence comparison has been recently demonstrated by other authors as well (Ghandi *et al.*, 2014; Leimeister *et al.*, 2014; Noé and Martin, 2014; Onodera and Shibuya, 2013). In this work, we specifically focused on the metagenomics setting characterized by very large volumes of data, both in terms of the number of reads and the size of genomic database.

This quantity of data precludes using some alignment-free comparison techniques and leaves room only for highly time- and space-efficient approaches. Note also that in our setting, we have to compare short sequences (reads) with long ones (whole genomes), which makes an important difference with problems considered in Leimeister *et al.* (2014), Morgenstern *et al.* (2015) and Noé and Martin (2014).

Another improvement considered in Leimeister *et al.* (2014), Morgenstern *et al.* (2015) and Noé and Martin (2014) is to use *multiple seeds*, i.e. several seeds simultaneously instead of a single one. This extension is known to bring an advantage in seed-and-extend sequence alignment (Li *et al.*, 2004; Sun and Buhler, 2004) and (Leimeister *et al.*, 2014; Noé and Martin, 2014) show that this improvement applies to alignment-free comparison as well. However, each seed requires building a separate index for database genomic sequences, and therefore it appears computationally difficult to use multiple seeds in metagenomics, unless some new indexing techniques are designed for this purpose.

In our work, we studied three estimators: hit number, coverage and Jaccard index. Hit number and coverage behave similarly in classification (Section 3.1), but Jaccard index generally improves on them in the case of mapping to real genomes (Section 3.1.2), due to the correction with respect to the *k*-mer-richness. Considered as an estimator of alignment quality (Sections 3.2 and 3.3), coverage provides a certain advantage over hit number. More subtle estimators can be considered as well, e.g. by taking into account the position of *k*-mer in the read (reflecting the sequencing error rate), and this provides an issue for further study.

Designing efficient seeds for metagenomic classification is another important issue that goes beyond the present study. Note that optimal spaced seeds for seed-and-extend alignment are generally not optimal for alignment-free *k*-mer-based comparison (Noé and Martin, 2014). In Noé and Martin (2014), the authors designed (sub-)optimal seeds maximizing the Pearson correlation between hit-number/coverage count and the alignment quality. Their solution is implemented in IEDERA software (The latest version of IEDERA performs design for Spearman correlation as well.) (<http://bioinfo.lifl.fr/yass/iedera>) (Kucherov *et al.*, 2006). On the other hand, recent work (Egidi and Manzini, 2015) introduces *quadratic residue seeds* (QR-seeds) for seed-and-extend alignment, which present a good performance and have the advantage of easy design, avoiding the computationally expensive enumeration of IEDERA. In our work, we used both IEDERA and QR-seeds adapted to our setting. We observed that in most cases, IEDERA seeds are superior (being designed specifically for our task) but in a few cases, QR-seeds demonstrated equal or even better performance (e.g. best results of Table 4 for weights 14–18 were obtained with QR-seeds). This may be due to their large span (cf. supplement material) for which applying IEDERA is computationally costly.

Real data experiments of Sections 3.1.2, 3.3 have been done using SNAKEMAKE (Köster and Rahmann, 2012).

Overall, all our experiments corroborate the thesis of better performance of spaced seeds for metagenomic classification. Experiments with KRAKEN (Section 3.4) give a strong evidence that spaced seeds can improve the classification accuracy in real-life large-scale metagenomic projects. One further improvement would be to implement coverage and Jaccard measures that showed, in general, a better performance compared with the hit number. Introducing spaced seeds rises new issues, such as the construction of an efficient index of the database or adapting the algorithm of computing the most likely node of the taxonomic tree from counts produced by individual genomes, i.e. leaves of the tree. These questions are a subject for future work.

## Acknowledgements

Many thanks to Michal Ziv-Ukelson for discussions that sparked our interest to this research and to Laurent Noé for help with seed design. We are very grateful to anonymous reviewers for their valuable remarks.

## Funding

This work was supported by ABS4NGS grant and by Labex Bézout of the French government (program *Investissement d'Avenir*).

*Conflict of Interest:* none declared.

## References

- Alcaraz, L.D. *et al.* (2010) Understanding the evolutionary relationships and major traits of *Bacillus* through comparative genomics. *BMC Genomics*, **11**, 332.
- Altschul, S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Ames, S.K. *et al.* (2013) Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, **29**, 2253–2260.
- Auch, A.F. *et al.* (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.*, **2**, 117–134.
- Bao, E. *et al.* (2011) SEED: efficient clustering of next-generation sequences. *Bioinformatics*, **27**, 2502–2509.
- Ben-Hur, A. *et al.* (2008) Support vector machines and kernels for computational biology. *PLoS Comput. Biol.*, **4**, e1000173.
- Benson, G. and Mak, D. (2008) Exact distribution of a spaced seed statistic for DNA homology detection. In: Amir, A. (ed.) *String Processing and Information Retrieval, 15th International Symposium, SPIRE 2008, Melbourne, Australia, November 10–12, 2008 Proceedings, volume 5280 of Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pp. 282–293.
- Brady, A. and Salzberg, S. (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat. Methods*, **8**, 367.
- Brown, D.G. (2008) A survey of seeding for sequence alignment. In: Mändouli, I. and Zelikovsky, A. (eds.) *Bioinformatics Algorithms: Techniques and Applications*. John Wiley & Sons, Hoboken NJ, pp. 126–152.
- Burkhardt, S. and Kärkkäinen, J. (2003) Better filtering with gapped *q*-grams. *Fundamenta Informaticae*, **56**, 51–70. Preliminary version in *Combinatorial Pattern Matching 2001*.
- Egidi, L. and Manzini, G. (2015) Multiple seeds sensitivity using a single seed with threshold. *J. Bioinform. Comput. Biol.*, **3**, 1550011.
- Ghandi, M. *et al.* (2014) Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLoS Comput. Biol.*, **10**, e1003711.
- Helgason, E. *et al.* (2000) *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*—one species on the basis of genetic evidence. *Appl. Environ. Microbiol.*, **66**, 2627–2630.
- Huson, D.H. *et al.* (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, **21**, 1552–1560.
- Kawulok, J. and Deorowicz, S. (2015) CoMeta: classification of metagenomes using *k*-mers. *PLoS One*, **10**, e0121453.
- Köster, J. and Rahmann, S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
- Kucherov, G. *et al.* (2006) A unifying framework for seed sensitivity and its application to subset seeds. *J. Bioinform. Comput. Biol.*, **4**, 553–569.
- Langmead, B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
- Leimeister, C.-A. *et al.* (2014) Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, **30**, 1991–1999.
- Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
- Li, M. *et al.* (2004) PatternHunter II: highly sensitive and fast homology search. *J. Bioinform. Comput. Biol.*, **2**, 417–439.
- Lindgreen, S. *et al.* (2015) An evaluation of the accuracy and speed of metagenome analysis tools. bioRxiv. doi:10.1101/017830, Cold Spring Harbor Laboratory.



- Ma, B. et al. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.
- Mande, S.S. et al. (2012) Classification of metagenomic sequences: methods and challenges. *Brief. Bioinform.*, **13**, 669–681.
- Marçais, G. and Kingsford, C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.
- Marco-Sola, S. et al. (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, **9**, 1185–1188.
- Methe, B.A. et al. (2012) A framework for human microbiome research. *Nature*, **486**, 215–221.
- Morgenstern, B. et al. (2015) Estimating evolutionary distances between genomic sequences from spaced-word matches. *Algorithms Mol. Biol.*, **10**, 5.
- Nath Meth Editorial (2009) Metagenomics versus Moore's law. *Nat. Methods*, **6**, 623–623.
- Noé, L. and Kucherov, G. (2004) Improved hit criteria for DNA local alignment. *BMC Bioinformatics*, **5**, 149.
- Noé, L. and Martin, D.E. (2014) A coverage criterion for spaced seeds and its applications to support vector machine string kernels and k-mer distances. *J. Comput. Biol.*, **21**, 947–963.
- Onodera, T. and Shibuya, T. (2013) The gapped spectrum kernel for support vector machines. In: Perner, P. (ed.) *Machine Learning and Data Mining in Pattern Recognition*, volume 7988 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 1–15.
- Ounit, R. et al. (2015) CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, **16**, 236.
- Sun, Y. and Buhler, J. (2004) Designing multiple simultaneous seeds for DNA similarity search. In: Burne, P.E. and Gusfield, D. (eds) *Proceedings of the 8th Annual International Conference on Computational Molecular Biology (RECOMB04), San Diego (California)*. ACM Press, pp. 76–84.
- Teeling, H. and Glockner, F.O. (2012) Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective. *Brief. Bioinform.*, **13**, 728–742.
- Tortoli, E. (2012) Phylogeny of the genus *Mycobacterium*: many doubts, few certainties. *Infect. Genet. Evol.*, **12**, 827–831.
- Vinga, S. (2014) Editorial: alignment-free methods in computational biology. *Brief. Bioinform.*, **15**, 341–342.
- Wood, D.E. and Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.