PaSiT: a novel approach based on short-oligonucleotide frequencies for efficient bacterial identification and typing

Gleb Goussarov, Ilse Cleenwerck, Mohamed Mysara, Natalie Leys, Pieter Monsieurs, Guillaume Tahon, Aurélien Carlier, Peter Vandamme, Rob van Houdt

Bioinformatics, 36(8), 2020, 2337–2344
doi: 10.1093/bioinformatics/btz964
Advance Access Publication Date: 3 January 2020
Original Paper

Genome analysis

PaSiT: a novel approach based on short-oligonucleotide frequencies for efficient bacterial identification and typing

Gleb Goussarov[1,2,*], Ilse Cleenwerck[2], Mohamed Mysara[1], Natalie Leys[1], Pieter Monsieurs[1,†], Guillaume Tahon[2,‡], Aurélien Carlier[2,3], Peter Vandamme[2] and Rob Van Houdt[1,*]

[1]Microbiology Unit, Belgian Nuclear Research Centre (SCK•CEN), Mol, Belgium, [2]Laboratory of Microbiology and BCCM/LMG Bacteria Collection, Department of Biochemistry and Microbiology, Faculty of Sciences, Ghent University, Ghent, Belgium and [3]LIPM, Université de Toulouse, INRAE, CNRS, Castanet-Tolosan, France

*To whom correspondence should be addressed.

†Present address: Veterinary Protozoology, Institute of Tropical Medicine, Antwerp, Belgium

‡Present address: Department of Microbiology, Wageningen University and Research, Wageningen, The Netherlands

Associate Editor: Alfonso Valencia

Received on August 19, 2019; revised on November 21, 2019; editorial decision on December 25, 2019; accepted on December 30, 2019

Abstract

Motivation: One of the most widespread methods used in taxonomy studies to distinguish between strains or taxa is the calculation of average nucleotide identity. It requires a computationally expensive alignment step and is therefore not suitable for large-scale comparisons. Short oligonucleotide-based methods do offer a faster alternative but at the expense of accuracy. Here, we aim to address this shortcoming by providing a software that implements a novel method based on short-oligonucleotide frequencies to compute inter-genomic distances.

Results: Our tetranucleotide and hexanucleotide implementations, which were optimized based on a taxonomically well-defined set of over 200 newly sequenced bacterial genomes, are as accurate as the short oligonucleotide-based method TETRA and average nucleotide identity, for identifying bacterial species and strains, respectively. Moreover, the lightweight nature of this method makes it applicable for large-scale analyses.

Availability and implementation: The method introduced here was implemented, together with other existing methods, in a dependency-free software written in C, GenDisCal, available as source code from https://github.com/LM-UGent/GenDisCal. The software supports multithreading and has been tested on Windows and Linux (CentOS). In addition, a Java-based graphical user interface that acts as a wrapper for the software is also available.

Contact: rvhoudto@sckcen.be or Gleb.Goussarov@sckcen.be

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The concept of bacterial species has long been defined through phenotypic characteristics (Rosselló-Mora and Amann, 2001), resulting in disagreements about what should constitute a bacterial species (Ward, 1998). Currently, a polyphasic approach is used for classification by integrating monophyly (16S rRNA gene similarity), genomic coherence (DNA–DNA hybridization) and phenotypic coherence (Mysara *et al.*, 2017; Rosselló-Móra and Amann, 2015). During the last decade, the introduction of high-throughput sequencing technologies has drastically reduced DNA sequencing costs (www.genome.gov/sequencingcostsdata), which allowed quick and affordable sequencing of whole bacterial genomes, and resulted in an explosive growth of bacterial genomic sequence data. The latter stimulated the development of computational methods to evaluate genomic coherence, with average nucleotide identity (ANI) being implemented most widely (Richter and Rosselló-Móra, 2009).

ANI splits the queried genome into fragments and attempts to align these to a reference genome. For each alignment, the percentage of identical nucleotides is computed and the average of these values is reported. Different methods for defining and aligning these fragments are the basis of multiple ANI variants. Whereas in the original definition of ANI, the DNA fragments were defined as genes (Konstantinidis and Tiedje, 2005), other methods, such as ANIb, ANIm (Richter and Rosselló-Móra, 2009) and OrthoANI (Lee et al., 2016) rely on arbitrary fragments with length of roughly 1000 base pairs. While ANI performs well when comparing a small number of genomes, it cannot be used for high-throughput analyses since it requires aligning entire genomes to each other.

In order to address this issue, a variety of alignment-free methods were developed. These rely on counting oligonucleotide sequences (further referred to as oligonucleotides), which in this context are also referred to as k-mers, ℓ-tuples and n-grams (Dubinkina *et al.*, 2016). Within these, it is possible to introduce a further separation between short oligonucleotide-based methods and long oligonucleotide-based methods. The former operate on the assumption that it is possible to differentiate genomes based on the frequency of each possible oligonucleotide, while the latter attempt to find which oligonucleotides are shared by both genomes. For bacteria, whose genome lengths are on the order of millions of base pairs, the transition between short and long oligonucleotides occurs around decanucleotides where there are 4^{10} possible sequences. Indeed, assuming a perfectly random sequence, each possible oligonucleotide would occur once in the whole genome on average, since the length of small bacterial genomes is around 4^{10}.

Long oligonucleotides have effectively been used by tools, such as MASH (Ondov *et al.*, 2016) and FastANI (Jain *et al.*, 2018), which rely on variants of the MinHash technique (Broder, 1997). These tools provide a good approximation of ANI, while also being computationally efficient, as they do not have an alignment step. However, in order to avoid wasting memory, these methods only work with a small subset of all possible oligonucleotides. Short oligonucleotide-based methods further reduce computational requirements by representing each genome as a numeric vector, where each value is associated with the occurrence frequency of a given oligonucleotide. However, this size reduction may incur a significant loss of information (Pride *et al.*, 2003; Yang *et al.*, 2010; Zhou *et al.*, 2008). Nevertheless, the existence of a species-specific short-oligonucleotide bias has been suggested in the early 1960s, when it was proposed that nucleotide and dinucleotide composition could be a useful tool for classifying microorganisms (De Ley and Van Muylem, 1963; Schildkraut *et al.*, 1962). In the 1990s, Karlin *et al.* outlined a formula to calculate the relative abundance of nucleotides with length 2, 3 and 4, which is referred to as an ‘odds ratio’ (Karlin and Cardon, 1994), and introduced the idea of a genomic signature (Karlin and Burge, 1995). In order to find general patterns, Karlin *et al.* (1998) also investigated frequent long oligonucleotides, codon biases and genomic signatures for a wide range of organisms, including prokaryotes. Their results showed that although some shared biases existed in bacteria, others were also specific to different species. Later investigations confirmed that tetranucleotide frequencies contain a phylogenetically meaningful signal (Reva and Tümmler, 2004; Teeling *et al.*, 2004a). Subsequently, tools such as TETRA (Teeling *et al.*, 2004b) were developed to exploit this knowledge and remain valuable for quickly identifying species (Richter and Rosselló-Móra, 2009).

Currently, web interfaces are user-friendly methods commonly used by perform genome identifications and comparisons. Although they provide an intuitive interface, they do not support large-scale comparisons due to the relatively large amount of memory required to run these programs, and settings tend to be limited. If offline implementations are available, their installation is often complex due to their dependencies on various operating systems. In this work, we propose a novel algorithm that solves the above issues without accuracy loss. Our method relies on short-oligonucleotide biases occurring in bacterial species to estimate taxonomic distances between genomes, and thereby infer the taxon to which a query sample belongs. We compared our implementation with TETRA and ANI (represented primarily by MASH and OrthoANI) for identifying bacterial species and types, respectively. In addition, we implemented a stand-alone command-line application, provided a Java-based graphical user interface and assessed computational requirements.

2 System and methods

We designed a new method for computing dissimilarity between bacterial genomes, which relies on specifically chosen genomic signatures and distance computation method. The signature is based on the transformation of an oligonucleotide frequency profile using a modified version of the method used to compute Karlin signatures. The signature defined by Karlin *et al.* (1994) was modified because its computation would require parsing the entire genome for each factor length, which can be inefficient, while our method allowed computing all the necessary frequencies in a single pass. Additionally, by relying on only one oligonucleotide length to compute frequencies, our modified signature removes the influence from all partial oligonucleotides of length $< k$. In general, this is not true for the original representation of the Karlin signatures, e.g.

$$\gamma_{XYN} = \frac{f_{XYN} f_X f_Y f_N}{f_{XY} f_{XN_1N} f_{YN}} \neq 1 \tag{1}$$

where f_{XYN} is the frequency of oligonucleotides with the sequence XYN. Indeed, unless the genome assembly consists exclusively out of circular sequences, $f_{XY} \neq f_{XYN}$. As a result, $\gamma_{XYN} = 1$ only for closed circular genomes, because in this case the number of oligonucleotides of length k is equal to the number of oligonucleotides with a length shorter than k (Supplementary Fig. S1).

The modified signature can be expressed as the product of the frequencies of all partial oligonucleotides of length k with an even number of undefined nucleotides divided by the product of the frequencies of all partial oligonucleotides of length k with an odd number of undefined nucleotides. In this manner, it can be calculated for any k, whereas Karlin signatures were only defined up to length 4.

We utilized the Manhattan distance for computing distances, with the additional constraint that the differences between the matching components of two signatures were not allowed to exceed a value designated as ‘similarity threshold’, which was optimized in a first step (see Section 3 for details). This single-pass signature threshold (PaSiT) distance was designed to be more locally sensitive than the other tested distances. In a next step, PaSiT was tested for identifying bacterial species as well as typing.

2.1 Datasets

The threshold optimization dataset consisted of 287 bacterial genomes, covering the major phylogenetic lineages (Forterre, 2015). This dataset was constructed by the Belgian Coordinated Collection of Microorganisms. Selected bacterial strains were cultured and genomic DNA was extracted using either a modification of the procedure of Pitcher *et al.* (1989), Gevers *et al.* (2001) and Wilson (2001) or using a Maxwell® 16 Tissue DNA Purification Kit, after a prior enzymatic lysis step in case of gram-positive strains (Supplementary Table S1). DNA integrity and purity were evaluated on a 1.0% (w/v) agarose gel and by spectrophotometric measurements at 234, 260 and 280 nm, respectively. DNA concentration was determined with the QuantiFluor® ONE dsDNA System (Promega Corporation, Madison, WI, USA). Library preparation and whole-genome sequencing were performed by the Oxford Genomic Center (University of Oxford, United Kingdom). Paired-end sequence reads (PE150) were generated using the Illumina HiSeq 4000 platform (Illumina Inc., San Diego, CA, USA). Quality check and trimming of the raw sequence reads was performed using the programs FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Trimmomatic (Bolger *et al.*, 2014). The genome assembly of the raw reads was performed using Shovill (https://github.com/tseemann/shovill), which relies on SPAdes (Bankevich *et al.*, 2012) and pilon (Walker *et al.*, 2014), and evaluated with QUAST (Gurevich *et al.*, 2013). Annotation was performed using RAST (Aziz *et al.*, 2008). Quality check of the assemblies was performed by determining the presence and absence of 107 single copy core genes using bcgTree (Ankenbrand and Keller, 2016), guanine-cytosine content distribution and 16S DNA-based identification.

To test the performance of the similarity threshold, a dataset consisting of 285 species with 20 bacterial genomes per species, and a dataset consisting of 16 228 species with one genome per species were analyzed (Table 1). These datasets are referred to as the ‘species-balanced’ and ‘distinct-species’ datasets, respectively. Both datasets were subsets from the National Center for Biotechnology

Table 1. Datasets used in this work

Dataset	Composition	Source
RefSeq	16 228 species, 112 181 genomes	NCBI RefSeq release 86[a]
Species-balanced	285 species, 20 genomes per species	NCBI RefSeq release 86[a]
Distinct species	16 228 species, 1 genomes per species	NCBI RefSeq release 86[a]
Threshold optimization	200 species, 287 genomes	This work
Typing	6 species, 63 genomes	This work

[a]Downloaded on April 25, 2018.

Information (NCBI)'s reference sequence database (RefSeq). Taxonomic relations between these genomes were obtained from the NCBI taxonomy database (accessed July 12, 2018).

To test the performance of our approach for differentiating strains within bacterial species, i.e. 'types' (Li *et al.*, 2009), a typing dataset (subset of the threshold optimization dataset) was built, containing 63 genomes of six species for which traditional molecular typing data were available. *Burkholderia stabilis* and *Burkholderia cepacia* strains were subjected to multi-locus sequence typing (MLST) based on *atpD*, *gltB*, *gyrB*, *lepA*, *phaC*, *recA* and *trpB* (Baldwin *et al.*, 2005; Spilker *et al.*, 2009). *Pandoraea apista* and *Pandoraea pnomenusa* were typed with *recA*. *Lactobacillus reuteri* and *Lactobacillus rhamnosus* were typed using amplified fragment length polymorphism (Vancanneyt *et al.*, 2006).

Individual accession numbers for the sequencing data are provided in Supplementary Table S1. The full lists of accession numbers for RefSeq subsets are provided in Supplementary Datasets S1 and S2.

2.2 Comparison with existing methods for purposes of identification

To compare our results with existing methods, we computed TETRA (own implementation), and MASH (v2.0) dissimilarities. For TETRA, the existing implementation in JSpeciesWS is not designed to handle large numbers of genomes efficiently, and as such had to be rewritten. Testing on a dataset containing 23 genomes obtained from RefSeq showed that our implementation did not differ substantially from the JSpeciesWS implementation, with differences only occurring in highly fragmented genomes (Supplementary Table S2). The relative accuracy of the methods was illustrated using the standard receiver operating characteristic (ROC) curves associated with the results and by computing Matthew's correlation coefficient (MCC) as a function of the choice of taxonomic cutoff. This was done to obtain a more general view of the accuracy of different methods rather than relying on pre-established cutoffs. For the ROC curves, the true positives (TP) correspond to comparisons where the distance is smaller than the selected species-level cutoff while also being labeled as belonging to the same species, while the false positives (FP) correspond to comparisons where the distance is smaller than the selected taxonomic cutoff but the samples are labeled as belonging to different species.

MCC is computed using the following equation:

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \tag{2}$$

where, similar to TP and FP, TN and FN represent true negatives and false negatives.

2.3 Comparison with existing methods for purposes of typing

Bacterial typing is performed on organism-by-organism and case-by-case basis, and as such, there does not currently exist a single typing scheme that can be applied in a more general context. We evaluated the applicability of our method as a predictor of typing-based comparisons on a dataset containing 63 genomes, belonging to 6 species (typing dataset). Unlike the dataset used for evaluating the identification capability of our method, this dataset was not designed to draw general conclusions, but still provides a good indication about the performance of the different methods as the selected species enable comparing multiple levels of taxonomic similarity.

Our results were compared with TETRA, MASH and OrthoANI. Moreover, since this dataset only contained 63 genomes, it was computationally feasible to obtain ANI values for paired comparison using OrthoANI. Each method yielded a distance matrix, which was used to perform case-by-case comparisons, illustrating the differences in behavior depending on each method.

Similar to what was done for identification, we used ROC curves and MCC to compare different methods. The TP and FP were defined, similar to how was done for identification, but based on whether genomes belonged to the same or a different type rather than species.

3 Algorithm

We considered that the computation of dissimilarity between genomes depends on two fundamental parts: a choice of signature, representing the genome, and a choice of distance, producing a single numeric value that gives an indication about the similarity of the compared genomes.

3.1 Defining the genomic signature

In this study, we defined a genomic signature as a vector of numbers that is unique to each genome. We refer to the vector containing the number of occurrences of all possible oligonucleotides of a given length k within a genome as its frequency profile, which consists of 4^k elements labeled $f_{X_1 \cdots X_k}$. When analyzing bacterial sequences, the profiles are computed based on double-stranded DNA, so the occurrence of an oligonucleotide contributes to the profile as well as its own reverse complement. Although a frequency profile can be used as a signature, we first performed a transformation derived from the one proposed by Karlin and Cardon (1994) who described methods to compute signatures up to length 4 using the following notation:

$$\rho_{XY} = \frac{f_{XY}}{f_X f_Y} \tag{3}$$

$$\gamma_{XYZ} = \frac{f_{XYZ} f_X f_Y f_Z}{f_{XY} f_{XNZ} f_{YZ}} \tag{4}$$

$$\tau_{XYZW} = \frac{f_{XYZW} f_{XY} f_{XNZ} f_{XN_1N_2W} f_{YZ} f_{YNW} f_{ZW}}{f_{XYZ} f_{XYNW} f_{XNZW} f_{YZW} f_X f_Y f_Z f_W}. \tag{5}$$

In the above formulae, ρ_{XY}, γ_{XYZ} and τ_{XYZW} represent the value of the genomic signature related to oligonucleotide XY, XYZ and $XYZW$, respectively, where X, Y, Z and W each represent one fixed nucleotide. The genomic signature is obtained by calculating this value for all possible combinations of nucleotides for a given length. $f_{X_1 \cdots X_k}$ correspond to the frequency at which a given oligonucleotide $(X_1 \cdots X_k)$ occurs in the genome or a part thereof. In this context, N, represents 'any nucleotide'.

The signature defined by Karlin for length k depends on oligonucleotides of lengths i $< k$. Its computation would thus require parsing the entire genome for each factor length in the equation, which is inefficient. We modified the formula proposed by Karlin so that for each order they depend on oligonucleotides of one length only, which allowed us to compute frequencies only once

$$K_{XY} = \frac{f_{XY}}{f_{X*} f_{*Y}} \tag{6}$$

$$K_{XYZ} = \frac{f_{XYZ} f_{X**} f_{*Y*} f_{**Z}}{f_{XY*} f_{X*Z} f_{*YZ}} \tag{7}$$

$$K_{XYZW} = \frac{f_{XYZW} f_{XY**} f_{X*Z*} f_{X**W} f_{**YZ} f_{*Y*W} f_{**ZW}}{f_{XYZ*} f_{XY*W} f_{X*ZW} f_{*YZW} f_{X***} f_{*Y**} f_{**Z*} f_{***W}}. \tag{8}$$

We used $*$ rather than N to represent ‘any nucleotide’ in order to make the equations more legible. Here, K_{XY} only depends on dinucleotides, K_{XYZ} only depends on trinucleotides and so on. Note that using these new formulae, if any nucleotide is undefined (i.e. X, Y, Z or W is replaced by $*$), the modified signature is always $=1$, whereas this is not guaranteed when using the original formula. For example, for $k = 3$

$$K_{XY*} = \frac{f_{XY*} f_{X**} f_{*Y*} f_{***}}{f_{XY*} f_{X**} f_{*Y*}} = 1. \tag{9}$$

3.2 Computing distances between signatures

To compute distances between signatures, we designed a locally sensitive distance (D_S). With D_S, the difference between two signature components is first passed through a saturation function h_t before being used to compute the Manhattan distance and normalized by the number of components in the signature, as follows:

$$h(x|t) = \begin{cases} 0, & x < 0 \\ x/t, & 0 \le x \le t \\ 1, & x > t \end{cases} \tag{10}$$

$$D_S(A,B) = \frac{1}{4^k} \sum_{i=1}^{4^k} h(A_i - B_i). \tag{11}$$

In the above equations, A and B are genomic signatures whose ith elements are written as A_i and B_i, respectively, k is the oligonucleotide length and t is an optimized similarity threshold (Sections 3.3 and 5.1). In order to avoid confusion, this threshold is referred to as the ‘similarity threshold’ while the values of the different distances associated with transitions between taxonomic similarity levels (e.g. 98.7% for 16S rRNA gene similarity) are referred to as ‘taxonomic cutoffs’.

D_S can be seen as a relaxed form of the Hamming distance. This approach reduces the effect of having a small number of highly dissimilar values by considering all large difference as equal to one, while smaller differences cover the range between zero and one. This is done for each component of the signature. The average of these differences is then reported as the distance between the two genomes. With an appropriate choice of similarity threshold, using the average ensures that the final output is always between 0 and 1. Accordingly, the dissimilarity between two signatures is $=1$ if all values are above the threshold (i.e. significantly different from each other), and 0 if all values are equal. As this method puts a threshold on large variations in individual components of the signature, it is sensitive to small differences between genomes.

Next to the Manhattan distance, the Euclidian distance could also have been used to compute D_S. However, for high dimensional data, the Manhattan distance is generally preferable over the Euclidian distance (Aggarwal *et al.*, 2001). In addition, we found that using the Euclidian distance decreased accuracy slightly (Supplementary Fig. S2).

3.3 Optimization of the similarity threshold

Optimization of the similarity threshold t was performed using the hill climb approach. The parameter to be maximized was defined as the distance between the centroids of distributions pertaining to two different classes of comparisons, e.g. ‘same species’ and ‘same genus’, divided by the sum of the standard deviation of both distributions.

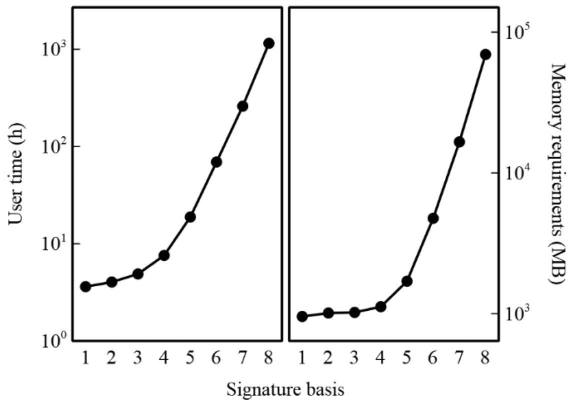

Fig. 1. Computational cost, measured as user time (i.e. excluding system calls) in hours, and required RAM for the PaSiT method as a function oligonucleotide length

$$d(A,B) = \frac{\mu_B - \mu_A}{\sigma_A + \sigma_B} \tag{12}$$

where A and B are distributions, μ_A and μ_B are their means, and σ_A and σ_B are their standard deviations.

We preferred this approach to a more typical classifier evaluation, such as the area under the ROC curve, because the distributions associated with comparing genomes belonging to the same species and genomes belonging to the same genus have almost no overlap in the datasets used to estimate the threshold.

In order to ensure that the selected threshold would not be over fitted, optimization was first performed on a limited dataset. Second, to verify that this threshold performed appropriately, we applied the resulting D_S with the optimized threshold to the ‘species-balanced’ and ‘distinct–species’ datasets. The two datasets ensured a balanced reference for obtaining distributions associated with distances between genomes belonging to the same species, and genomes belonging to different species, respectively.

4 Implementation

We implemented the software, named GenDisCal, as a standalone command-line application, and provided a Java-based graphical user interface wrapper with a simplified use that provides appropriate presets and also includes most of the options available through the command-line. The computational requirements of this software rely almost exclusively on the choice of signature length (Fig. 1). While lengths above eight would remain feasible if a smaller number of genomes were used, these provide neither a computational advantage nor an accuracy advantage over other methods (data not shown) and thus were not considered here.

In addition, we performed the same analysis using MASH. For FastANI and OrthoANI, computational costs for the RefSeq dataset were prohibitively large, and smaller datasets were used instead (Table 2). For TETRA, the publically available implementation in JSpeciesWS is not designed to handle more than 15 genomes at once and hence is not considered here. MASH has roughly the same speed as PaSiT6.

In order to demonstrate GenDisCal’s performance, it was tested against other tools on a Linux server with an Intel Xeon CPU E5-2690 v4, running at 2.6 GHz, with 56 cores and 252 GB of RAM (16 cores were used in all tests).

5 Discussion

5.1 Parameter optimization

In a first step, the similarity threshold was optimized using a curated dataset containing 287 genomes with an unambiguous taxonomic

Table 2. Computational costs of different methods, using default settings

Dataset [samples]	Method	User time (min)	Memory
RefSeq [112 181]	MASH sketch	565	2690 MB
RefSeq [112 181]	MASH dist	4765[a]	2716 MB
RefSeq [112 181]	PaSiT4	455	1120 MB
RefSeq [112 181]	PaSiT6	4156	4762 MB
Species-balanced [5700]	FastANI	10410[a]	97.04 GB
Typing [64]	OrthoANI	4180	5452 MB

[a] All comparisons had to be performed twice due to input format limitations, so time could potentially be halved.

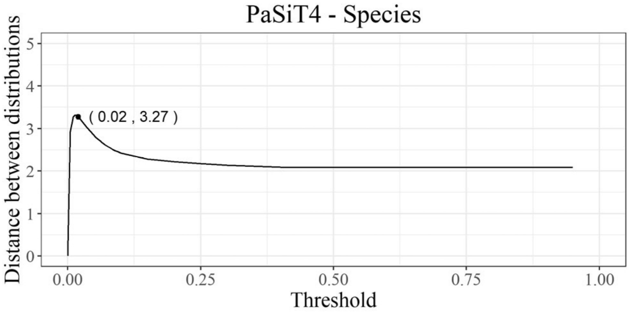

Fig. 2. Distance between 'same species' and 'same genus' distributions for the PaSiT4 method. Distances were measured as a function of the similarity threshold used to calculate D_S. The distance between distributions was defined as $|\mu_A - \mu_B|/(\sigma_A + \sigma_B)$, where μ and σ represent the mean and standard deviation of a distribution, respectively

assignment. For different oligonucleotide lengths as well as similarity threshold values, we computed the distance between the distributions associated with 'same species' comparisons and 'same genus' comparisons (Fig. 2). The optimal values for each oligonucleotide length were refined using the hill climb method (Table 3). We found that for tetranucleotides, a similarity threshold of 0.0144 produced the most distinct distributions of any oligonucleotide length. However, we decided to round this value up to 0.02, since performance is more affected by underestimation than overestimation (Fig. 2). With the threshold set to 0.02, the distance between means of the distributions being 3.27 times the sum of the standard deviations of each distribution. Therefore, there should be little ambiguity when testing if two genome sequences belong to the same species when applying PaSiT with tetranucleotides (i.e. PaSiT4) and a 0.02 similarity threshold. In addition, PaSiT with hexa and heptanucleotides should not be used for species identification since the maximal distance between distributions is comparatively quite low (Table 3).

5.2 Identification of bacterial species

We defined a species-level cutoff value based on the output of the threshold optimization dataset and a similarity threshold of 0.02. In this dataset, all genomes with distances below 0.36 belong to the same species (Figs 3a and 4b), with the exception of comparisons between *Lactobacillus plantarum* and *Lactobacillus paraplantarum* genomes which also fall into this range. The 0.36 species-level cutoff value is also valid when applied to the species-balanced dataset (Fig. 3b). However, the distribution from genomes belonging to the same genus ('same genus' distribution) overlaps with the distribution associated with comparisons of genomes belonging to the same species ('same species' distribution). The number of inter-genera distances below 0.6 is smaller for the distinct-species dataset (Fig. 3c) than for the species-balanced dataset, indicating that the overlap of the same-genus and same-species distributions in the latter is due to the inclusion of 'outlier' genera.

In order to assess whether or not our results improve the state-of-the-art, we compared the relative accuracy of PaSiT to TETRA and MASH using the species-balanced dataset by assessing the ROC curves and MCC associated with the results (Fig. 4). The performance of PaSiT4 is comparable with the other methods when it comes to distinguishing species.

Table 3. Optimal thresholds obtained for different oligonucleotide lengths and the associated distances between the distributions

Oligonucleotide length	Optimal threshold	Distance
2	0.0194	2.74
3	0.0129	2.94
4	0.0144	3.31
5	0.0150	2.86
6	0.0250	2.38
7	0.0687	1.97

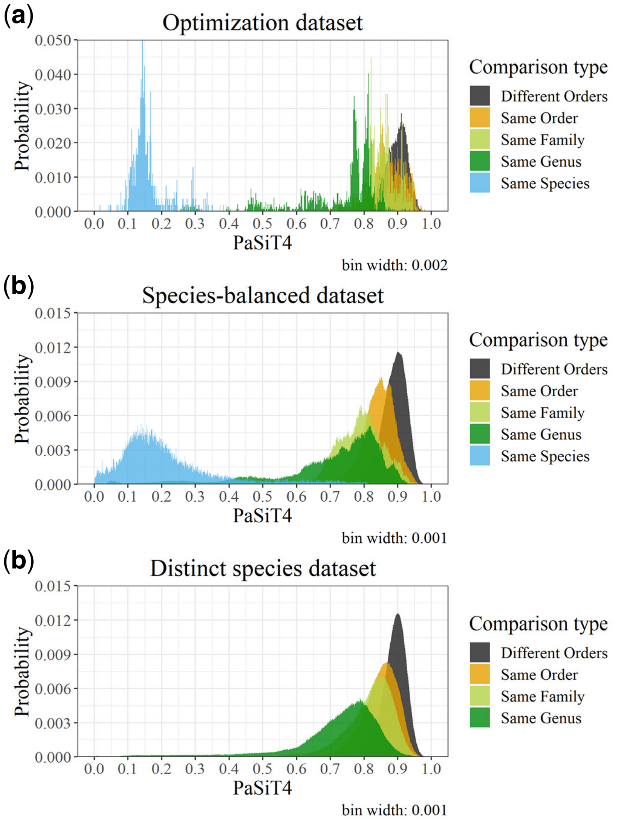

Fig. 3. Histograms showing the distributions of the PaSiT4 dissimilarity values. These graphs were obtained from all-against-all comparisons for the (**a**) optimization dataset, (**b**) species-balanced dataset and (**c**) distinct-species dataset. Histograms are colored according to the taxonomic relation between these genomes

5.3 Bacterial typing

In a next step, we used a dataset with genomes of known bacterial types (typing dataset) to evaluate whether PaSiT could be used for bacterial typing. To illustrate the bacterial typing capabilities of the different methods, heatmaps of complete distance matrices are shown (Fig. 5) as well as histograms (Fig. 6).

In terms of 'same species' comparisons, values for the genomes belonging to *B.stabilis* and *Burkholderia cenocepacia* are clearly above the 'same species' cutoffs using PaSiT4 and OrthoANI-based dissimilarity, but fall into the ambiguous zone when using TETRA (see Supplementary Tables S3–S8 for exact values). On the other hand, values for the *L.reuteri* genomes are below the 'same species'

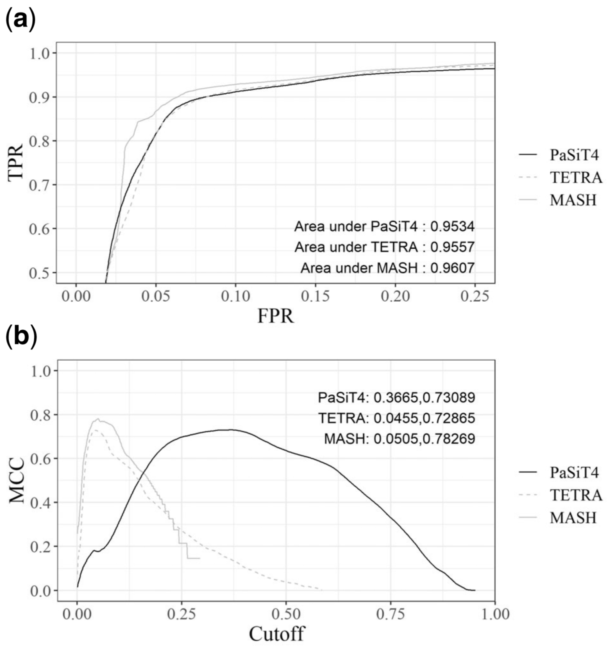

Fig. 4. Comparison of the accuracy of PaSiT4, TETRA and MASH on the species-balanced dataset by means of (**a**) ROC curves created by plotting the TP rate against the FP rate for different taxonomic cutoff values, and (**b**) MCC as a function of the taxonomic cutoff used to differentiate 'same species' comparisons from 'same genus' comparisons. Labels correspond to the area under the ROC curve (a) and the best cutoff and associated MCC value (b), respectively. For TETRA, these graphs are based on transformed values, defined as square root of $(1-T_{\mathrm{orig}})/2$, since these values correlate well with ANI (Richter and Rosselló-Móra, 2009)

cutoffs when using PaSiT4 and TETRA, while OrthoANI-based dissimilarity values are above the cutoff in case of LMG 13088 with LMG 9213^{T} and LMG 18238. It is worth nothing that for all other genome pairs of the *L.reuteri* strains the OrthoANI-based dissimilarity values are below the 'same species' cutoff. These results indicate that PaSiT4 is most congruent with the current classification of the typing dataset.

For typing, PaSiT6 and OrthoANI both provide results that are largely congruent with the data obtained through the classical typing methods MLST and amplified fragment length polymorphism, and to each other. However, PaSiT6 and OrthoANI do provide a different outcome for the comparison of two *Burkholderia* strains of type 122 (R-13114 and R-18887). The OrthoANI-based dissimilarity value for these two strains was calculated as 0.0005, assigning both strains to the same type, while the PaSiT6 distance was 0.33, suggesting that these strains belong to different types. The PaSiT6 value is higher than expected based on the OrthoANI value. A whole-genome alignment of the two assemblies using Mauve (Darling *et al.*, 2004) revealed that one of the genomes contained contigs with a total length over 800 kbp that had no equivalent in the other, mostly related to mobile genetic elements [e.g. (mega)plasmids]. This illustrates a key difference between both methods, as PaSiT6 uses the entirety of the input data, while with OrthoANI and other ANI methods only shared regions are investigated. Numeric representations of the accuracy of the methods can be found in Figure 7.

5.4 Conclusion

ANI was proposed as a new standard for bacterial identification based on genome coherence by Richter and Rosselló-Móra (2009) and remains to date one of the most effective ways to compare entire genomes to each other. However, until recently the computational cost of such methods was prohibitive for large datasets such as the ones presented here. While this problem has largely been solved by

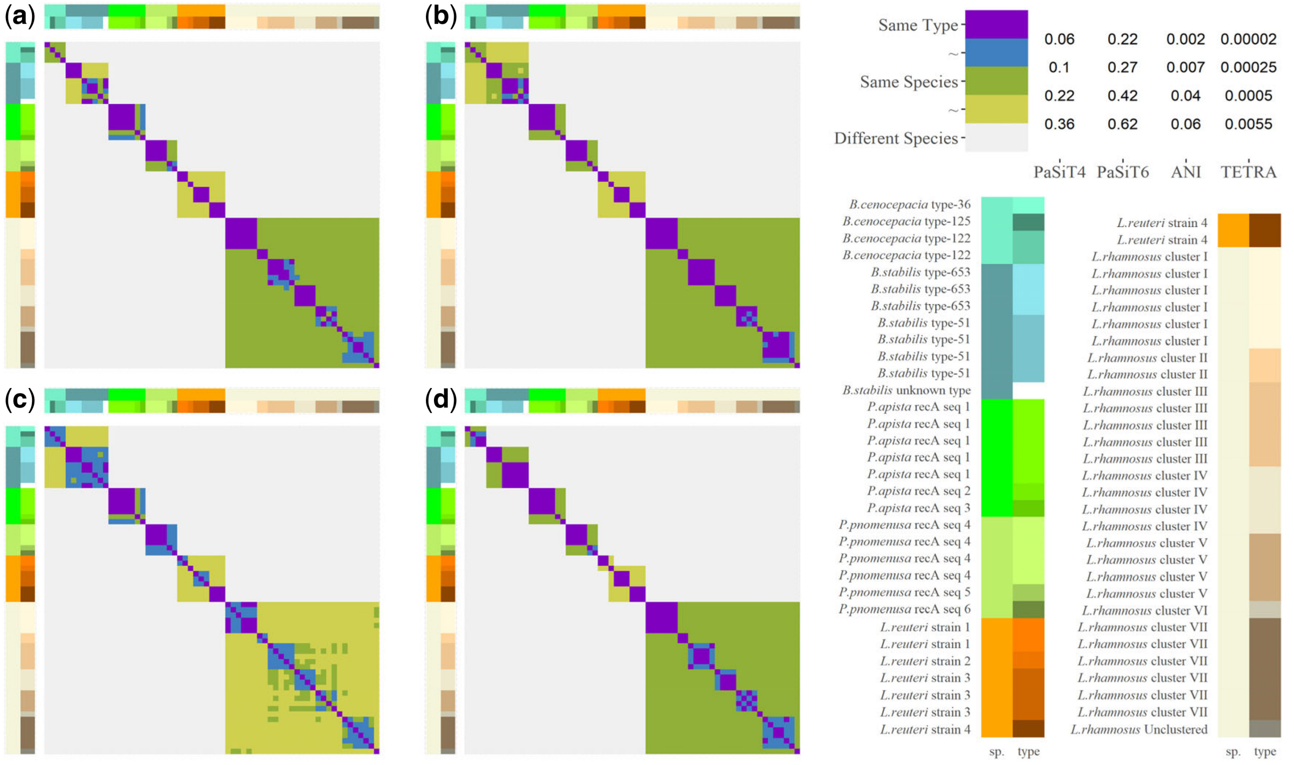

Fig. 5. Distance matrices for the typing dataset calculated with (**a**) PaSiT4, (**b**) PaSiT6 (**c**) TETRA and (**d**) OrthoANI. The color code on the top right represents taxonomic cutoffs: purple—same type, blue—likely same type, green—same species, yellow—likely same species, white—likely different species, as well as the associated values. Only the cutoffs for 'likely same species' are defined in the existing literature for ANI (94∼96% identity) and TETRA (correlation of 0.989∼0.999), other cutoffs were chosen manually to fit the comparison types with as few incorrectly colored cells as possible. Each sample in the distance matrix is represented by two colors, corresponding to the species and type, respectively. The exact values as well as results obtained with MASH and FastANI are available in Supplementary Tables S7 and S8, respectively. (Color version of this figure is available at *Bioinformatics* online.)

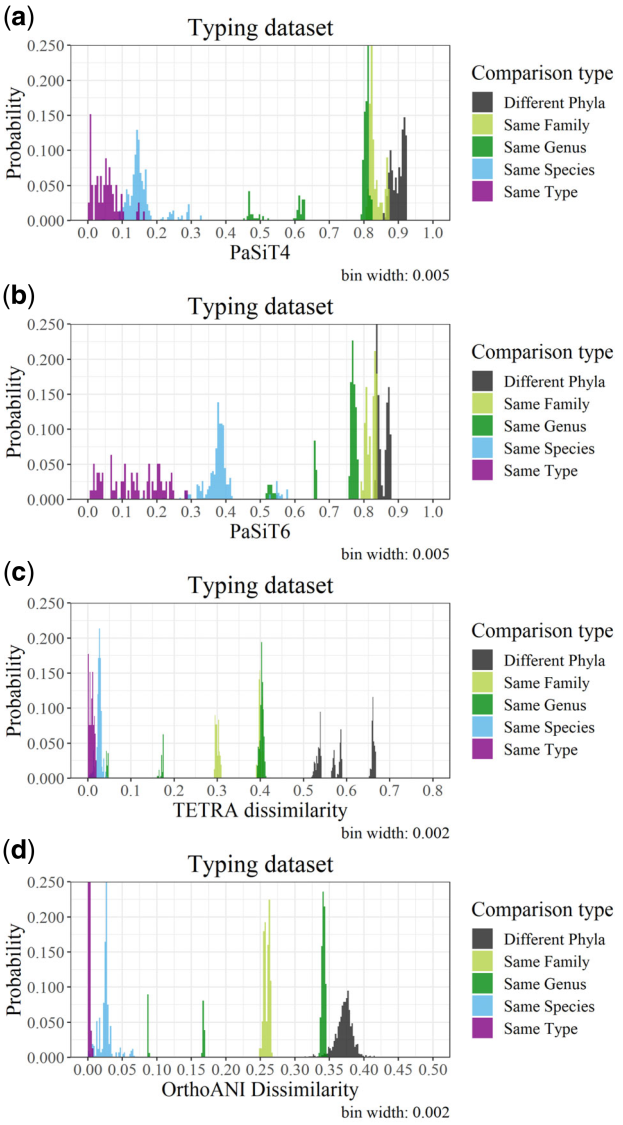

Fig. 6. Histograms showing the distributions observed on the typing dataset for the (**a**) PaSiT4 dissimilarity value, (**b**) PaSiT6 dissimilarity values, (**c**) the square roots of normalized TETRA values and (**d**) ANI dissimilarity values computed using OrthoANI for all-against-all comparisons. Histograms are colored according to the taxonomic relation between these genomes. TETRA values are based on transformed values defined as $\sqrt{(1 - T_{orig})/2}$

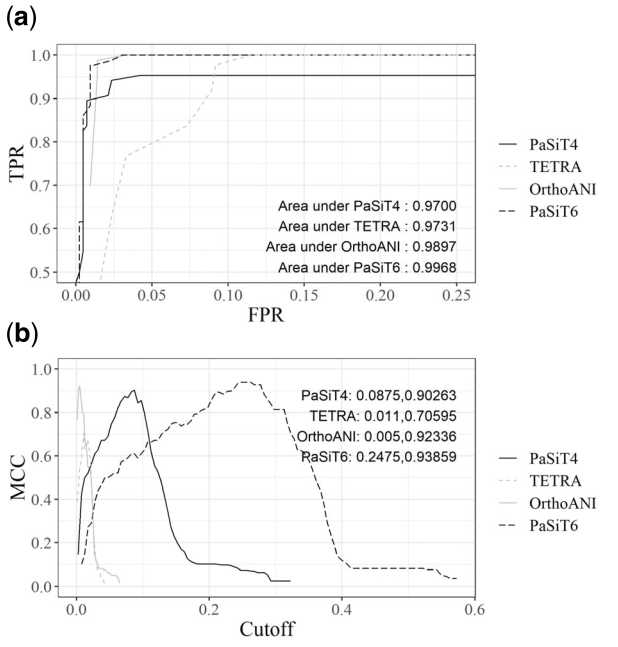

Fig. 7. Comparison of the accuracy of PaSiT4, PaSiT6, TETRA and OrthoANI on the typing dataset by means of (**a**) ROC curves created by plotting the TP rate against the FP for different taxonomic cutoff values and (**b**) MCC as a function of the taxonomic cutoff used to differentiate 'same type' comparisons from 'same species' comparisons. Labels correspond to the area under the ROC curve (a), and the best cutoff and associated MCC value (b), respectively. TETRA values are based on transformed values, defined as $\sqrt{(1 - T_{orig})/2}$

the introduction of ANI approximations using MinHash techniques, it can be further reduced by using tetranucleotide frequencies as a basis for comparison. Moreover, these MinHash-based methods utilize only part of the entire genome to perform comparisons, as opposed to frequency methods.

The PaSiT method is on par with TETRA in terms of accuracy when using tetranucleotides (PaSiT4). A stand-alone version of TETRA, which is able to handle large-scale comparisons, is currently not available. Therefore, a rigorous comparison is not possible. Nevertheless, GenDisCal contains an implementation of TETRA (Teeling *et al.*, 2004a), which was used to compare its accuracy to the proposed method. Interestingly, while both methods have a similar accuracy, they are not strongly correlated, with an R^2 of 0.7 on the 'same species' distribution from the species-balanced dataset (Supplementary Fig. S3). This indicates that the proposed method exploits the available data differently. Indeed, by reducing the influence of large differences between components of signatures on the overall distance, cases where only a small part of the signature differs produce a small distance, even if the parts that differ do so significantly. Based on the distributions obtained using the typing, species-balanced and distinct-species datasets, taxonomic cutoffs can be proposed for tetranucleotide dissimilarities. Values below 0.22 indicate that two samples belong to the same species, while values above 0.36 indicate that two samples belong to different species. The 0.02 threshold, while not optimal for all oligonucleotide lengths up to hexanucleotides, is sufficiently close to the best value and as such can also be used for those lengths without impacting the results significantly.

When using hexanucleotides, the accuracy of the PaSiT method is similar to that of ANI (represented here by OrthoANI or MASH, depending on the size of the dataset) for distinguishing bacterial types while requiring less computational resources than even the modern equivalents. Moreover, although it is less accurate than PaSiT6, PaSiT4 offers an improvement over TETRA in this use-case, as shown by its MCC of 0.9, compared to 0.7 for TETRA (see also the number of values within the ambiguous type zone in Fig. 4).

An important consideration is that the proposed method is sensitive to large insertions or deletions with nucleotide composition diverging from the rest of the genome, such as DNA acquired by horizontal gene transfer (Dufraigne *et al.*, 2005; Garcia-Vallvé *et al.*, 2000), and will therefore provide results that diverge from ANI values when comparing samples belonging to species with genomic islands or plasmids (Smillie *et al.*, 2010). A clear example of such a case was provided here by the *B.cenocepacia* type-122 assemblies. Indeed, these two genomes showed a significant difference using the proposed method, which was not observed using ANI. A similar case was also observed for *B.stabilis* type-51, but was less pronounced. While this difference is likely due to mobile genetic elements, and as such does not necessarily impact the bacterial type when it is obtained using MLST, this difference would not have been apparent when using only ANI values.

We also noted that comparison of the genomes of the different *L.reuteri* strains revealed ANI values below even the less conservative 94% similarity cutoff when comparing *L.reuteri* LMG 13088 with LMG 9213^T and LMG 18238, while values above this cutoff were found for all other pairs of the *L.reuteri* strains of the typing dataset. For both TETRA and our PaSiT4 method, all values are within the range of uncertainty. This shows that PaSiT4 provides an added value to ANI.

To conclude, we developed a new alignment-free approach for computing genomic distances, called PaSiT. This method utilizes the entirety of the genomic information and improves on the speed and—for selected datasets—on the accuracy of existing methods. PaSiT, as well as other oligonucleotide-based methods (a full list is provided in the software help) are implemented in a dependency-free program called GenDisCal and an intuitive companion graphical user interface. Combined these tools can be used to produce a comprehensive types of signatures and calculates distances between them based on a specified algorithm.

Acknowledgements

We thank the Oxford Genomics Centre at the Wellcome Centre for Human Genetics (funded by Wellcome Trust grant reference 203141/Z/16/Z) for the generation and initial processing of the sequencing data. We are indebted to BCCM/LMG Bacteria Collection staff for the cultivation and DNA extraction for whole genome sequence analysis of all strains included in this study.

Funding

This work was jointly supported by the BCCM/LMG Bacteria Collection and Belgian Nuclear Research Centre (SCK•CEN). The BCCM/LMG Bacteria Collection is supported by the Federal Public Planning Service—Science Policy, Belgium.

Conflict of Interest: none declared.

References

Aggarwal,C.C. *et al.* (2001) On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche Jan and Vianu Victor (eds.) *Database Theory—ICDT 2001*. Springer, Berlin Heidelberg, pp. 420–434.

Ankenbrand,M.J. and Keller,A. (2016) bcgTree: automatized phylogenetic tree building from bacterial core genomes. *Genome*, **59**, 783–791.

Aziz,R.K. *et al.* (2008) The RAST server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

Baldwin,A. *et al.* (2005) Multilocus sequence typing scheme that provides both species and strain differentiation for the *Burkholderia cepacia* complex. *J. Clin. Microbiol.*, **43**, 4665–4673.

Bankevich,A. *et al.* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Broder,A.Z. (1997) On the resemblance and containment of documents. In *Compression and Complexity of SEQUENCES Proceedings 1997 (Cat. No.97TB100171)*, pp. 21–29.

Darling,A.C.E. *et al.* (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14**, 1394–1403.

De Ley,J. and Van Muylem,J. (1963) Some applications of deoxyribonucleic acid base composition in bacterial taxonomy. *Antonie Van Leeuwenhoek*, **29**, 344–358.

Dubinkina,V.B. *et al.* (2016) Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics*, **17**, 38.

Dufraigne,C. *et al.* (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. *Nucleic Acids Res.*, **33**, e6.

Forterre,P. (2015) The universal tree of life: an update. *Front. Microbiol.*, **6**, 717.

Garcia-Vallvé,S. *et al.* (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.*, **10**, 1719–1725.

Gevers,D. *et al.* (2001) Applicability of rep-PCR fingerprinting for identification of Lactobacillus species. *FEMS Microbiol. Lett.*, **205**, 31–36.

Gurevich,A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

Jain,C. *et al.* (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 5114.

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, **11**, 283–290.

Karlin,S. and Cardon,L.R. (1994) Computational DNA sequence analysis. *Annu. Rev. Microbiol.*, **48**, 619–654.

Karlin,S. *et al.* (1994) Heterogeneity of genomes: measures and values. *Proc. Natl. Acad. Sci. USA*, **91**, 12837–12841.

Karlin,S. *et al.* (1998) Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, **32**, 185–225.

Konstantinidis,K.T. and Tiedje,J.M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc. Natl. Acad. Sci. USA*, **102**, 2567–2572.

Lee,I. *et al.* (2016) OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.*, **66**, 1100–1103.

Li,W. *et al.* (2009) Bacterial strain typing in the genomic era. *FEMS Microbiol. Rev.*, **33**, 892–916.

Mysara,M. *et al.* (2017) Reconciliation between operational taxonomic units and species boundaries. *FEMS Microbiol. Ecol.*, **93**, fix029.

Ondov,B.D. *et al.* (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.

Pitcher,D.G. *et al.* (1989) Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. **8**, 151–156.

Pride,D.T. *et al.* (2003) Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.*, **13**, 145–158.

Reva,O.N. and Tümmler,B. (2004) Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformatics*, **5**, 90.

Richter,M. and Rosselló-Móra,R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. USA*, **106**, 19126–19131.

Rosselló-Mora,R. and Amann,R. (2001) The species concept for prokaryotes. *FEMS Microbiol. Rev.*, **25**, 39–67.

Rosselló-Móra,R. and Amann,R. (2015) Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.*, **38**, 209–216.

Schildkraut,C.L. *et al.* (1962) Deoxyribonucleic acid base composition and taxonomy of some protozoa. *Nature*, **196**, 795–796.

Smillie,C. *et al.* (2010) Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–452.

Spilker,T. *et al.* (2009) Expanded multilocus sequence typing for burkholderia species. *J. Clin. Microbiol.*, **47**, 2607–2610.

Teeling,H. *et al.* (2004a) Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.*, **6**, 938–947.

Teeling,H. *et al.* (2004b) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163.

Vancanneyt,M. *et al.* (2006) Intraspecific genotypic characterization of *Lactobacillus rhamnosus* strains intended for probiotic use and isolates of human origin. *Appl. Environ. Microbiol.*, **72**, 5376–5383.

Walker,B.J. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.

Ward,D.M. (1998) A natural species concept for prokaryotes. *Curr. Opin. Microbiol.*, **1**, 271–277.

Wilson,K. (2001) Preparation of genomic DNA from bacteria. *Curr. Protoc. Mol. Biol.*, **56**, 2.4.1–2.4.5.

Yang,B. *et al.* (2010) Unsupervised binning of environmental genomic fragments based on an error robust selection of l-mers. *BMC Bioinformatics*, **11**, S5.

Zhou,F. *et al.* (2008) Barcodes for genomes and applications. *BMC Bioinformatics*, **9**, 546.