

## Genome analysis

# *i*DeLUCS: a deep learning interactive tool for alignment-free clustering of DNA sequences

Pablo Millan Arias <sup>1,\*</sup>, Kathleen A. Hill <sup>2</sup>, Lila Kari<sup>1</sup>

<sup>1</sup>Cheriton School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

<sup>2</sup>Department of Biology, University of Western Ontario, London, ON N6A 5B7, Canada

\*Corresponding author. Cheriton School of Computer Science, University of Waterloo, 200 University Avenue, West Waterloo, ON N2L 3G1, Canada.  
E-mail: pmillana@uwaterloo.ca (P.M.A.)

Associate Editor: Tobias Marschall

### Abstract

**Summary:** We present an *interactive* Deep Learning-based software tool for Unsupervised Clustering of DNA Sequences (*i*DeLUCS), that detects genomic signatures and uses them to cluster DNA sequences, without the need for sequence alignment or taxonomic identifiers. *i*DeLUCS is scalable and user-friendly: its graphical user interface, with support for hardware acceleration, allows the practitioner to fine-tune the different hyper-parameters involved in the training process without requiring extensive knowledge of deep learning. The performance of *i*DeLUCS was evaluated on a diverse set of datasets: several real genomic datasets from organisms in kingdoms Animalia, Protista, Fungi, Bacteria, and Archaea, three datasets of viral genomes, a dataset of simulated metagenomic reads from microbial genomes, and multiple datasets of synthetic DNA sequences. The performance of *i*DeLUCS was compared to that of two classical clustering algorithms (*k*-means++ and GMM) and two clustering algorithms specialized in DNA sequences (MeShClust v3.0 and DeLUCS), using both intrinsic cluster evaluation metrics and external evaluation metrics. In terms of unsupervised clustering accuracy, *i*DeLUCS outperforms the two classical algorithms by an average of ~20%, and the two specialized algorithms by an average of ~12%, on the datasets of real DNA sequences analyzed. Overall, our results indicate that *i*DeLUCS is a robust clustering method suitable for the clustering of large and diverse datasets of unlabeled DNA sequences.

**Availability and implementation:** *i*DeLUCS is available at <https://github.com/Kari-Genomics-Lab/iDeLUCS> under the terms of the MIT licence.

## 1 Introduction

Clustering algorithms for DNA sequences play a fundamental role in bioinformatics, as they can be used to study the structural composition of DNA sequence datasets, to discover novel operational taxonomic units, and to complement phylogenetic analysis. The development of high throughput sequencing technologies has raised several challenges to many clustering methodologies, as most of them cannot keep up with the exponential increase in the number of sequences available for analysis. One of the reasons is that many clustering methods rely on the computationally expensive process of sequence alignment. To address these limitations, several alignment-assisted and alignment-free methodologies were proposed, see, e.g. James *et al.* (2018) and Ghodsi *et al.* (2011). The majority of these methods also face scalability issues, as they are reliant on classic clustering algorithms that perform well in the low data regime but have poor performance when large amounts of data are available. While in the aforementioned approaches the exponential increase of data is a hindrance to good performance, other approaches, e.g. deep-learning-based methods, benefit from the availability of large amounts of data. In particular, multiple deep learning algorithms have been developed for classification and inference using both alignment-based and alignment-free methodologies (Tampuu *et al.* 2019, Vu *et al.* 2020, Nissen *et al.* 2021). It has also been shown recently by Millán Arias *et al.*

(2022) that deep learning provides a significant improvement over classical unsupervised learning algorithms in discovering genomic-signature-based clusters, at different taxonomic levels. These promising initial results motivated the development of *i*DeLUCS, which takes advantage of the capabilities of deep learning, and is capable of clustering datasets comprising more than 400 Mbp. In addition, *i*DeLUCS exhibits several novel features which enhance the interpretability of its results: confidence scores of the final cluster assignments, a graphical user interface (GUI), dynamic visualization of the underlying training process, and incorporated evaluation metrics.

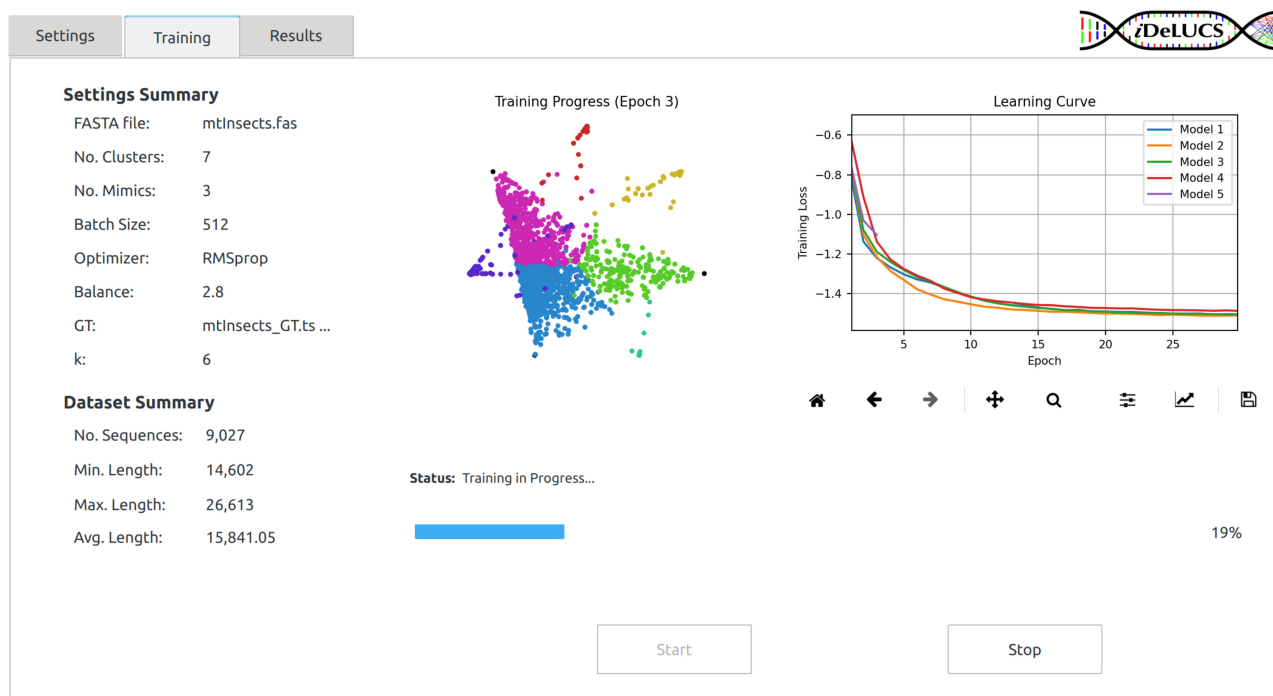
## 2 Software description

*i*DeLUCS is a standalone software tool that exploits the capabilities of deep learning to cluster genomic sequences. It is agnostic to the data source, making it suitable for genomic sequences taken from any organism in any kingdom of life. *i*DeLUCS assigns a cluster identifier to every DNA sequence present in a dataset, while incorporating several built-in visualization tools that provide insights into the underlying training process and the composition of the datasets (Fig. 1). *i*DeLUCS offers an evaluation mode to compare the ground-truth label assignments (or hypothesized label assignments) of the dataset sequences with their discovered cluster labels. This is accompanied by a visual qualitative assessment of the

Received: September 13, 2022. Revised: July 18, 2023. Editorial Decision: August 8, 2023. Accepted: August 16, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



**Figure 1.** Training tab of *iDeLUCS*. The left panel displays a summary of the main training parameters, as well as some statistics about the dataset under study. The center panel contains a qualitative assessment of the learning progress. In this particular case, the figure illustrates the clustering of mitochondrial genomes of insects into seven different clusters, each corresponding to one corner of the heptagon. Each point represents a genome, and its position indicates the probability that it is assigned to a different cluster/corner. The right panel contains a dynamic plot with the learning curves of the different models and serves as an indicator of whether or not the contrastive loss function is being minimized during training.

clustering, through the use of the uniform manifold approximation (UMAP, see [McInnes \*et al.\* 2018](#)) of the learned lower dimensional embedding. Finally, *iDeLUCS* outputs confidence scores for all of its cluster-label predictions, for enhanced interpretability. The software was developed using Python 3.9 and can be deployed with or without a graphics processing unit (GPU) (see [Supplementary Appendix SA](#) for implementation details).

### 3 Materials and methods

*iDeLUCS* builds upon the pipeline proposed in [Millán Arias \*et al.\* \(2022\)](#), consisting of: (i) calculating the  $k$ -mer frequencies for each DNA sequence, (ii) computing the data augmentations (mimic sequences), (iii) training multiple deep neural networks to learn the cluster assignments, and (iv) computing the majority voting cluster assignment for each sequence. In addition to multiple algorithmic optimizations to the pipeline, *iDeLUCS* significantly extends it in four main aspects. First, it uses the contrastive learning framework introduced by [Chen \*et al.\* \(2020\)](#), and incorporates an additional contrastive term in the loss function, which enforces the consistency of the hidden representations learned by the artificial neural networks. These hidden representations are learned simultaneously with the cluster assignments via backpropagation. Second, it replaces the majority voting scheme by a more robust clustering ensemble based on information theory, which reduces the variance and boosts the accuracy. Third, it uses the information provided by the ensemble and the consistency of the hidden representations to provide an intrinsic quantitative assessment of the clustering assignment (silhouette coefficient, Davies–Bouldin Index), as well as to output the confidence score for the cluster assignment of each sequence in the

dataset. Finally, the new contrastive learning framework can be combined with nonparametric clustering algorithms, such as HDBSCAN ([McInnes \*et al.\* 2017](#)), to automatically determine the number of clusters. This *iDeLUCS* option is recommended only for fine-grained clusterings, due to the fact that HDBSCAN is a density-based method (see [Supplementary Appendix SB](#) for details).

To assess the performance and applicability of *iDeLUCS*, we first analyzed 14 real datasets with known ground-truth annotations [described in (a) and (b)]: nine datasets of mitochondrial DNA from various Kingdoms of life (Animalia, Protista, Fungi) totaling 18 810 sequences, two Bacteria datasets totaling 4800 sequences, and three viral datasets totaling 4144 sequences. In addition, we analyzed one dataset of simulated reads from microbial genomes (Bacteria and Archaea) comprising 432 333 reads [described in (c)], and 12 synthetic datasets totaling 246 625 artificial DNA sequences [described in (d)]. Each dataset was selected for its unique characteristics, as described herein:

- Eight datasets from Kingdom Animalia, Kingdom Bacteria, and three datasets of viral sequences, obtained from Millán Arias *et al.* (2022): six mitochondrial DNA datasets of vertebrates at taxonomic levels from Subphylum to Family; two bacterial datasets to be clustered into families; and three viral datasets (Dengue, Influenza-A, Hepatitis B) to be clustered into virus subtypes. The maximum number of clusters per dataset is 12, and the maximum cluster size is 500 sequences, with the average sequence length of 16 700 bp for mtDNA, 433 882 bp for bacterial, and 5058 bp for viral sequences.*
- Three new mitochondrial DNA datasets (Table 1) created to enhance the representation across Kingdoms of*

**Table 1.** Description of the new mitochondrial DNA datasets (b).<sup>a</sup>

Dataset	Total no. sequences	Min. seq. length (bp)	Avg. seq. len. (bp)	Max. seq. len. (bp)	Total no. clusters	Cluster min. size	Cluster avg. size	Cluster max. size
Insects	9027	14 602	15 841	26 613	7	652	1290	1976
Fungi	1759	20 063	62 644	99 976	3	335	586	889
Protists	2581	5493	17 141	69 503	3	315	860	1642
Insects—balanced	4550	14 602	15 897	25 011	7	650	650	650
Fungi—balanced	1005	21 684	60 657	99 976	3	335	335	335
Protists—balanced	945	5498	24 697	69 503	3	315	315	315

<sup>a</sup> Note that there is a balanced version of each new dataset (Fungi, Protists, Insects). For the balanced version, the number of sequences per cluster was selected according to the number of sequences available in the smallest cluster.

life: a dataset of 2581 mitochondrial genomes from Kingdom Protista (average sequence length 17 141 bp) clustered into three phyla/subphyla; a dataset of 9027 mitochondrial genomes from class Insecta (average sequence length 15 841 bp) clustered into seven orders; and a dataset of 1759 mitochondrial genomes from Kingdom Fungi (average sequence length 62 644 bp), clustered into three phyla/subphyla.

- c) *One dataset of simulated metagenomic reads from eight microbial genomes, obtained from Wickramarachchi and Lin (2022)*. This dataset comprises more than 430 000 reads to be clustered into 8 species (7 Bacteria and one Archaea). The reads were simulated using the PacBio sequencing simulation parameters, with maximum cluster size of 119 330 sequences, and average sequence length of 8511 bp.
- d) *Twelve synthetic datasets from Girgis (2022)*. These are artificial datasets, each consisting of 100 random template sequences, representing the true clusters, and a random number of mutated copies that were generated from each template according to a predefined identity threshold. Each dataset contains at most 25 000 sequences, with a minimum dataset size of 18 210. The maximum number of clusters for each dataset is 12, the maximum cluster size is 400 sequences, and the average sequence length is 20 552 bp.

A detailed description of the datasets can be found in the [Supplementary Tables S1–S3](#) in [Supplementary Appendix SC](#).

To assess the performance of *iDeLUCS* on all the datasets analyzed in this study, we utilize both intrinsic (Davies–Bouldin Index, Silhouette Coefficient) and external [Homogeneity, Completeness, Unsupervised Clustering Accuracy (ACC)] clustering evaluation metrics. Of these, ACC arguably is the best indicator of performance, as it reflects the correspondence between cluster assignments and the ground-truth.

## 4 Applications and results

The performance of *iDeLUCS* was compared against two classic clustering algorithms, *k*-means++ and Gaussian Mixture Models (GMM), as well as two recent clustering methods specific to DNA sequences, DeLUCS (Millán Arias *et al.* 2022) and MeShClust v3.0 (Girgis 2022). The performance results for all four algorithms, in terms of intrinsic and external evaluation metrics as well as running time and memory usage, on the new mitochondrial datasets (b), are summarized in [Table 2](#). Note that MeShClust v.3.0 was run both with the recommended option of automatic identification of

the “identity threshold” parameter, as well as with the manually optimized value of the identity threshold parameter obtained by searching the interval [0.5, 0.9], with a step size of 0.05. The performance results for all algorithms on the other, previously published, datasets can be found in [Supplementary Tables S4–S7](#) in [Supplementary Appendix SC](#). Overall, *iDeLUCS* has a robust performance across these very different types of datasets: small (113 sequences) or large (432 000 reads); real, simulated, or synthetic; at different taxonomic levels ranging from phyla to subtypes; with balanced clusters or with unbalanced clusters; with cluster number varying from 3 to 100 clusters; comprising long sequences (500 000 bp) or short sequences (650 bp); consisting of homologous sequences or of nonhomologous sequences. On these datasets, the unsupervised clustering accuracy (ACC) obtained by *iDeLUCS* ranges from 78% to 100%, with an average accuracy of 90%.

In particular, *iDeLUCS* outperforms the other four clustering algorithms on the real datasets in (a) and (b), most of which consist of nonhomologous sequences. For example, for the mitochondrial genome datasets the average accuracy (ACC) of *iDeLUCS* is 92.4%, while classic clustering algorithms obtain an average accuracy of 70.2%, and the second-best performant algorithm has an average accuracy of only 75.42%.

As seen in [Table 2](#), for the new mitochondrial datasets (b), *iDeLUCS* obtains unsupervised clustering accuracies ranging from 78% to 89.7%. Specifically, *iDeLUCS* outperforms all the other clustering algorithms for the balanced and unbalanced versions of the Insects and Fungi datasets, and has a comparable performance with the other classifiers for both the balanced and unbalanced versions of the Protist dataset. Note that the improved clustering ensemble and the new contrastive loss function of *iDeLUCS* significantly enhance its capability to cluster unbalanced datasets, compared to DeLUCS. These improvements become apparent in the clustering of the dataset of simulated long metagenomic reads (c), where the accuracy of *iDeLUCS* is ~16% higher than that of DeLUCS, and ~7% higher than that of *k*-means++.

For the synthetic datasets in (d), *iDeLUCS* obtains an average accuracy of 98.5% when the number of clusters is given as a parameter, and of 97.3% when the option of using HDBSCAN is selected to automatically determine the number of clusters. This is slightly lower than, but comparable to, the performance of MeShClust v3.0, which achieves an average accuracy of 99.3% for these synthetic datasets. Note that not all synthetic datasets analyzed in [Girgis \(2022\)](#) were included in this comparison, since *iDeLUCS* was not optimized for some types of datasets. In particular, due to existing

**Table 2.** Comparison of the performance of *i*DeLUCS against *k*-means++, GMM, DeLUCS, and MeShClust v3.0 clustering algorithms on the new mtDNA datasets (b), using intrinsic cluster evaluation metrics (Davies–Bouldin Index, Silhouette Coefficient) and external evaluation metrics (homogeneity, completeness, unsupervised clustering accuracy ACC), as well as time and memory.<sup>a</sup>

Dataset	Model	DB (↓) Index	Silhouette (↑)	Homogeneity (↑)	Completeness (↑)	ACC (↑)	Time (↓)	Memory (↓) (GB)
Insects (9027 sequences)	k-means++	1.62	0.22	0.50	0.52	64.50%	<b>0:04:58</b>	<b>1.51</b>
	GMM	1.61	0.21	0.51	0.54	60.60%	1:16:15	4.57
	DeLUCS	1.87	0.25	0.64	0.63	72.80%	0:15:50	6.74
	MeshClust—auto	<b>0.44</b>	0.44	0.00	0.26	21.90%	1:06:13	1.81
	MeshClust—0.7	0.91	0.21	<b>0.98</b>	0.37	49.40%	0:25	
→	<b><i>i</i> DeLUCS</b>	3.75	<b>0.52</b>	0.78	<b>0.76</b>	<b>84.10%</b>	0:38:18	6.48
Fungi (1759 sequences)	k-means++	1.03	0.44	0.46	0.53	70.80%	<b>0:00:20</b>	<b>0.55</b>
	GMM	1.04	0.44	0.49	0.57	70.20%	0:18:30	1.94
	DeLUCS	<b>0.88</b>	<b>0.46</b>	0.50	0.48	63.50%	0:05:31	2.36
	MeshClust—auto	1.15	0.11	0.81	0.19	34.50%	0:47:35	1.70
	MeshClust—0.6	1.07	0.28	<b>0.96</b>	0.28	42.00%	0:13:33	1.15
→	<b><i>i</i> DeLUCS</b>	1.67	0.28	0.67	<b>0.63</b>	<b>78.00%</b>	0:04:32	2.28
Protists (2581 sequences)	k-means++	0.91	0.53	0.69	0.70	81.30%	<b>0:0:08</b>	0.61
	GMM	0.93	0.53	0.69	0.69	<b>81.90%</b>	0:07:40	2.00
	DeLUCS	1.16	0.51	0.54	0.45	62.20%	0:06:16	2.82
	MeshClust—auto	0.99	0.44	<b>1.00</b>	0.43	70.70%	0:04:16	<b>0.34</b>
	MeshClust—0.5	1.13	0.38	0.88	0.42	76.70%	0:01:40	0.23
→	<b><i>i</i> DeLUCS</b>	<b>0.39</b>	<b>0.81</b>	0.60	<b>0.79</b>	<b>81.40%</b>	0:13:56	2.77
Insects—balanced (4550 sequences)	k-means++	1.56	0.23	0.51	0.53	63.60%	<b>0:00:20</b>	0.89
	GMM	1.64	0.22	0.52	0.55	62.70%	0:55:40	3.90
	DeLUCS	1.33	0.37	0.67	0.68	78.30%	0:11:56	4.01
	MeshClust—auto	<b>0.43</b>	0.43	0.00	0.21	21.90%	1:28:35	1.81
	MeshClust—0.6	1.02	0.02	0.63	0.41	47.50%	0:03:20	<b>0.53</b>
→	<b><i>i</i> DeLUCS</b>	1.90	<b>0.57</b>	<b>0.82</b>	<b>0.83</b>	<b>89.70%</b>	0:19:11	3.90
Fungi—balanced (1005 sequences)	k-means++	<b>1.09</b>	0.37	0.38	0.43	59.50%	<b>0:00:12</b>	0.44
	GMM	<b>1.09</b>	0.37	0.40	0.46	60.10%	0:05:41	1.95
	DeLUCS	<b>1.09</b>	0.32	0.52	0.52	76.50%	0:04:47	1.91
	MeshClust—auto	1.14	0.11	<b>0.81</b>	0.18	34.45%	1:04:08	1.70
	MeshClust—0.5	1.41	−0.34	0.51	0.20	44.20%	0:01:44	<b>0.30</b>
→	<b><i>i</i> DeLUCS</b>	2.30	<b>0.39</b>	<b>0.81</b>	<b>0.81</b>	<b>86.20%</b>	0:12:12	2.17
Protists—balanced (945 sequences)	k-means++	0.97	0.42	0.55	0.62	74.10%	<b>0:0:08</b>	0.43
	GMM	1.00	0.41	0.56	0.62	74.50%	0:05:46	1.98
	DeLUCS	<b>0.85</b>	<b>0.53</b>	<b>0.70</b>	<b>0.70</b>	<b>88.10%</b>	0:07:42	1.85
	MeshClust—auto	1.20	0.09	0.94	0.51	74.92%	0:03:20	0.26
	MeshClust—0.7	1.02	0.06	0.97	0.44	71.85%	0:01:03	<b>0.23</b>
→	<b><i>i</i> DeLUCS</b>	<b>0.85</b>	<b>0.53</b>	<b>0.70</b>	<b>0.70</b>	<b>88.10%</b>	0:08:02	2.14

<sup>a</sup> Boldface indicates the best result, (↑)/(↓) indicate that higher/lower is better, “balanced” indicates the balanced version of the datasets. “MeShClust—auto” denotes MeshClust v3.0 run with the option of automatic identification of the identity threshold parameter, and “MeshClust—*p*” denotes MeshClust v3.0 run with a manually optimized identity threshold  $p \in [0.5, 0.9]$ .

restrictions on the size of the output layer in deep learning models, the synthetic datasets where the expected number of clusters was large (5000) were excluded. In addition, since the real genomic datasets analyzed in this study do not include short sequences (< 500 bp), the synthetic datasets with average sequence length < 500 bp were also excluded from this comparison. Future work is needed to systematically test and optimize *i*DeLUCS for datasets with short reads, or datasets where more than 200 clusters are expected.

All computational tests were performed on the Beluga cluster of the Digital Research Alliance of Canada (16 x Intel Gold 6148 Skylake @ 2.4 GHz CPU, 32 GB RAM) with NVIDIA V100SXM2 GPU (16 GB memory). The results obtained without GPU for hardware acceleration can be found in [Supplementary Table S8](#) in [Supplementary Appendix SC](#).

In summary, this study shows that *i*DeLUCS outperforms other algorithms in clustering sizeable datasets of unlabeled DNA sequences, especially when homology may or may not be present, and when the user has some prior knowledge of the expected number of clusters. Overall, our analysis shows that *i*DeLUCS is an accurate and scalable clustering method, performant on datasets of long, homology-free DNA

sequences, not tractable via alignment-based methods due to either lack of alignment or excessive time complexity.

## Acknowledgements

The authors thank Shane Ding for assistance with testing and software release, Daniel Olteanu, Joseph Butler, and Connor Holmes for software testing, and Zihao Wang for additional computational resources.

## Supplementary data

[Supplementary data](#) are available at *Bioinformatics* online.

## Conflict of interest

None declared.

## Funding

This work was supported by the Natural Sciences and Engineering Research Council of Canada [grant numbers

RGPIN-2023-03663 to L.K., R3511A12 to K.A.H.]; Compute Canada [grant number 616 to K.A.H.].

## References

- Chen T, Kornblith S, Norouzi M *et al.* A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning*, 1658–9. 2020.
- Ghods M, Liu B, Pop M *et al.* DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* 2011;12:271.
- Girgis HZ. MeShClust v3.0: high-quality clustering of DNA sequences using the mean shift algorithm and alignment-free identity scores. *BMC Genomics* 2022;23:423.
- James BT, Luczak BB, Girgis HZ *et al.* MeShClust: an intelligent tool for clustering DNA sequences. *Nucleic Acids Res* 2018;46:e83.
- McInnes L, Healy J, Astels S *et al.* HDBSCAN: hierarchical density based clustering. *JOSS* 2017;2:205.
- McInnes L, Healy J, Saul N *et al.* UMAP: uniform manifold approximation and projection. *JOSS* 2018;3:861.
- Millán Arias P, Alipour F, Hill KA *et al.* DeLUCS: deep learning for unsupervised clustering of DNA sequences. *PLoS One* 2022;17:e0261531.
- Nissen JN, Johansen J, Allesøe RL *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol* 2021;39:555–60.
- Tampuu A, Bzhalava Z, Dillner J *et al.* ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLoS One* 2019;14:e0222271.
- Vu D, Groenewald M, Verkley G *et al.* Convolutional neural networks improve fungal classification. *Sci Rep* 2020;10:12628.
- Wickramarachchi A, Lin Y. Binning long reads in metagenomics datasets using composition and coverage information. *Algorithms Mol Biol* 2022;17:14.