Rapid and Accurate Estimation of Genetic Relatedness Between Millions of Viral Genome Pairs Using MANIAC

Wanangwa Ndovie[1,2], Jan Havránek[3,†], Janusz Koszucki[1,2], Jade Leconte[1], Leonid Chindelevitch[4], Evelien M. Adriaenssens[5] & Rafal J. Mostowy[1,†]

1) Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland

2) Doctoral School of Exact and Natural Sciences, Jagiellonian University, Krakow, Poland

3) Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Krakow, Poland

4) Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, UK

5) Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom

† Correspondence: j.havranek@gmail.com, rafal.mostowy@uj.edu.pl

Abstract

The estimation of Average Nucleotide Identity (ANI) plays a pivotal role in microbial and viral research, facilitating species delineation, taxonomy, genome dereplication in metagenomics and even detection of horizontal gene transfer. Traditional tools, optimised for bacterial genomes, fall short in addressing the complexities of phage genomics such as high sequence variability, mosaicism or the absence of universally shared genes. To bridge this gap, we introduce MANIAC (MMseqs2-based ANI Accurate Calculator), aiming to accurately estimate ANI and alignment fraction (AF) between pairs of viral genomes, using the MMseqs2 software which combines alignment-free and alignment-based approaches. We evaluated MANIAC against the gold-standard ANIb using complete phage genomes and further validated its performance with simulated and real genomic data. MANIAC demonstrated a near-perfect correlation with ANIb ($R^2 = 0.999$), outperforming existing tools like fastANI and Mash, especially for genomes below 80% ANI. When applied to hundreds of millions of pairs of phage genomes, MANIAC revealed a bimodal ANI distribution amongst phage populations, pointing to the existence of an 'ANI gap' similar to that observed in bacterial populations, albeit with quantitative differences. We then used a machine learning approach to classify same-genus pairs by combining both ANI and AF metrics, showing its strong predictive power (PR-AUC=0.972), particularly in virulent phages (PR-AUC=0.991). These findings underscore MANIAC's potential to significantly advance viral genomics by providing a more accurate framework for quantifying genetic relatedness between viral genomes. MANIAC can be accessed under `https://github.com/bioinf-mcb/MANIAC`.

Introduction

Average Nucleotide Identity, or ANI, is one of the most used ways to estimate genetic relatedness between microbial organisms. ANI is calculated by aligning the nucleotide sequences of orthologous genes and then determining the percentage of nucleotides that are identical between the two sequences. This value is then averaged over all the orthologous genes that are compared[1,2]. While it only approximates evolutionary distance – and only for closely related strains – ANI has played an important role in multiple areas of microbial research, including species delineation and taxonomy[3,4,5,6,7,8,9,10,11], detecting horizontal gene transfer events[12,13] and in metagenomics[14,15,16].

Initially, the identification of orthologous genes in pairs of microbial genomes relied on local alignment tools like BLAST[17]. However, with the advent of next-generation sequencing technologies, the number of publicly available microbial genomes to date made such approaches computationally unfeasible. Hence, in recent years various new tools have been proposed to facilitate millions of pairwise calculations. These approaches can be broadly subdivided into two categories: alignment-based and alignment-free methods. Alignment-based approaches have attempted to substitute BLAST with more efficient approaches like MUMmer[18,19] but those are typically less sensitive than BLAST and are still relatively computationally expensive. Alignment-free approaches on the other hand use k-mers to either estimate ANI directly[14,20] or to find areas of local alignment[7,21]. While much more efficient and scalable, alignment-free approaches tend be less accurate in calculating sequence similarity between distantly related genomes. Nevertheless, the scalability of these tools has facilitated large-scale comparisons on unprecedented scales, providing important data in relation to ongoing debates about the universality of genetic boundaries and a single threshold for species delineation in bacteria[7,4,22].

Even though applications of ANI in evolutionary microbiology have been extensive, most of the focus has been on bacteria rather than viruses. At the same time, in recent years ANI has been increasingly used in viral genomics, with important applications like identification of new viruses from metagenomes[23,24], decontamination of bacterial strains from viral sequences[25], taxonomic assignment of new viral strains[26,27,28,29] and studying genetic boundaries between viral populations[30,31,32]. However, currently available tools have been primarily optimised for bacterial genomes, mainly to delineate them at the species level of approximately 95% of ANI. Genomes of viruses, particularly those infecting bacteria, differ from bacterial genomes in important ways. For example, viruses–compared to bacteria–have a much

higher variability in nucleotide sequences, are often highly mosaic, are much shorter and do not universally share genes[33,34,35]. These characteristics present challenges that may not be adequately addressed by methods primarily developed for bacterial genomes. Indeed, some methods have been specifically developed (or used) to estimate genetic relatedness between viral genomes, including PASC[36], DEmARC[37], Gegenees[38] or SDT[39]. However, these tools do not calculate the proportion of the aligned genome for which genetic similarity is calculated, which can bias estimates of mosaic viral pairs[40]. An alternative approach was thus recently proposed, called VIRIDIC[40], which estimates both ANI and Alignment Fraction (AF) between pairs of viral genomes but, relying solely on BLAST, it becomes very costly to scale beyond a few hundred genomes or 10^5 pairs. Here, we argue that a large-scale quantification of genetic relatedness in viruses should (a) estimate both ANI as well as AF to account for genetic mosaicism, (b) accurately estimate ANI at the level of 70% to better delineate viral taxonomic units and (c) be scalable to datasets of at least several thousand (and thus millions of pairs of) viral genomes.

To this end, we introduce a new approach called MANIAC (MMseqs2-based, ANI Accurate Calculator) that efficiently tackles these challenges. MANIAC is both fast and sensitive by employing MMseqs2[41] which filters the DNA sequence space with an alignment-free approach before accurately estimating ANI using an alignment-based approach. By tuning its parameters and using simulated data, we demonstrate that MANIAC can accurately estimate ANI even for viruses belonging to different genera (<70% ANI). We then apply MANIAC to a dataset of 21,754 complete viral genomes from NCBI Genbank. By accurately estimating ANI and AF between hundreds of millions of pairs of viral genomes, we investigate the existence of an 'ANI gap' in phage populations. Using the Virus Metadata Resource from the International Committee on Taxonomy of Viruses (ICTV), we then apply machine learning to classify same-genus pairs of phages using ANI and AF and make recommendations about thresholds that are best suited for large-scale taxonomic assignment of dsDNA bacterial viruses.

Methods

MANIAC Workflow

MANIAC is written in Python3 and implemented in Snakemake[42]. MANIAC runs in three modes: (i) Genome Mode, which takes entire genome sequences as input, (ii) Coding Sequence (CDS) Mode, which takes as input nucleotide sequences of predicted genes, commonly obtained using tools like Glimmer or

Prodigal, and (iii) Protein Mode, which uses amino acid sequences for analysis and focuses on calculating Amino Acid Identity (AAI) rather than ANI.

ANI calculation

In the Genome Mode, MANIAC follows the approach proposed by Goris and colleagues[1]. First, genomes are divided into non-overlapping fragments of length $F_L = 1020$ bp. Next, the `search` module of MMseqs2[41] is used to find homologues between these fragments (query) and the full genomes (reference), filtering all hits that show less that $F_I = 30\%$ identity over the full fragment length and less than $F_C = 0.7$ fragment (query) coverage. MMseqs2 operates through a multi-stage process to efficiently identify similar sequences. During the prefiltering stage, it utilises a precomputed index table containing k-mers from the reference database, allowing for a rapid comparison against the query sequences. This enables the system to quickly identify potential matches, which are then subjected to a more detailed alignment stage. Once aligned, ANI is then calculated as the mean percentage identity over the length of aligned nucleotides. For each pair of genomes, A and B, we first calculate ANI when A is the query and B is the reference, and then when B is the query and A is the reference. We assume a single value of ANI per pair which is the average of these two values. In the CDS and Protein Modes, MANIAC takes as input user-specified CDSs (in nucleotide or amino-acid space, respectively) and uses a best-bidirectional hits approach to approximate orthology[43,44]. The values of the parameters F_I and F_C are the same as in the Genome Mode.

AF calculation

MANIAC also calculates an alignment fraction (AF) for each pair of genomes. Specifically, for each pair of genomes, A and B, we first calculate the total alignment length (i.e., sum of the length of all aligned fragments) as the average of the two scenarios: when A is the query and B is the reference, and then when B is the query and A is the reference. AF is then calculated as a ratio of this total alignment length to the length of the corresponding genome A or B. We report both values as well as their mean, minimum and maximum AF. We consider only pairs with a minimum alignment length of 500 (the remaining ones are filtered out which means that their AF is considered to be 0). The final output of MANIAC is a table with a unique occurrence of each genome pair, reporting all metrics calculated by the programme for this pair.

Controlling sensitivity and specificity

In MMseqs2, the k-mer size flag (`-k`) controls the specificity and sensitivity of the search by specifying the length of the k-mers used in both the query and the precomputed MMseqs2 index table. For nucleotide sequences, such as those used in MANIAC's Genome and CDS Modes, the k-mer size plays a critical role in balancing the accuracy of the results. While a higher k-mer size can enhance specificity by allowing only exact or very close k-mer matches, it may also reduce sensitivity by missing alignments where sequences may be more divergent. Conversely, using shorter k-mers increases sensitivity by identifying a broader range of potential alignments but it also may reduce specificity by incorporating matches that are less precise. This trade-off necessitates careful selection of the k-mer size to achieve an optimal balance between sensitivity and specificity, tailored to the specific requirements of each analysis.

In contrast, the sensitivity flag (`-s`) primarily influences the search in MANIAC's Protein Mode. Higher sensitivity values in this context generate a longer list of similar k-mers, enabling the identification of sequence pairs with lower sequence identity. This increases the ability to find more remote homologues but can also reduce the speed of the search. In nucleotide searches, the sensitivity parameter may have less impact due to the nature of nucleotide matching and the limited nucleotide alphabet, making k-mer length the more critical factor in controlling the search. In this study, we did not see any evidence of the sensitivity parameter affecting the results of nucleotide searches. This is probably due to the nature of nucleotide matching and the limited nucleotide alphabet, making k-mer length the more critical factor in controlling the search.

MANIAC parameters

By default, MANIAC is run with the following parameters for Genome and CDS Modes:

```
--search-type 3 -a --max-seqs 10000 --max-seq-len 100000
-s 7.5 --mask 0 -e 1e-15 -k 11 --zdrop 150
```

and with the following parameters for the Protein Mode:

```
--search-type 1 -a --max-seqs 10000 -s 7.5 --mask 0 -e 1e-15 --zdrop 150
```

These parameters are explained in detail by the authors of MMseqs2. Briefly, `--search-type 3` refers to the genetic code used (nucleotide or amino-acid); `-a` specifies the alignment mode that contains

information necessary to compute ANI and AF; `--max-seqs` specifies the maximum number of sequences aligned with the query and was strongly increased from the default value of 300 to avoid the loss of aligned pairs; `--max-seq-len` specifies the longest unsplit sequence in the database and was increased from its default value to prevent splitting of viral genomes shorter than 100kb which could slightly bias the AF estimates; sensitivity `-s` was chosen as the highest value to maximise sensitivity; `-mask` controls masking of the low complexity regions which was turned off to ensure a comprehensive assessment of genomic similarity; `-e` specifies the e-value chosen based on[18]; `-k` was chosen as 11 based on the analyses conducted in this paper (see Results); `-zdrop` is the nucleotide drop-off value chosen based on the xdrop value used in[18].

The final output of MANIAC consists of pairs of genomes aligned in either of the three modes (Genome, CDS, Protein). If a pair that was aligned in one mode but was missing in another, it was added to the final output with the zero values of ANI/AAI and AF. This approach ensures the inclusion of pairs with any degree of relatedness identified in at least one mode, while pairs not aligned in any mode are excluded from the final output to manage the data size effectively.

Estimating ANI and/or AF using other approaches

pyani

We calculated ANI and AF using the ANIb method implemented in pyani[18]. The ANIb method fragments one genome (query) in pieces of 1020 bp and compares it to a target genome using BLAST+. For each comparison between genomes, pyani calculates ANI following the approach of Goris et al.[1], namely as the percentage identity of the aligned fragments of at least 0.3 percentage identity over the full fragment length and at least 0.7 fragment coverage; pyani also calculates alignment coverage (fraction of query that is aligned). The average of query coverage and subject coverage returned by pyani was directly comparable to our definition of alignment fraction (AF).

Mash

Mash performs pairwise genome comparisons using the ultra-fast MinHash technique to create small signatures known as sketches[14]. These sketches, rather than whole genomes, are then compared to each other. The method then estimates Mash distance based on the rate of sequence mutation, assuming

a simple evolutionary model with a Poisson-distributed mutation rate[45]. Such a distance can then be approximated as $1 - \text{ANI}$ for closely related strains[14]. Following the approach recommended by Mavrich & Hatfull[30], we ran Mash (version 2.3) with `k=15` and further doubled the sketch-size (`-s`) to 50,000 to maximise the accuracy; we also used a probability threshold of $p = 10^{-10}$. Henceforth, in this work we refer to Mash distances as Mash-estimated ANI values. Mash does not calculate sequence coverage and hence cannot be compared to MANIAC's estimate of AF.

FastANI

FastANI is an alignment-free tool that calculates ANI and genome coverage[46]. For a given pair of genomes, FastANI fragments the query genome into non-overlapping fragments and then maps them onto the reference genome using mashmap. Spatially proximate mappings are grouped into bins, and ANI is determined by calculating the weighted mean of the average nucleotide identities across these bins. The coverage for a genome pair using FastANI is quantified as the proportion of query genome fragments that successfully map to the reference genome, and hence directly comparable to MANIAC's estimate of AF. We ran FastANI (version 1.33) with the following parameters: `k=15` (following the recommended length for Mash), `fragLen=500` (based on the fragment length analysis by[32]), and `minFraction=0.1` (to capture pairs with low AF but avoiding noise stemming from short alignments). The ANI of each pair was then taken as the mean of the query-reference ANI and query/fragment AF was reported as the coverage calculated by FastANI.

Data used for benchmarking

Genomic data

In October 2022, we downloaded a comprehensive dataset consisting of 22,774 phage genomes, along with associated metadata, from the Inphared database[47]. Genes were predicted with prodigal using the `-p meta` option[48]. To refine the dataset for accuracy and relevance, we incorporated the "exclusion list" from the most recent Inphared update as of March 2024, resulting in the exclusion of certain isolates and a revised count of 22,616 genomes. Of those, 4,655 were part of RefSeq and are henceforth referred to as the **RefSeq dataset**; 500 of those genomes were randomly selected to benchmark the accuracy and efficiency of MANIAC against established methods such as ANIb[18], FastANI[7], and Mash[14] (the

RefSeq 500 dataset). We also downloaded taxonomic information from the Virus Metadata Resource from the International Committee on Taxonomy of Viruses (ICTV), version 38[49]. We then employed CheckV[25] to evaluate the completeness of the Inphared genomes, setting a threshold for inclusion at greater than 90%, resulting in a curated dataset of 21,754 high-quality phage genomes, among which 7,752 are acknowledged in ICTV VMR v38. Additionally, we constructed a distinct taxonomic dataset from the VMR v38 and removing any isolates not present in the Inphared database or listed on the exclusion list. In cases of isolates appearing in both RefSeq and GenBank, we identified and eliminated segmentations and duplications to avoid redundancy. This approach yielded a refined dataset of 4,163 unique species of bacterial viruses, henceforth referred to as the **ICTV dataset**, which was used for machine learning of the taxonomic same-genus predictions. All four datasets with the corresponding accessions and relevant metadata are provided as Supplementary Data.

Simulated Data

The simulated dataset contained simulated genome-like string pairs (henceforth called simulated genome pairs) with predefined ANI and AF values. Each genome pair thus shared a proportion of orthologous gene-like strings (henceforth called genes) and a proportion of unrelated genes, the balance between which determined the AF (e.g., if all genes were orthologous, the AF was 1). To simulate gene lengths, we investigated the lengths of protein sequences from the NCBI RefSeq database and found them to follow a log-normal distribution with a mean of 2.7 and a standard deviation of 0.33 using base-10 logarithms. The protein lengths were then sampled from this log-normal distribution and multiplied by 3 to represent the corresponding nucleotide sequence lengths. For simplicity, we assumed identical genome lengths for each pair, with each genome consisting of $N = N_O + N_U$ genes, where N is the total number of genes per genome, N_O is the number of orthologous genes, and N_U is the number of unrelated genes, such that $AF = \frac{N_O}{N}$. For orthologous genes, we used SeqGen[50] to simulate genetic divergence within a gene with the F81 model and a given value of evolutionary distance d, which was then translated into ANI by directly counting the proportion of identical bases to all bases in the gene. Unrelated genes were simulated as independent instances of random DNA strings. To imitate the Genome Mode, orthologous genes were concatenated in the same order in both genomes while unrelated genes were added in a random order. To simulate the deviation from synteny, we shuffled a proportion s of orthologous genes in one genome by randomly selecting $\text{floor}(s \cdot N_O)$ genes and resampling them without replacement. We varied

the evolutionary distance d from 0 to 0.8 (corresponding to ANI between 1 and 0.5), and AF from 1 to 0.1, in increments of 0.05, with 20 genome pairs simulated for each combination of ANI/AF, yielding a total of 6,840 simulated genome pairs from 13,680 simulated genomes. By default, we assumed $N = 20$ and $s = 0$.

Measuring accuracy of prediction and runtime

We calculated the Root Mean Squared Error (RMSE) as the mean difference between the predicted and actual values of ANI and AF values. When assessing RMSE on real data, the actual value was the one estimated by ANIb. Confidence intervals for linear regression were calculated by performing 1000 bootstrap iterations to estimate the coefficients. We then calculated the median and 95% quantile range for these coefficients. To estimate the run times for ANI calculations, we executed all commands on a Linux machine with 32 cores and 300GB of RAM and used the `time` command to measure the actual run times of each programme.

Lifestyle and Taxonomic Prediction

We used BACPHLIP v0.9.6 to predict the phage lifestyle[51]. Phages were assigned as virulent if their score was 0.95 and greater and as temperate if their score was less than 0.05, yielding 6,036 isolates predicted as virulent, 4,908 as temperate, and 10,655 without lifestyle prediction. We used different releases of the Virus Metadata Resource from the International Committee on Taxonomy of Viruses (ICTV) to assign genus-level taxonomy to the phages under study[49]: VMR 38 (default) and 37, 36, and 35 for comparison.

Same-genus classification with machine learning

We employed a logistic regression approach to classify pairs of phage genomes as belonging to the same or different genera based on their Average Nucleotide Identity (ANI) and Alignment Fraction (AF) values. To this end, we used the "ICTV dataset", which consisted of 4,163 representative genomes of bacterial virus species. We ran MANIAC in all three modes (Genome, CDS, Protein) and found 745,209 pairs for which ANI or AAI estimates were made. Logistic regression was chosen for its simplicity, interpretability, and effectiveness in binary classification tasks, making it well-suited for analysing the relationship between genomic similarity metrics (ANI and AF) and taxonomic classification.

Model Description

The logistic regression model was formulated with ANI (or AAI) and AF as the independent variables and the binary outcome variable indicating whether a pair of genomes belongs to the same genus (`TRUE`) or not (`FALSE`). The model estimates the probability of the `TRUE` outcome based on their ANI (or AAI) and AF values, using a logistic function to map linear combinations of the input variables to probabilities.

Data Splitting and cross-validation

To ensure a rigorous assessment of model performance, the ICTV dataset comprising pairs of phage genomes, was initially divided into two distinct sets: (i) a combined training/validation set and (ii) a separate testing set. This division was executed at the genome level, not at the level of pairs, adhering to a ratio of 65% for training/validation and 35% for the test set. Consequently, the distribution of genome pairs followed an approximate 70%/30% split, respectively. For the process of model optimisation and parameter tuning, we exclusively utilised the training/validation set, implementing a 10-fold cross-validation strategy. Specifically, this set was further divided into ten equal parts, or 'folds.' During each iteration of cross-validation, nine of these folds were used for training the model, while the remaining single fold served as a validation set to assess model performance. This cycle was repeated ten times, ensuring that each fold acted as the validation set once, thereby allowing for a comprehensive internal evaluation of the model's predictive capabilities. It is important to note that the results derived from this 10-fold cross-validation process pertain solely to the training/validation phase, aimed at refining our models within a controlled subset of the data. In contrast, the independent testing set was reserved for a final, unbiased evaluation of the model's performance. The metrics reported for the testing set, therefore, reflect the model's ability to generalise to new, unseen data, distinct from the internal validation achieved through cross-validation.

Model Evaluation

For model evaluation, we focused on precision-recall metrics, given the class imbalance typically observed in genomic datasets. In the context of our study, precision (the proportion of true positive predictions in all positive predictions) and recall (the proportion of true positive predictions in all actual positives) are particularly informative:

- **Precision** is critical in ensuring that when pairs are classified as belonging to the same genus, this classification is likely correct, minimising false positives.
- **Recall** assesses the model's ability to identify all relevant pairs that truly belong to the same genus, minimizing false negatives.

The choice of the precision-recall (PR) curve and area under the PR curve (PR-AUC) as metrics for model performance is justified by the imbalanced nature of our dataset. This imbalance is evident, as the number of pairs belonging to the same genus is significantly lower than the number of pairs from different genera, rendering accuracy or the traditional ROC-AUC metric less suitable for our analysis. Both precision and recall take into consideration only cases where either the true class or the predicted class is positive, and neither metric relies on the true negative pairs, which comprise the majority of pairs for which the ANI estimation failed. The uncertainty in the PR-AUC was estimated using bootstrap resampling (1,000 iterations), where predictions and true labels were randomly sampled with replacement. The PR-AUC scores derived from these bootstrap samples were then used to calculate the 95% confidence interval.

Results

MANIAC Achieves Near-Perfect Correlation with BLAST in ANI and AF Estimation

To evaluate MANIAC's performance, we benchmarked its estimates of Average Nucleotide Identity (ANI) and Alignment Fraction (AF) against ANIb, which is widely regarded as the gold standard in the field. We used pyani, an implementation of the original methodology by Goris et al., to calculate ANIb values (see Methods for details). Our analysis employed a dataset comprising 500 randomly sampled phage genomes from the NCBI RefSeq database (RefSeq 500 dataset; see Methods). In addition, we compared MANIAC's results to those from fastANI (for both ANI and AF) and Mash (for ANI only).

As illustrated in Figure 1, MANIAC achieved an exceptionally high correlation with ANIb in estimating both ANI (Figure 1A) and AF (Figure 1B). Specifically, the R-squared (R^2) value, indicating the proportion of variance in Pyani's ANI estimates explained by MANIAC, was 0.9989 [95% CI: 0.9986, 0.9993], while for AF, the R^2 was 0.9997 [95% CI: 0.9996, 0.9997]. Further supporting these findings, the

coefficients of the linear regression model for MANIAC, expressed as $Y = \beta_0 + \beta_1 X$ (where X are values given by ANIb and Y values given by MANIAC), were determined as $\beta_0 = -0.0044$ [95% CI: -0.0052, -0.0035] and $\beta_1 = 1.0046$ [95% CI: 1.0036, 1.0055] for ANI, and $\beta_0 = 0.0005$ [95% CI: 0.0003, 0.0007] and $\beta_1 = 0.9928$ [95% CI: 0.9917, 0.9938] for AF, thus in both cases closely approximating the $Y = X$ relationship.

By contrast, both fastANI and Mash displayed lower levels of correlation with ANIb, with $R^2 = 0.9141$ [95% CI: 0.8969, 0.9298] for fastANI and $R^2 = 0.8237$ [95% CI: 0.8059, 0.8406] for Mash. Moreover, these tools deviated significantly from the ideal $Y = X$ relationship between ANIb and estimated ANI values, (fastANI: $\beta_0 = 0.307$ [95% CI: 0.297, 0.318] and $\beta_1 = 0.674$ [95% CI: 0.659, 0.686]; Mash: $\beta_0 = -0.216$ [95% CI: -0.236, -0.197] and $\beta_1 = 1.213$ [95% CI: 1.187, 1.240]), suggesting that these approaches are prone to over- or underestimating ANI for more distantly related pairs of genomes. Finally, fastANI's AF estimates underperformed relative to those of MANIAC with $R^2 = 0.927$ [95% CI: 0.919, 0.934] (see Figure 1B).

In spite of near-perfect correlations achieved for MANIAC, we noticed that there were 18 pairs from 29 genomes where the ANI estimated by MANIAC deviated by 2-6 percentage points from the value estimated by ANIb. Upon further investigation, we found that these pairs were characterised by a low AF estimate (AF$<$ 0.1 by both ANIb and MANIAC). We thus next checked how much these correlations would improve when discarding pairs with a very low value of AF, recalculating all correlation coefficients for pairs with the estimated value AF$\geqslant$ 0.1 by ANIb. The results (Figure 1C and 1D) show that while all correlation coefficients improved (MANIAC: $R^2 = 0.9997$ [95% CI: 0.9996, 0.9997], fastANI: $R^2 = 0.978$ [95% CI: 0.975, 0.981], Mash: $R^2 = 0.922$ [95% CI: 0.911, 0.932]), MANIAC was still superior in reproducing the estimates of ANIb.

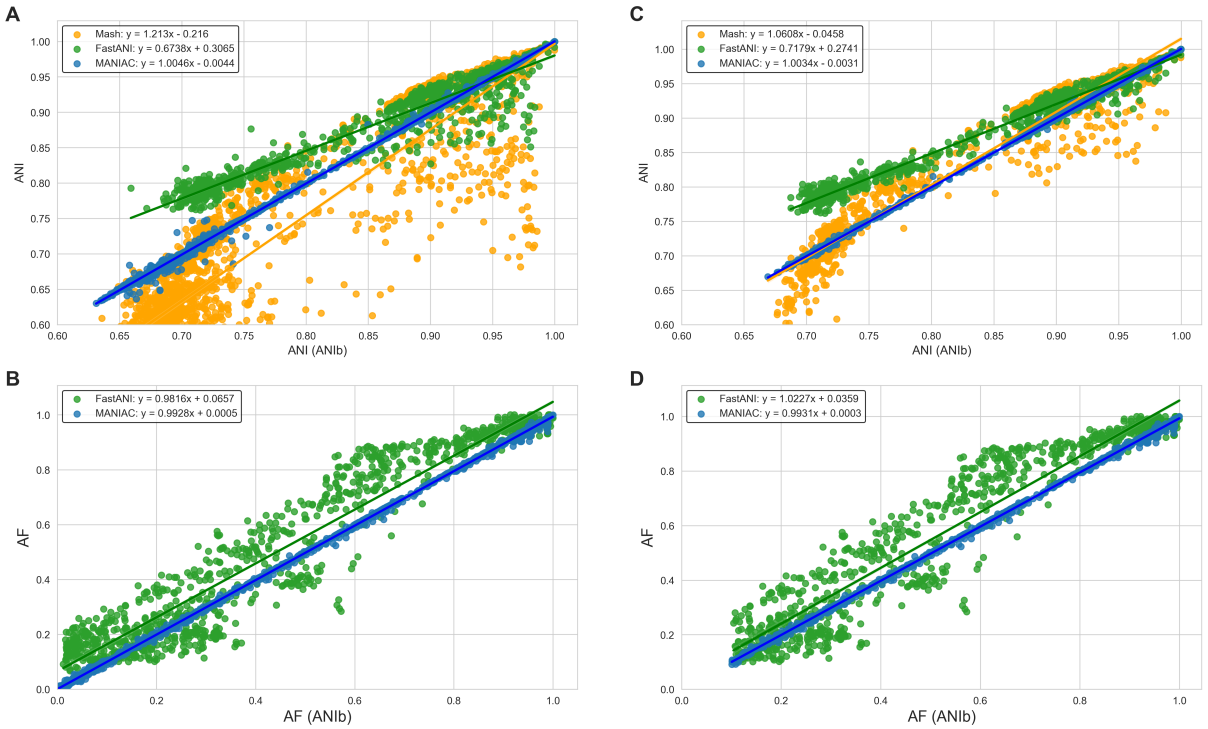

Figure 1. Comparison of estimated ANI and AF values using the RefSeq 500 dataset. (A) ANI estimated by pyani (ANIb; X-axis) vs. ANI estimated by MANIAC (blue), Mash (orange) and fastANI (green). (B) AF estimated by pyani (X-axis) vs. AF estimated by MANIAC (blue) and fastANI (green). (C) Same as panel A but for AF$\geqslant$ 0.1. (D) Same as panel B but for AF$\geqslant$ 0.1.

MANIAC is fast and scalable to millions of pairs of viral genomes

We next examined how MANIAC's speed and scalability compared to other ANI-estimation tools. Additionally, we explored the balance between speed and accuracy depending on the k-mer size, a crucial parameter in MMseqs2 that influences the dimensions of the precomputed index table (see Methods for details). We varied the k-mer sizes from 7 to 15 and assessed the accuracy of ANI estimates through the Root Mean Squared Error (RMSE) metric, using ANIb as a reference. The same dataset used in Figure 1 was employed for this analysis. Our findings, depicted in Figure 2, highlight three key observations.

First, across all tested k-mer sizes, MANIAC consistently outperformed both fastANI and Mash in terms of accuracy for ANI and AF calculations (Figure 2A). Second, MANIAC's accuracy saw noticeable improvements when the k-mer size was increased from 7 to 9 (Figure 2A). Beyond this point, accuracy plateaued for k-mer sizes of 9, 11, and 13, and then exhibited a slight decline at a k-mer size of 15. This pattern aligns with the intuition that excessively small k-mers introduce noise due to random matches, while overly large k-mers risk overlooking nearly identical sequence fragments. Third, MANIAC's computational speed was inversely related to k-mer size, becoming increasingly comparable to fastANI's runtime when the k-mer size reached 15, although with superior accuracy (Figure 2B).

In summary, when benchmarked against ANIb, the gold standard for ANI estimation, MANIAC emerges as a viable option for evaluating ANI and AF in viral genomes. It offers a combination of computational efficiency and accuracy, in some instances matching the speed of fastANI while retaining a level of precision on par with ANIb. This makes MANIAC a compelling alternative to alignment-free tools like fastANI and Mash.

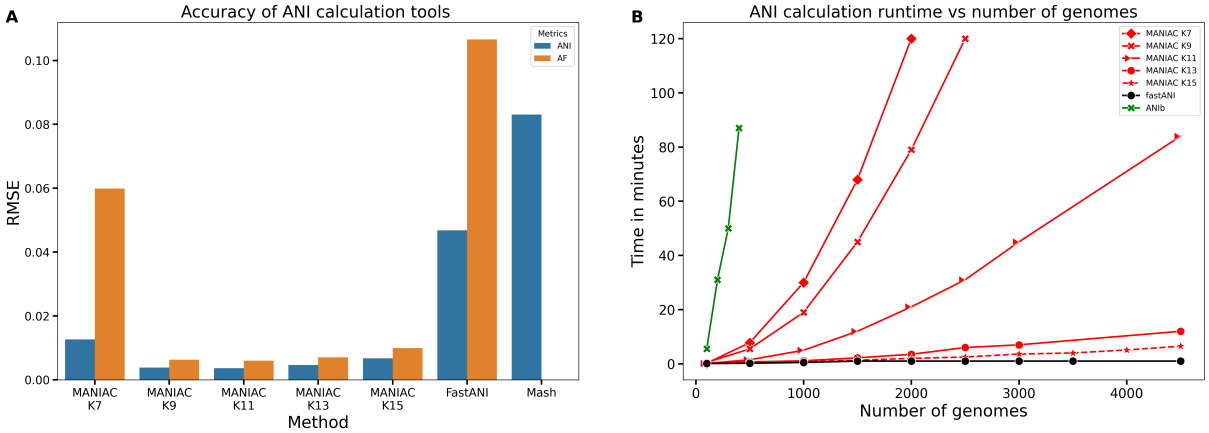

Figure 2. (A) Root Mean Squared Error (RMSE) of ANI (blue) and AF (orange) calculation when using ANIb as the benchmark using the same dataset as in Figure 1 for three approaches: MANIAC (with odd k values ranging from 7 to 15), fastANI, and Mash (only ANI). Data obtained with a random subset of 500 genomes from RefSeq. (B) Runtime as a function of the number of genomes when calculating all-by-all distances for different approaches, with the gold standard ANIb plotted in green and other approaches in red with different point shapes. ANIb calculation is capped at 500 genomes (dataset from panel A), and the remaining are calculated on the entire RefSeq ($n = 4655$).

MANIAC accurately estimates ANI for simulated genome pairs

An ideal benchmark for any method to estimate ANI (Average Nucleotide Identity) and AF (Alignment Fraction) should involve knowing the full evolutionary history and hence exact alignments of any pair of sequences, which is impossible. Consequently, the scientific community has largely defaulted to using BLAST-based approaches as a proxy gold standard. However, those approaches, such as the method by Goris et al. implemented in pyani, were originally developed with bacterial genomes in mind. These tools may not offer equivalent accuracy for viral genomes, which often present higher levels of sequence divergence, extensive genetic mosaicism and much shorter alignable regions.

One pivotal but sometimes overlooked aspect of BLAST-based approaches is the scoring system used for DNA sequence alignment. Scoring matrices not only gauge the quality of an alignment but also influence the method's sensitivity and specificity. Importantly, the scoring system inherently reflects the expected level of sequence divergence, essentially setting the stage for how effectively the algorithm will identify related sequences[52]. For example, BLASTN, a commonly used algorithm, uses a scoring system that adds 2 points for nucleotide matches and subtracts 3 points for mismatches. Such a scoring system presumes the level of divergence at ANI around 90% and is indeed optimal for studying bacterial species delineation but may not be suitable for viral genomes[52,53]. Specifically, in viral taxonomy, a 70% ANI threshold is recommended as the first criterion for classifying double-stranded DNA tailed phages into genera[27], while in eukaryotic viruses single-stranded DNA viruses which belong to the same species can often have ANI of 70% or even lower[54]. We thus hypothesised that alternative scoring systems might better estimate ANI at this critical 70% level, pivotal for taxonomic classification across the spectrum of viral diversity.

To address this hypothesis, we decided to benchmark MANIAC by generating a simulated dataset consisting of pairs of genome-like strings (here called genome pairs; see Methods) with predefined values of ANI and AF. This approach allows us to explore the impact of different scoring systems on the accuracy of ANI and AF calculation, while also providing a tailored benchmark that is more relevant for viral genomes. We considered three scoring schemes: the standard BLASTN scoring for divergence levels of around 90% (match +3, mismatch -2), UNIT scoring for divergence levels of around 75% (match +1, mismatch -1), and WU scoring for divergence levels of around 65% (match +5, mismatch -4). We then used the RMSE metric to investigate how well MANIAC estimated the simulated values of ANI (via evolutionary distance parameter d) and/or AF depending on their range, its parameters as well as

departure from synteny (modelled by randomisation of the gene order; see Methods). We also defined a range of ANI $\geqslant 0.7$ and AF $\geqslant 0.5$ as a 'taxonomic genus range' (due to the fact that the large majority of viral genomes used in this study are tailed phages).

The comparison of the simulated and estimated values of ANI and AF are shown in Figure 3. In our comprehensive analysis, MANIAC demonstrated superior accuracy in estimating ANI compared to fastANI or Mash, particularly for genomes with an ANI below 80%, as depicted in Figure 3A. Notably, MANIAC exhibited near-perfect performance within the range of $d < 0.4$ (or ANI $\geqslant 0.7$) and AF $\geqslant 0.5$, as shown in Figure 3B. Furthermore, the UNIT and WU scoring schemes proved to be more accurate than BLASTN in estimating ANI, especially in the range of ANI approximately 0.6 to 0.7 and even below an ANI of 0.6, as evidenced in Figure 3A and Supplementary Figure S1. Results for the WU scoring scheme are omitted from Figure 3 because they are quantitatively identical as those for the UNIT scoring. Consistent with previous findings using ANIb as a benchmark, MANIAC's optimal performance was observed for k-mer sizes of 9 and 11, with the latter offering a faster processing speed (refer to Supplementary Figure S1 and Figure 2B).

In analogy to ANI, simulations showed that MANIAC in the Genome Mode was superior to fastANI in estimating AF for more distant genomes (see Figure 3C and 3D). However, we found that the accuracy of AF estimation by MANIAC was modulated by two effects. First, deviation from synteny, modelled by randomisation of the order of orthologous genes, led to an underestimation of AF even for genomes with a high value of ANI (Supplementary Figure S2). Second, low ANI may lead to the failure of fragment alignment (particularly for short fragments), hence also underestimating AF (Figure 3E and 3F). These effects can be to some extent rectified by using the CDS mode in MANIAC, which estimated ANI and AF (but especially AF) more accurately than the genome mode (Figure 3C-F). However, these results have to be interpreted with caution due to the limitations of our simulation approach (see Discussion). Finally, the underestimation of AF can be partly remedied by using a shorter fragment size in the genome mode (see Supplementary Figure S3). However, this approach is not recommended for two reasons: (i) an increased runtime for shorter fragment sizes and (ii) a phenomenon when distantly related fragments are less likely to align when shorter (e.g., due to effects like sequence complexity or composition) means it may not work for distantly related genomes. Hence, the prediction from the simulation approach is that AF is likely to be underestimated between distantly related genomes due to the failure of fragment alignment close the the twilight zone and the effect of the departure from synteny.

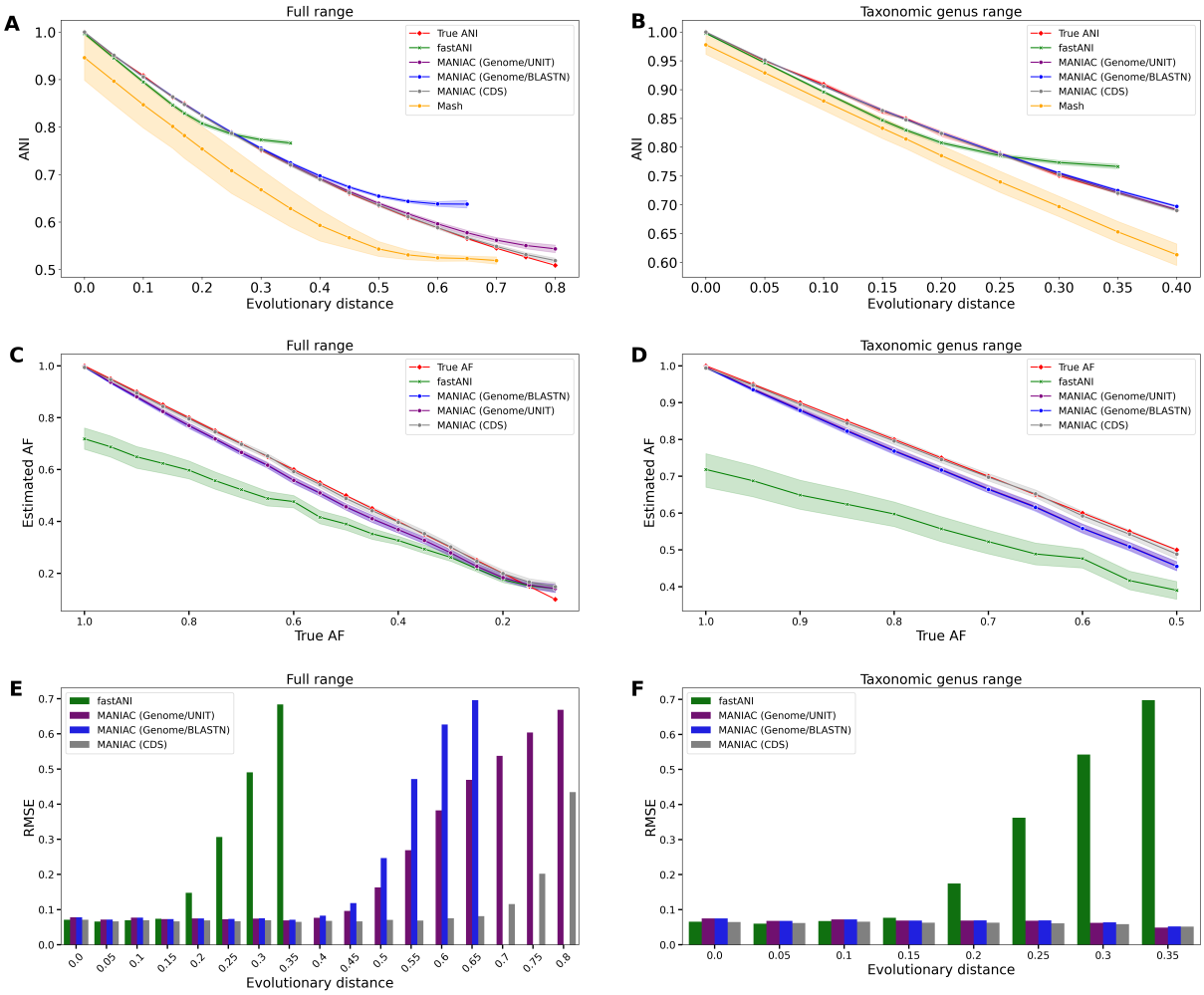

Figure 3. (A) Simulated evolutionary distance (X-axis) vs. true ANI (red) and ANI estimated with fastANI (green), Mash (orange) and MANIAC in Genome Mode with BLASTN scoring (blue), Genome Mode with UNIT scoring (purple) and CDS Mode in UNIT scoring (grey). (B) Same as panel A but recalculated for the 'taxonomic genus range' ($d < 0.4$ or ANI $\geqslant 0.7$ and AF $\geqslant 0.5$). (C) Simulated AF (X-axis) vs. estimated AF (Y-axis). Methods and colours are analogous to panels A and B. (D) Same as panel C but recalculated for the 'taxonomic genus range'. (E) RMSE of AF estimate as a function of evolutionary distance, with colours and methods the same as in panels C and D. (F) Same as panel E but recalculated for the 'taxonomic genus range'. In all panels we assumed $s = 0$.

Validating simulation predictions with real genomic data

The main prediction of the simulation approach is that the choice of the scoring scheme and MANIAC's operational mode (Genome vs. CDS) can affect the bias in estimates of ANI and AF for more distantly related pairs of genomes. Before proceeding to apply MANIAC to specific biological questions, we tested whether these types of bias can be reproduced when using AAI (Average Amino Acid Identity) as a more sensitive approach to estimate genetic relatedness between viral genomes. To this end, we executed MANIAC on the RefSeq dataset ($n = 4655$) in all three modes. We employed a k-mer size of 11 for both the Genome and CDS Modes and a fragment size of 1020 nucleotides for the Genome Mode. In the Protein Mode, which calculates AAI across all genome pairs, we correlated these AAI values with ANI values derived from the Genome and CDS modes. To ensure robustness in our ANI/AAI comparisons, we only included genome pairs with substantial coverage (total alignment length exceeding 5000 amino acids).

Confirming our simulation-based predictions, the scoring scheme significantly impacted the ANI estimations for genome pairs at the borderline of the taxonomic genus range (ANI of 0.7 and below). Figures 4A illustrates a strong correlation between ANI and AAI estimates; however, this correlation varies notably between the BLASTN and UNIT scoring schemes, especially for ANI values below 0.7 (Figure 4A and 4B). The disparity in ANI estimates between these two scoring schemes could surpass five percentage points for genome pairs with an AAI less than 0.8, which could affect ANI-based taxonomic assignment in many viruses.

Also in line with our simulation findings, AF estimation in the Genome Mode was found to be less reliable for more distantly related genomes. This underestimation of AF, as evidenced by comparisons with the Protein Mode (considered here as 'true AF'), became more pronounced with increasing genomic distance (Figures 4C and 4D). The fact that the underestimation of AF, when measured in this way, occurs also for closely related genome pairs, suggests that deviation from synteny (via rampant genetic mosaicism) is one important factor contributing to this effect, but the failure to align fragments is also very likely to contribute as ANI gets close to the twilight zone of 0.5.

Finally, we explored the hypothesis that reduced fragment sizes in the Genome Mode could enhance AF estimation accuracy, particularly for closely related genome pairs as predicted by simulations. (ANI$\geqslant$ 0.7 and AF$\geqslant$ 0.5). Upon analysis, we found that AF estimates in the Genome Mode, when utilising shorter fragment sizes, correlated better with those in the Protein Mode (see Supplementary Figure S4). However,

as expected, neither fragment size matched the accuracy of AF estimation achieved in the CDS Mode (as detailed in Supplementary Figure S4).

In conclusion, through the integration of both simulated and real datasets, we demonstrate that MANIAC, when accurately parameterised, serves as the optimal tool for assessing genetic relatedness among phage genome pairs at the threshold of around 70% of ANI. We thus next employed it to assess the relationship between genetic relatedness and taxonomic assignment using genetic information.

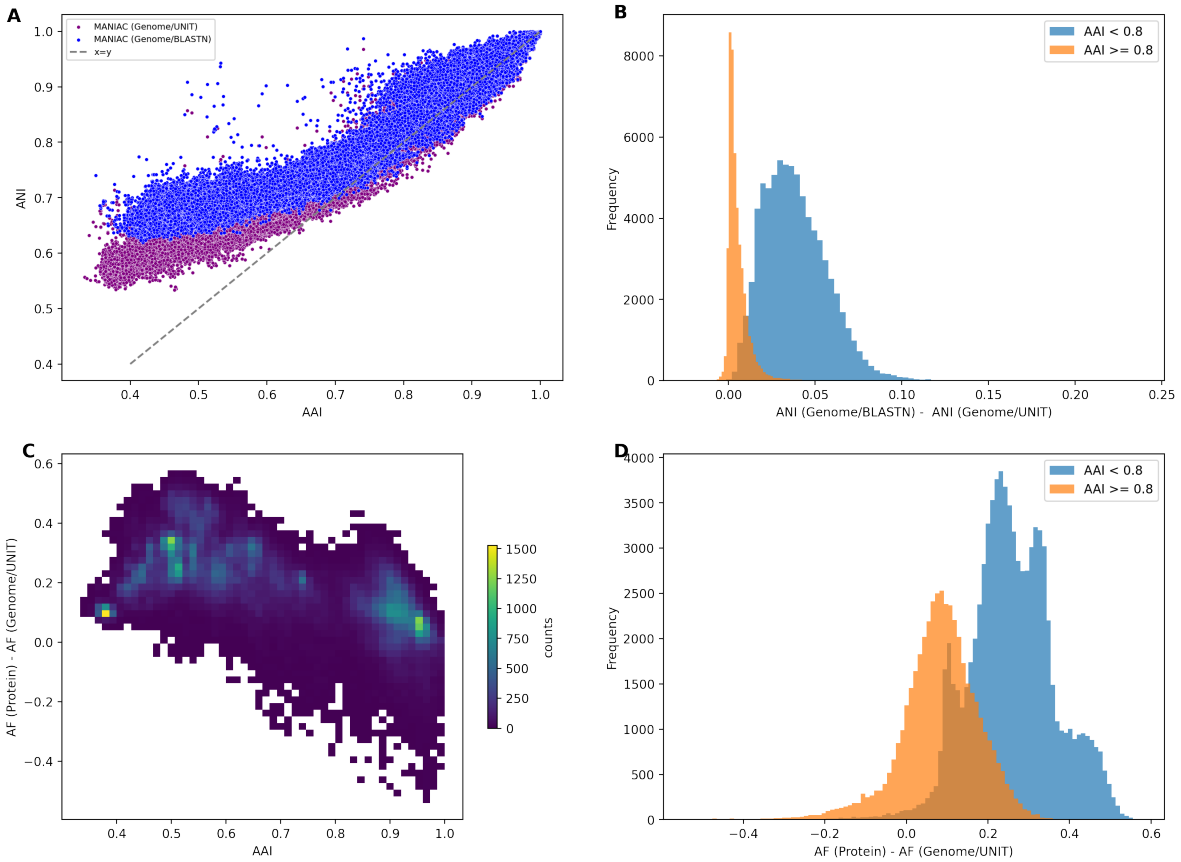

Figure 4. (A) ANI (Y-axis) vs AAI (X-axis) estimated for the RefSeq dataset by MANIAC in the Genome Mode using two scoring schemes: BLASTN (blue) and UNIT (magenta). Dashed line shows the Y=X relationship. (B) Histogram of ANI difference between BLASTN and UNIT scoring schemes for all pairs with AAI⩾0.8 (orange) and AAI<0.8 (blue). (C) 2D histogram of differences between AF measured in Protein Mode and AF measured in Genome Mode (Y-axis) vs the estimated value of AAI (X-axis). (D) Histogram of differences between AF measured in Protein Mode and AF measured in Genome Mode for various ranges of AAI: AAI⩾0.8 (orange) and 0.8>AAI (blue). Only pairs with a minimum alignment length of 5000 amino-acids are shown.

Distribution of ANI and AF values across hundreds of millions of genome pairs

After calibrating and validating MANIAC through both simulated and genomic datasets, we next applied the pairwise calculation of ANI and AF between phage genomes to two biological problems, the first of which is the investigation of the ANI discontinuity in phage populations.

Large-scale comparisons of microbial genomes have revealed that bacterial and archaeal populations exhibit a distribution of genomic sequence similarity, as measured by ANI, which features a distinct gap—commonly referred to as the 'ANI gap'[7,55,56]. In bacteria, this gap typically occurs in the range of 85%-95%, with most intra-population pairs displaying high similarity of ANI $\geqslant$ 95%. Despite suggestions from previous research that phages exhibit a similar type of discontinuity[31], there has been a lack of large-scale comparisons involving complete phage genomes with precise estimations of ANI values below 70%. To help bridge this gap, we ran MANIAC on the Inphared dataset of $n = 21,754$ genomes to calculate ANI and AF values between hundreds of millions of pairs of complete phage genomes. Our analysis, depicted in Figure 5, reveals several observations.

First, our findings align with those of Mavrich & Hatfull, showcasing two evolutionary modes within the phage populations, characterized by a high rate or a low rate of gene flux[30]. Importantly, it should be noted that Mavrich & Hatfull employed Mash to estimate nucleotide distance, which does not distinguish well between ANI and AF for distant genomes. Consequently, their nucleotide distance measurements are influenced by gene content distance, affecting the shape of their plot compared to ours. Despite these methodological differences, the dense clustering of data points in our analysis, indicating high ANI but low AF (see Figure 5A), suggests the presence of many recent horizontal gene transfer (HGT) events, corroborating the two evolutionary modes identified by Mavrich & Hatfull. Thus, while the approaches differ, the underlying signal from the data is consistent across both studies, reaffirming the identification of these two modes in the evolution of phage populations.

Second, the observed ANI distribution among pairs of phage genomes resembles a bimodal pattern, akin to the one observed in bacterial populations (see Figure 5B). One notable difference is that the 'ANI gap' is shifted down to the range of around 78% to 85%, with only 3% of genome pairs residing within that range. To address the hypothesis that the over-representation of pairs for ANI $\geqslant$ 85% may be driven by recent mosaicism observed between many pairs of genomes (high ANI, low AF), stemming from the high rate of gene flux mode, we re-examined the ANI distribution considering only data points with an AF $\geqslant$ 0.5 (see Figure 5C). This analysis confirms that the observed distribution pattern for ANI $\geqslant$ 85%

is not a mere result of a recent gene flux known to occur in phage populations, hence supporting the existence of the 'ANI gap' in phages (see also Discussion).

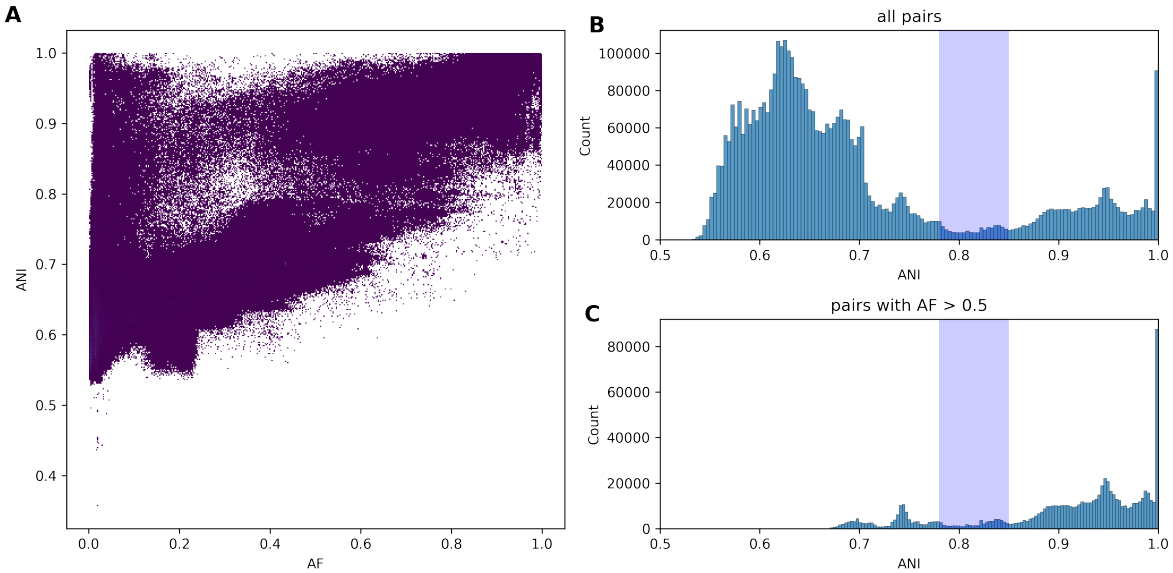

Figure 5. (A) 2D histogram of ANI (Y-axis) vs. AF (X-axis) for all pairs of 21,754 complete phage genomes from the Inphared dataset estimated by MANIAC using Genome Mode. After filtering (see Methods), we found 4,211,185 pairs with a non-zero value of ANI and AF. (B) Histogram of ANI values for all pairs as displayed in panel A, with the 'ANI gap' indicated as the shaded area. (C) Same as panel B but only for 778,330 pairs with AF $\geqslant$ 0.5.

Assessing the quality of ANI and AF as predictors of phage genus taxonomy

Our second application of MANIAC was to apply the pairwise calculation of ANI and AF between phage genomes to the problem of taxonomic assignment, which is challenging in bacterial viruses[57]. Given that establishing viral taxonomy traditionally hinges on a multifaceted approach—incorporating phenotypic and genotypic characteristics such as morphology, genetic content, host range and phylogenetic relationships—metrics like ANI have increasingly become pivotal for standardising taxonomic classifications[58,54,59]. This trend is particularly critical given the surge in sequenced viral genomes from metagenomic studies, many of which remain uncultivated. Notably, the ICTV's Bacterial Viruses Subcommittee recommends using nucleotide identity cut-offs across the genome lengths at 95% for the species rank and 70% for the genus rank for tailed dsDNA phages of the class Caudoviricetes. While threshold-based analyses have been previously used for species delineation among bacterial viruses[31,32], to our knowledge there have been no attempts to systematically examine the genus assignment in dsDNA bacterial viruses depending on the values of both ANI and AF for various modes of their calculation.

To carry out this analysis, we created a representative dataset consisting of 4,163 species of bacterial viruses with the full taxonomic classification provided by ICTV and a corresponding genome sequence (the ICTV dataset; see Methods). We processed these genomes through MANIAC in all three modes (Genome, CDS, Protein). Our visual representation of the dataset (Figure 6A) shows the distribution of ANI and AF among genome pairs by their genus assignment (same or different). The analysis yields several insights. First, while a strong correlation is evident between genetic relatedness and taxonomic genus classification, the absence of a singular, universal cutoff underscores the complexity of viral taxonomy, which relies on multiple layers of evidence. Second, the presence of genome pairs with high ANI yet low AF highlights the prevalence of genetic mosaicism, supporting the notion that ANI alone is not an ideal predictor of bacterial genus. Finally, the bimodality of the ANI distribution on a representative set of ICTV phages supports the hypothesis that this bimodality is not just driven by biased sampling.

To evaluate the predictive utility of ANI and AF for genus classification, we applied a statistical learning approach via logistic regression (see Methods), treating ANI and/or AF as predictors and genus assignment as the binary outcome (same vs. different). After partitioning the dataset into training/validation (65%) and testing (35%) sets, we conducted 10-fold cross-validation across five models (see Methods): ANI in Genome Mode alone (model 1), ANI+AF in Genome Mode (model 2), ANI+AF in CDS Mode (model 3), AAI+AF in Protein mode (model 4), and ANI in Genome Mode + AF in Protein Mode (model

5). Finally, performance of these models was assessed by calculating the area under the precision-recall curve on both training/validation and test datasets (PR-AUC; see Methods and Table 1).

Our results show that while the PR-AUC for model 1 (ANI alone) is relatively high, estimated as 0.814 [95% CI: 0.804, 0.823], the predictive power for model 2 (ANI+AF) is substantially and significantly higher, estimated as 0.970 [95% CI: 0.967, 0.972] (see Figure 6B and Table 1). This finding emphasizes the criticality of both ANI and AF in taxonomic classification, as indeed recommended by ICTV's Bacterial Viruses Subcommittee[27]. Second, we hypothesised that models 3-5 which incorporate more information about orthology and/or are more sensitive in detecting it, may be more accurate in predicting the taxonomic genus than the Genome Mode. Our results did not show any support for this hypothesis, with no significant difference between PR-AUC for model 2 and model 4 or 5 (although model 2 significantly outperformed model 3; see Figure 6B and Table 1). Thus, contrary to our initial hypothesis, the computational intricacies involved in ORF prediction or protein sequence alignment (CDS and Protein modes) did not confer any additional predictive advantage for genus prediction over the straightforward Genome Mode calculation of ANI. Thus henceforth we continue the analyses for model 2 (ANI+AF in Genome Mode) since it offers the best combination of accuracy and simplicity.

Dataset	**Model**	**PR-AUC**	**95% CI**
Train/Valid	ANI	0.8180	(0.8121, 0.8250)
Train/Valid	ANI+AF (Genome)	0.9692	(0.9672, 0.9709)
Train/Valid	ANI+AF (CDS)	0.9672	(0.9652, 0.9691)
Train/Valid	ANI+AF (Protein)	0.9700	(0.9679, 0.9719)
Train/Valid	ANI+AF (Mixed)	0.9701	(0.9681, 0.9719)
Test	ANI	0.8138	(0.8038, 0.8229)
Test	ANI+AF (Genome)	0.9700	(0.9672, 0.9724)
Test	ANI+AF (CDS)	0.9682	(0.9653, 0.9707)
Test	ANI+AF (Protein)	0.9710	(0.9682, 0.9733)
Test	ANI+AF (Mixed)	0.9713	(0.9685, 0.9738)

Table 1. PR-AUC estimates with 95% confidence intervals for 5 principal models obtained on both training/validation dataset and test dataset.

The precision and recall values, influenced by the chosen prediction threshold, highlight a trade-off: higher thresholds improve precision at the cost of recall, and vice versa. Optimal balance, as indicated by the F1 metric, was achieved at a probability threshold of approximately 0.44 (Supplementary Figure S5). To understand how the choice of ANI and AF values impacts the precision and recall scores, we also applied a simple ANI/AF threshold analysis on the entire dataset and calculated the resulting metrics (Supplementary Figure S6). The results suggest that, in the case of model 2, the optimal F1 score is

obtained for ANI ~ 0.8 and AF ~ 0.3. Regardless of the threshold, however, the model is not perfect and produces false positives even at a high probability threshold and false negatives at a low probability threshold. To investigate those outliers, we looked for pairs of different viral genera predicted to be the same with $p \geqslant 0.95$, as well as for pairs of viruses within the same genus predicted as the same with $p < 0.05$ (or, in other words, predicted as different genera with $p \geqslant 0.95$). In the test dataset, the first group (false positives) represented 34 out of 1325 genera for which another genome with a value of ANI/AF was found (see Supplementary Figure S7). The second group (false negatives) in the test dataset was dominated by two genera: *Fromanvirus* (142 out 180 pairs) and *Phietavirus* (27 out of 180 pairs). It is unclear whether these outliers reflect complex taxonomic assignment (e.g., due to genetic mosaicism or extensive genus diversity) or imperfect taxonomic classification in groups like *Fromanviruses* and *Phietaviruses*. To display all possible false positive predictions, we generated a network of false-positive genera from the entire ICTV dataset and assigned them to putative families or subfamilies to facilitate potential verification in the future (Supplementary Figure S8).

ICTV version	**PR-AUC**	**95% CI**
ICTV 35	0.8804	(0.8811, 0.8988)
ICTV 36	0.9334	(0.9277, 0.9371)
ICTV 37	0.9689	(0.9660, 0.9716)
ICTV 38	0.9700	(0.9672, 0.9724)

Table 2. PR-AUC estimates with 95% confidence intervals for model 2 on different ICTV releases. Results were obtained on the test dataset.

We next compared the performance of model 2 (trained on release 38) on different ICTV releases, hypothesising that the genus prediction accuracy may improve with subsequent releases. Indeed, the data presented in Table 2 indicate a significant improvement in the model's performance between releases 35, 36, and 37/38, with no noticeable or significant differences between versions 37 and 38. This observation implies that while the expanding corpus of the Virus Metadata Resource inherently influences model assessment, the principal enhancements likely originate from the ICTV's continuous refinements and updates in taxonomic classification, which take ANI and AF into consideration.

Lifestyle	**PR-AUC**	**95% CI**
Temperate	0.9415	(0.9337, 0.9492)
Virulent	0.9896	(0.9863, 0.9922)

Table 3. PR-AUC estimates with 95% confidence intervals for model 2 on different phage lifestyles. Results were obtained on the test dataset.

Finally, we investigated performance of model 2 depending on the phage lifestyle (temperate or virulent). We found that virulent phage pairs yielded a significantly stronger predictive outcomes of model 2 compared to temperate pairs, as illustrated in Figure 6D and Table 3. This variance may not only reflect the compositional bias of our dataset—temperate phages being less represented—but may also hint at the intrinsic complexity of taxonomic classification amongst temperate phages. The latter are more likely to undergo horizontal gene transfer (HGT) than virulent phages, leading to increased genetic mosaicism[60,61]. Combining this information with the false positive dataset (Supplementary Figure S6), we can clearly observe genera of temperate phages that belong to the same subfamily cluster together, such as members of the subfamily *Hendrixvirinae* and genera related to the genus *Lambdavirus*. Mosaicism in the "Lambda supercluster" has been well documented, with different genus pairs sharing different sections of the genome[62]. This module shuffling makes it more challenging to define clear taxonomic boundaries to use as demarcation criteria, and poses a sophisticated challenge in using ANI and AF to predict genus classification. Selecting a set of hallmark genes as the unit of evolutionary information is the current gold standard in virus taxonomy to overcome this issue[27,59].

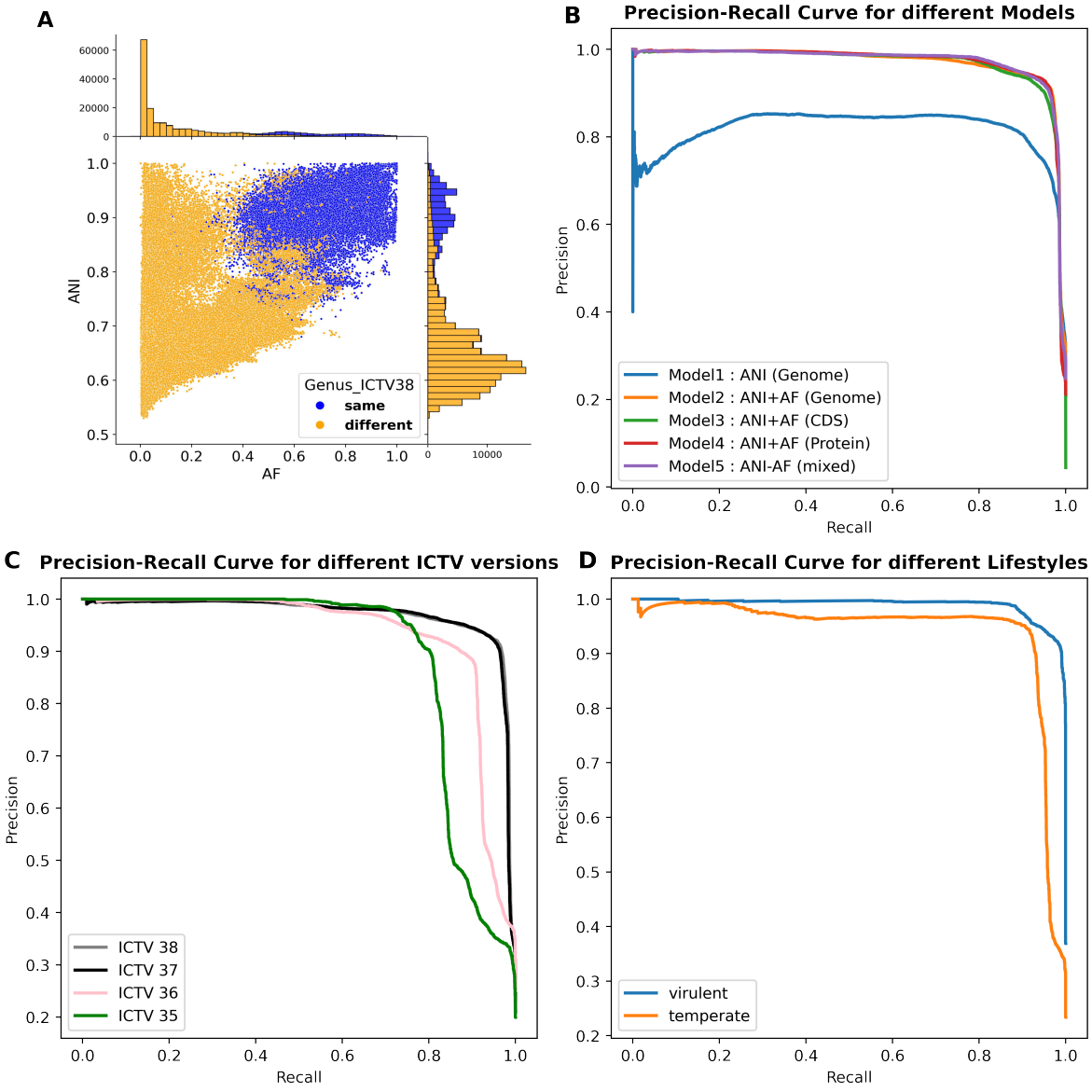

Figure 6. ANI (Y-axis) vs. AF (X-axis) for 745,209 pairs from the ICTV dataset. (A) Points are colored based on genus classification: same genus (blue), different genus (yellow). (B) Precision-Recall curves for five models evaluated on the test dataset: Model 1 (ANI in Genome Mode, blue), Model 2 (ANI+AF in Genome Mode, orange), Model 3 (ANI+AF in CDS Mode, green), Model 4 (AAI+AF in Protein mode, red), Model 5 (ANI in Genome Mode + AF in Protein Mode, purple). (C) Precision-Recall curves of Model 2 across different ICTV releases: 35 (orange), 36 (green), 37 (red), and 38 (blue). (D) Precision-Recall curves for Model 2, stratified by the predicted phage lifestyle: virulent (blue) and temperate (orange). PR-AUC values for curves in panels B, C and D are provided in Tables 1, 2 and 3, respectively.

Discussion

In this study we have introduced MANIAC: an approach for accurate and rapid calculation of Average Nucleotide Identity (ANI) and Alignment Fraction (AF) between pairs of viral genomes. MANIAC demonstrates a near-perfect correlation with the gold-standard ANIb in estimating ANI and AF, highlighting its exceptional precision in genomic analysis. When compared to other tools like fastANI and Mash, MANIAC provided more precise estimations, especially for genomes with ANI below 80%, while maintaining scalability when applying to millions of pairs of viral genomes. Simulated datasets were then used to optimise MANIAC, predicting that with the right parametrisation the tool accurately predicts genetic divergence for ANI values of approximately 60-70% and even below, a hypothesis supported with the use of real genomic data and the comparison with the average amino-acid identity (AAI) metric. This achievement underscores the tool's potential in enhancing the standardisation of taxonomic methodologies and resolving classification conflicts within the burgeoning field of virology.

One important application of having an accurate and scalable method to estimate ANI and AF between pairs of viral genomes was to investigate the existence of the 'ANI gap' in phage populations. The existence and implications of such a gap have been widely discussed in bacterial genomics[7,22]. The current consensus is that the ANI gap is not just an artefact of sampling and most likely reflects a true species discontinuity in bacteria, with species demarcation threshold of around 95%. However, the concept of a species is much more difficult to define in viruses due to their high mutation rates, absence of universal genetic markers, frequent recombination, dependence on host cells and diverse host ranges that blur traditional species boundaries[63]. With all of this in mind, our analysis points to the existence of an 'ANI gap' in phage populations around 78%-85%, akin to the one present in bacteria. We found that the bimodality of the ANI distribution is not solely driven by recently emerged mosaicism of phages, nor that it is a mere result of sampling (although it will inadvertently bias the shape of the distribution). Nevertheless, there are important differences in the observed distribution compared to the one found in bacteria, including its wider range, a downward shift of the 'ANI gap' with a noticeable continuity of thousands of pairs of genomes found in that range, as well as rampant genetic mosaicism. These differences are unsurprising given the rapid evolution of bacterial viruses, permissive thresholds for homologous recombination and frequent horizontal gene transfer, all of which suggests that strict ANI thresholds my not be the best approach to define cohesive viral populations, particularly in temperate

phages[63]. Hence, while our study finds support for the existence of bimodality in ANI distribution in phages, it should not be necessarily interpreted as evidence for the existence of sequence-discrete species in dsDNA bacterial viruses, or in any viruses for that matter.

The second important application of this study was to inquire about the usefulness in applying ANI and AF metrics for taxonomic classification of newly identified viral genera. Our prediction approach using logistic regression shows that while ANI alone is a suboptimal predictor of a viral genus due to significant genetic mosaicism, ANI and AF together work very well, achieving a 99% precision with the recall of 42% or 95% precision with the recall of 88%. Interestingly, false positives were limited to a relatively small number of genera: at 95% probability threshold it was 59 genera out of 1373 genera in the ICTV dataset, and at 99% probability threshold there were 14 unique genera, most of which belonged to the *Strabovdiridae* family, suggesting a potential incorrect taxonomy. *Fromanvirus* was also the main genus for which the model predicted false-negative assignments (i.e., genome pairs predicted as different genera). While this genus is one of the most diverse in the ICTV dataset ($n = 39$ genomes), for other genera there was a much stronger accordance between the ANI/AF distribution and the genus assignment, highlighting the complexity of taxonomic assignment in viruses, particularly in highly mosaic genomes of temperate phages.

Notably, the optimal thresholds for taxonomic assignments inferred here are mostly suited for dsDNA viruses, most of which are tailed phages. Such thresholds could potentially vary within different viral sub-populations, depending on the underlying biological properties like the rate of gene flow[63] or genome composition[27]. In any case, MANIAC offers the most accurate to-date tool to investigate the relationship between taxonomic demarcation and genetic relatedness in bacterial viruses.

Comparatively, MANIAC offers a superior alternative to other ANI-estimation tools, merging computational efficiency with unmatched accuracy. While tools like fastANI and Mash provide valuable insights, they tend to under- or overestimate ANI for distantly related genome pairs, a limitation MANIAC notably circumvents. One limitation of MANIAC is its substantial memory requirements when processing datasets comprising hundreds of millions of viral genome pairs. This issue arises from the necessity to amalgamate extensive alignment tables for ANI and AF computation. Given that MANIAC's accuracy in identifying closely related genomes is comparable to that of FastANI for such genomes, FastANI might be the preferred choice for projects aimed at identifying nearly identical genome pairs, such as data dereplication tasks. However, for investigations focused on elucidating more distant genomic relationships —

such as identifying genera that might be candidates for consolidation or detecting signs of recent genetic mosaicism — MANIAC emerges as the superior tool.

In conclusion, MANIAC offers a valuable advancement in the genomic analysis of phages, providing a reliable framework for taxonomic classification that balances accuracy with computational efficiency. On a laptop with Apple M1 chip with 8 cores, 16GB of RAM memory and SSD hard drive, running MANIAC in the Genome mode on the RefSeq dataset ($n = 4655$ genomes) took us 2 hours and 50 minutes. As a useful tool in virological research, its further development, informed by ongoing improvements in methodology and the broadening of datasets, will be crucial for fully realising its capabilities in capturing the extensive diversity of viruses.

Funding

This work was financed by the Polish National Agency for Academic Exchange (JK, RM), the Polish National Science Centre OPUS Grant (grant 2020/37/B/NZ8/03492; WN, RM), the EMBO Installation Grant (WN, JK, JL, RM), the Polish National Science Centre Sonata Bis Grant (grant 2020/38/E/NZ8/00432; JL, RM) and the International Visegrad Fund scholarship (JH). LC acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC). This UK funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking. EMA gratefully acknowledges funding by the Biotechnology and Biological Sciences Research Council (BBSRC); this research was funded by the BBSRC Institute Strategic Programme Microbes and Food Safety BB/X011011/1 and its constituent projects, BBS/E/F/000PR13634, BBS/E/F/000PR13635 and BBS/E/F/000PR13636, and the BBSRC Institute Strategic Programme Food Microbiome and Health BB/X011054/1 and its constituent project BBS/E/F/000PR13631. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: CI TASK, ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016857.

Author contributions

The following contributions can be distinguished by different authors based on the CRediT (Contributor Roles Taxonomy) system:

- **Conceptualization**: JH, RM
- **Methodology**: WN, JH, LC, RM
- **Software**: WN, JK, JL
- **Validation**: JK, JL, RM
- **Formal Analysis**: WN, JH, EA, RM
- **Investigation**: WN, JH, LC, EA, RM
- **Resources**: RM
- **Data Curation**: WN, JL, RM
- **Writing – Original Draft**: WN, JH, RM
- **Writing – Review & Editing**: WN, JH, JK, JL, LC, EA, RM
- **Visualization**: WN, JH

- **Supervision**: RM
- **Project Administration**: RM
- **Funding Acquisition**: RM

References

[1] Goris J, Konstantinidis K, Klappenbach J, Coenye T, Vandamme P, Tiedje J. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 2007;57:81–91.

[2] Konstantinidis K, Tiedje J. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 2005;102:2567–72.

[3] Gosselin S, Fullmer M, Feng Y, Gogarten J. Improving Phylogenies Based on Average Nucleotide Identity, Incorporating Saturation Correction and Nonparametric Bootstrap Support. Syst Biol. 2022;71:396–409.

[4] Murray C, Gao Y, Wu M. Re-evaluating the evidence for a universal genetic boundary among microbial species. Nat Commun. 2021;12:4059.

[5] Parks D, Chuvochina M, Chaumeil P, Rinke C, Mussig A, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38:1079–1086.

[6] Palmer M, Steenkamp E, Blom J, Hedlund B, Venter S. All ANIs are not created equal: implications for prokaryotic species boundaries and integration of ANIs into polyphasic taxonomy. Int J Syst Evol Microbiol. 2020;70:2937–2948.

[7] Jain C, Rodriguez-R L, Phillippy A, Konstantinidis K, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.

[8] Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K, Turner S, et al. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Int J Syst Evol Microbiol. 2018;68:2386–2392.

[9] Varghese N, Mukherjee S, Ivanova N, Konstantinidis K, Mavrommatis K, Kyrpides N, et al. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 2015;43:6761–71.

[10] Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106:19126–31.

[11] Konstantinidis K, Ramette A, Tiedje J. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 2006;361:1929–40.

[12] Sheinman M, Arkhipova K, Arndt P, Dutilh B, Hermsen R, Massip F. Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain. Elife. 2021;10.

[13] Li X, Tong W, Wang L, Rahman S, Wei G, Tao S. A Novel Strategy for Detecting Recent Horizontal Gene Transfer and Its Application to Rhizobium Strains. Front Microbiol. 2018;9:973.

[14] Ondov B, Treangen T, Melsted P, Mallonee A, Bergman N, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.

[15] Olm M, Brown C, Brooks B, Banfield J. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 2017;11:2864–2868.

[16] Koslicki D, White S, Ma C, Novikov A. YACHT: an ANI-based statistical test to detect microbial presence/absence in a metagenomic sample. bioRxiv. 2023 Apr;.

[17] Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

[18] Pritchard L, Glover R, Humphris S, Elphinstone J, Toth I. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Analytical Methods. 2016;8:12–24.

[19] Marçais G, Delcher A, Phillippy A, Coston R, Salzberg S, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14:e1005944.

[20] Deorowicz S, Gudys A, Dlugosz M, Kokot M, Danek A. Kmer-db: instant evolutionary distance estimation. Bioinformatics. 2019;35:133–136.

[21] Tian L, Mazloom R, Heath L, Vinatzer B. LINflow: a computational pipeline that combines an alignment-free with an alignment-based method to accelerate generation of similarity matrices for prokaryotic genomes. PeerJ. 2021;9:e10906.

[22] Rodriguez-R L, Jain C, Conrad R, Aluru S, Konstantinidis K. Reply to: Re-evaluating the evidence for a universal genetic boundary among microbial species. Nat Commun. 2021;12:4060.

[23] Roux S, Adriaenssens E, Dutilh B, Koonin E, Kropinski A, Krupovic M, et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). Nat Biotechnol. 2019;37:29–37.

[24] Camarillo-Guerrero L, Almeida A, Rangel-Pineros G, Finn R, Lawley T. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184:1098–1109.e9.

[25] Nayfach S, Camargo A, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides N. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol. 2021;39:578–585.

[26] Meier-Kolthoff J, Göker M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics. 2017;33:3396–3404.

[27] Turner D, Kropinski A, Adriaenssens E. A Roadmap for Genome-Based Phage Taxonomy. Viruses. 2021;13.

[28] Dong Y, Zheng K, Zou X, Liang Y, Liu Y, Li X, et al. Characterization and Genomic Analysis of the First Podophage Infecting Shewanella, Representing a Novel Viral Cluster. Front Microbiol. 2022;13:853973.

[29] Neri U, Wolf Y, Roux S, Camargo A, Lee B, Kazlauskas D, et al. Expansion of the global RNA virome reveals diverse clades of bacteriophages. Cell. 2022;185:4023–4037.e18.

[30] Mavrich T, Hatfull G. Bacteriophage evolution differs by host, lifestyle and genome. Nat Microbiol. 2017;2:17112.

[31] Accetto T, Janež N. The lytic Myoviridae of Enterobacteriaceae form tight recombining assemblages separated by discontinuities in genome average nucleotide identity and lateral gene flow. Microb Genom. 2018;4.

[32] Deng Z, Xia X, Deng Y, Zhao M, Gu C, Geng Y, et al. ANI analysis of poxvirus genomes reveals its potential application to viral species rank demarcation. Virus Evol. 2022;8:veac031.

[33] Hendrix R, Smith M, Burns R, Ford M, Hatfull G. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci U S A. 1999;96:2192–2197.

[34] Peck K, Lauring A. Complexities of Viral Mutation Rates. J Virol. 2018;92.

[35] Mahmoudabadi G, Phillips R. A comprehensive and quantitative exploration of thousands of viral genomes. Elife. 2018;7.

[36] Bao Y, Kapustin Y, Tatusova T. Virus Classification by Pairwise Sequence Comparison (PASC). Encyclopedia of Virology. 2008;p. 342–348.

[37] Lauber C, Gorbalenya A. Partitioning the genetic diversity of a virus family: approach and evaluation through a case study of picornaviruses. J Virol. 2012;86:3890–3904.

[38] Agren J, Sundström A, Håfström T, Segerman B. Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups. PLoS One. 2012;7:e39107.

[39] Muhire B, Varsani A, Martin D. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One. 2014;9:e108277.

[40] Moraru C, Varsani A, Kropinski A. VIRIDIC-A Novel Tool to Calculate the Intergenomic Similarities of Prokaryote-Infecting Viruses. Viruses. 2020;12.

[41] Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35:1026–1028.

[42] Köster J, Rahmann S. Snakemake–a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2522.

[43] Kristensen D, Wolf Y, Mushegian A, Koonin E. Computational methods for Gene Orthology inference. Brief Bioinform. 2011;12:379–391.

[44] Wolf Y, Koonin E. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol Evol. 2012;4:1286–1294.

[45] Fan H, Ives A, Surget-Groba Y, Cannon C. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC Genomics. 2015;16:522.

[46] Jain C, Dilthey A, Koren S, Aluru S, Phillippy A. A Fast Approximate Algorithm for Mapping Long Reads to Large Reference Databases. J Comput Biol. 2018;25:766–779.

[47] Cook R, Brown N, Redgwell T, Rihtman B, Barnes M, Clokie M, et al. INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. Phage (New Rochelle). 2021;2:214–223.

[48] Hyatt D, Chen G, Locascio P, Land M, Larimer F, Hauser L. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

[49] Lefkowitz E, Dempsey D, Hendrickson R, Orton R, Siddell S, Smith D. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 2018;46:D708–D717.

[50] Rambaut A, Grassly N. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 1997;13:235–238.

[51] Hockenberry A, Wilke C. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ. 2021;9:e11396.

[52] Eddy S. Where did the BLOSUM62 alignment score matrix come from? Nat Biotechnol. 2004;22:1035–1036.

[53] States DJ, Gish W, Altschul SF. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. Methods. 1991;3(1):66–70.

[54] Dutilh B, Varsani A, Tong Y, Simmonds P, Sabanadzovic S, Rubino L, et al. Perspective on taxonomic classification of uncultivated viruses. Curr Opin Virol. 2021;51:207–215.

[55] Rodriguez-R L, Conrad R, Viver T, Feistel D, Lindner B, Venter S, et al. An ANI gap within bacterial species that advances the definitions of intra-species units. mBio. 2024;15:e0269623.

[56] Konstantinidis K. Sequence-discrete species for prokaryotes and other microbes: A historical perspective and pending issues. mLife. 2023;2.

[57] Lawrence JG, Hatfull GF, Hendrix RW. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol. 2002 Sep;184(17):4891–4905.

[58] Aiewsakun P, Adriaenssens E, Lavigne R, Kropinski A, Simmonds P. Evaluation of the genomic diversity of viruses infecting bacteria, archaea and eukaryotes using a common bioinformatic platform: steps towards a unified taxonomy. J Gen Virol. 2018;99:1331–1343.

[59] Simmonds P, Adriaenssens E, Zerbini F, Abrescia N, Aiewsakun P, Alfenas-Zerbini P, et al. Four principles to establish a universal virus taxonomy. PLoS Biol. 2023;21:e3001922.

[60] Dion M, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. Nat Rev Microbiol. 2020;18:125–138.

[61] Moura de Sousa JA, Pfeifer E, Touchon M, Rocha EPC. Causes and Consequences of Bacteriophage Diversification via Genetic Exchanges across Lifestyles and Bacterial Taxa. Mol Biol Evol. 2021 May;38(6):2497–2512.

[62] Grose JH, Casjens SR. Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. Virology. 2014 Nov;468-470:421–443.

[63] Bobay L, Ochman H. Biological species in the viral world. Proc Natl Acad Sci U S A. 2018;115:6040–6045.