

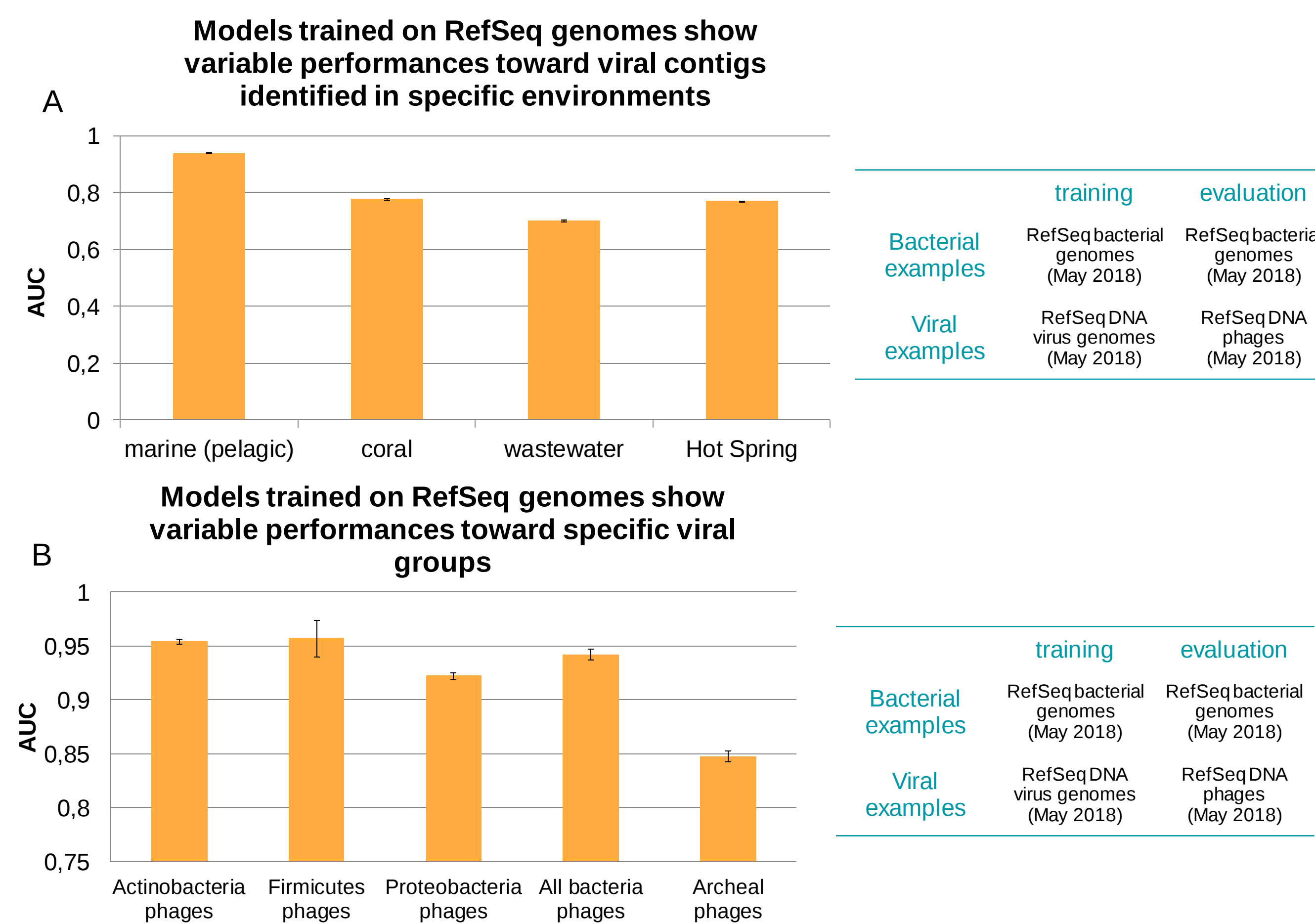
Ecologically informed viral sequence detection in aquatic metagenomes using machine learning

Introduction

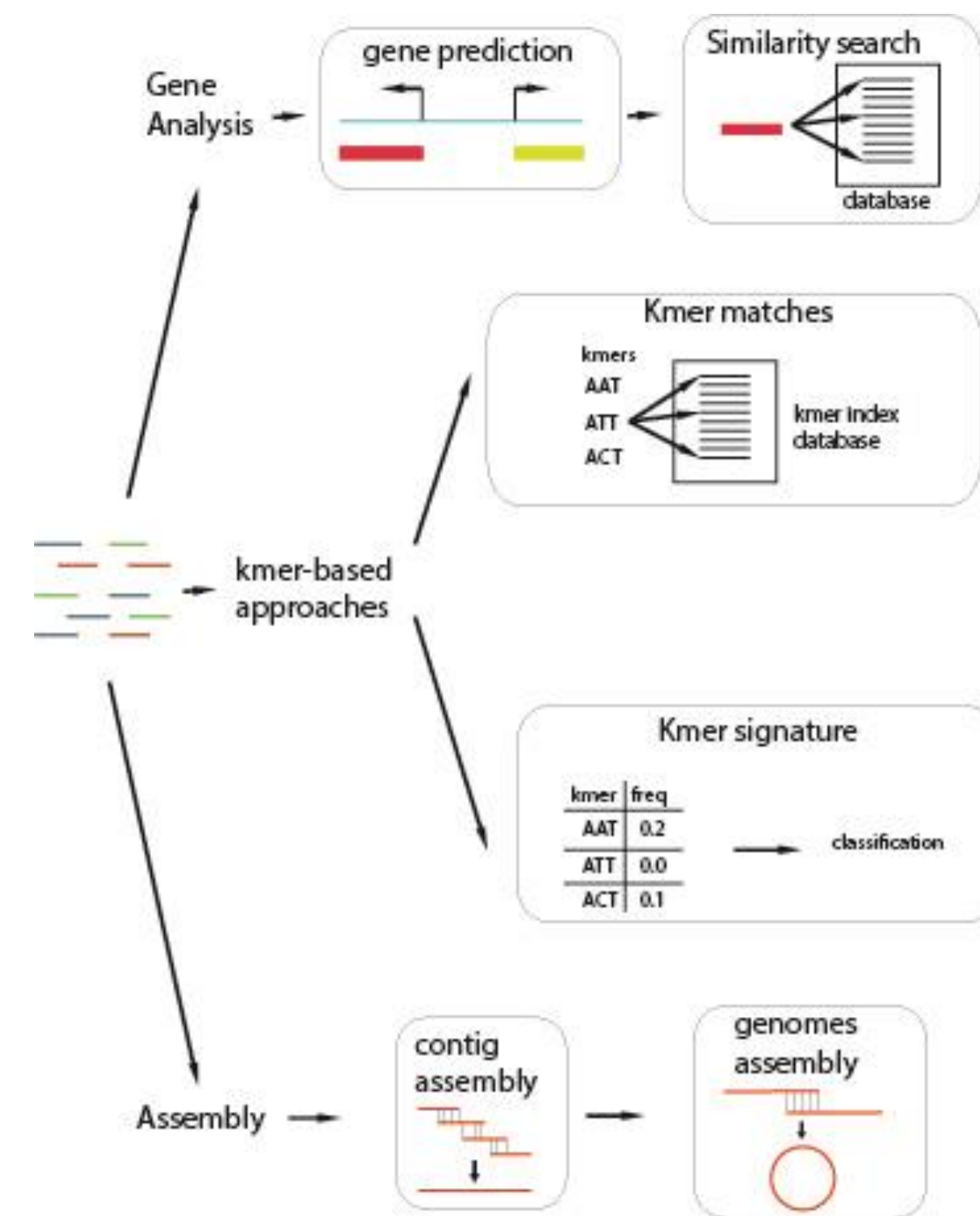
Modern 'omics allows for the exploration of in situ relationships between phages and their hosts, to provide new insights into the impact of viral populations on numerous biological systems. Retrieving viral sequences from bacterial metagenomes is critical to understanding host-viruses interactions in a given ecosystem. Homology to known viral genes is a primary method to retrieve viral contigs from complex metagenomes. However these approaches limit the discovery of novel viral sequences with no similarity to previously known viruses.

Recently, VirFinder, a new tool to detect viral sequences in bacterial metagenomes using a machine learning method, was released [Ren *et al.* 2017]. This method distinguishes viral from bacterial sequences based on their k-mer signatures, rather than through homology based searches to viral genes.

Current Limitations: VirFinder relies on a model trained on known viral and bacterial genomes from the RefSeq database, that shows a bias toward the detection of the most abundant viral groups in reference databases [Ren *et al.* 2017].

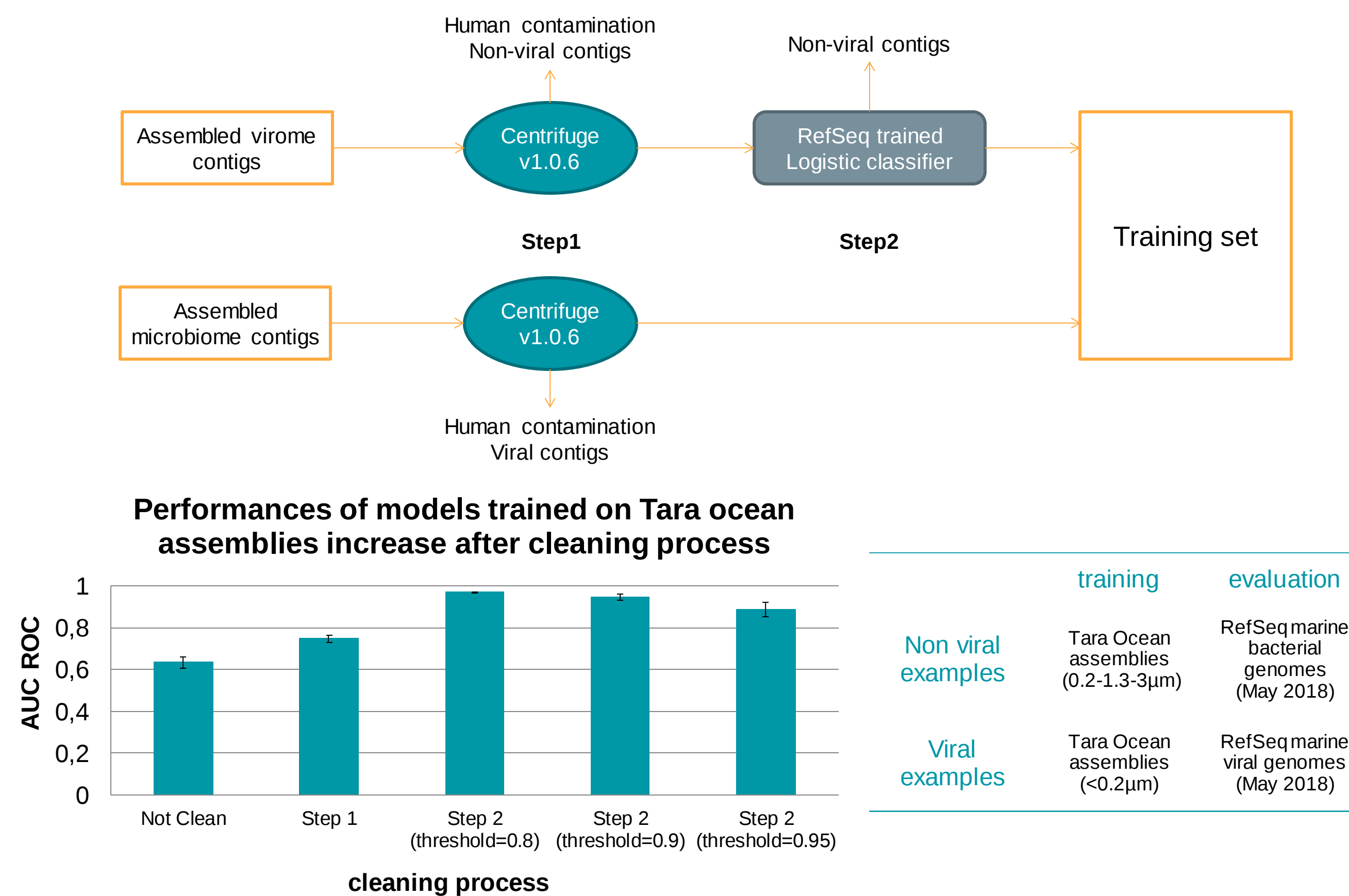


Viromes represent a large collection of viral sequences that are unbiased by cultivation methods and cover a wide variety of ecosystems. These sequences are a vast and interesting source of information about viral k-mer signatures. We present a **scalable computational framework** to train machine learning models directly on curated ecosystem-specific metagenomic contigs. This approach aims to ensure **reliable viral sequence detection even in ecosystems less studied**, where a smaller amount of phages have been previously isolated and sequenced, and provide the user with ecosystem specific predictions. Finally, our approach takes into account the possibility of **eukaryotic contamination** that is fundamental in various environments.



Creating marine-specific models

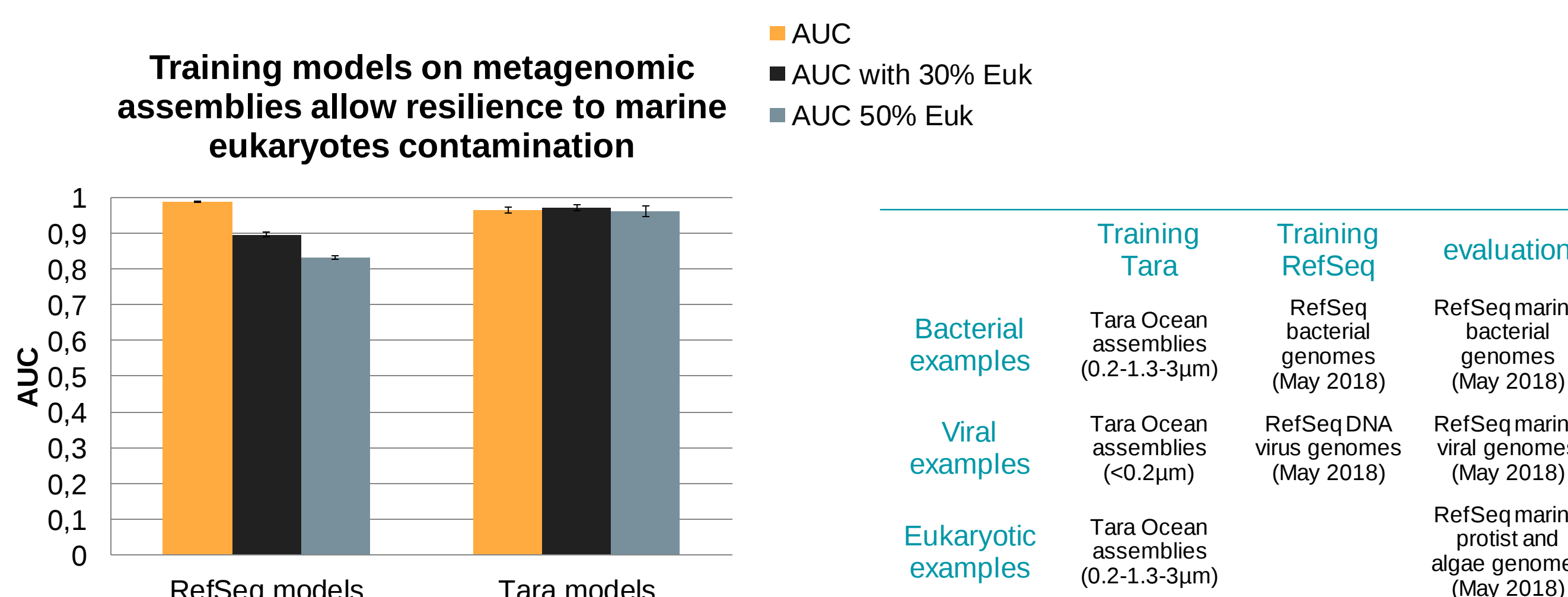
As a proof of concept, we provide the results obtained for a marine specific model, trained using the Tara Ocean assemblies dataset [Sunagawa *et al.* 2015]. Viromes and microbiomes always carry small amount of sequence contamination. However, in order to ensure good performance for the models, it is necessary to provide a clean training set.



Better resilience to contaminations

VirFinder training set is void of any Eukaryotic training examples. The tool is thus naïve to sequences of Eukaryotic origin, which can lead to an increased false positive rate in certain ecosystems where sequences from micro-eukaryotes are naturally present in metagenomes. Including protists and micro-eukaryotes genomes in a RefSeq trained model is limited by the low number of sequences currently available for those organisms.

Training classifiers using metagenomes allows the model to take into account micro-eukaryotes sequences, when they are present naturally in the ecosystems.



An ensemble method to improve detection of rare events

Even when using state-of-the-art tools, the detection of rare events often leads to a low precision : $TP/(TP+FP)$. This can be easily explained using the Bayes's theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

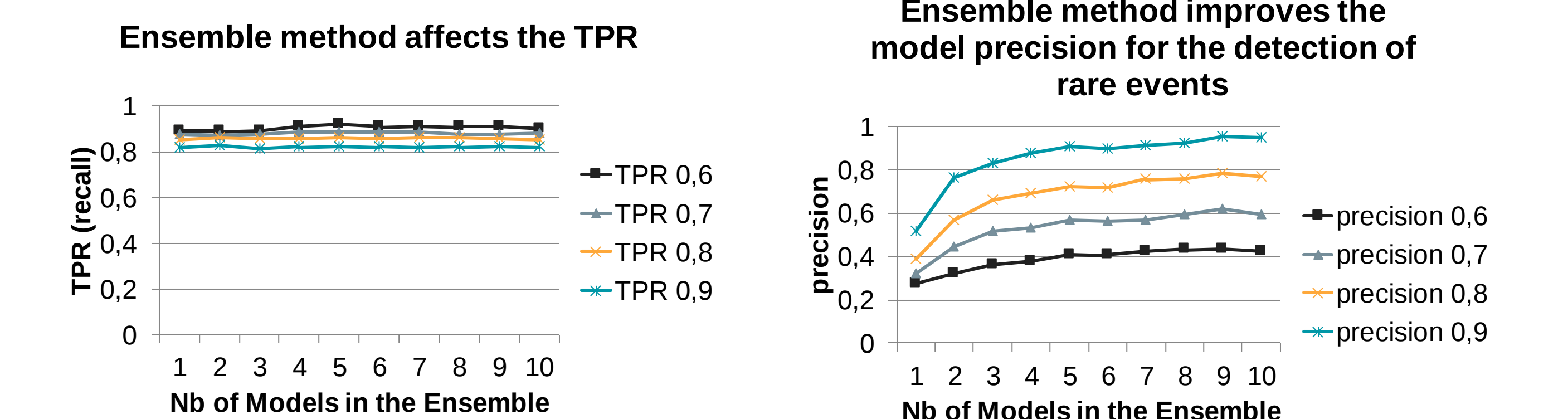
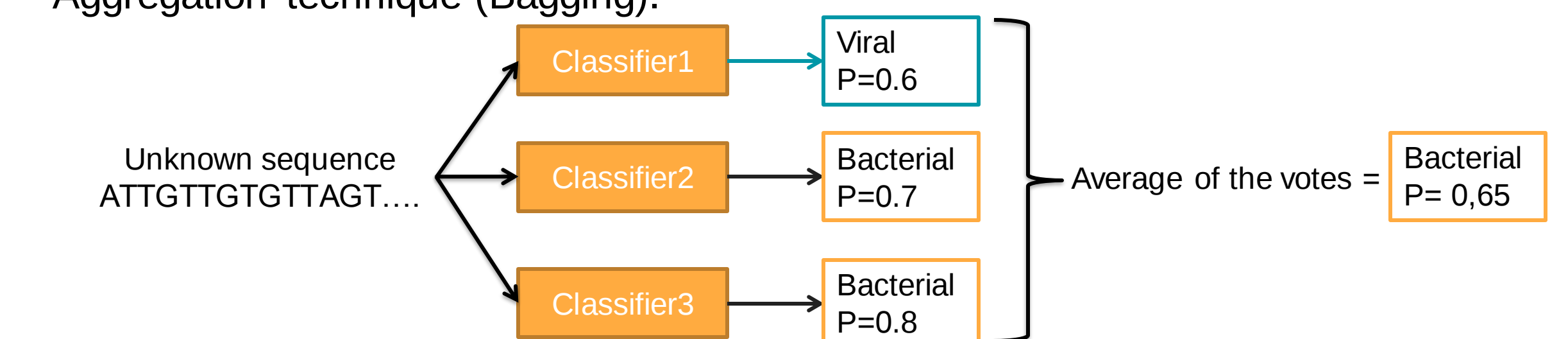
Eq 1 : Bayes's theorem formula

	expected		total
	viral	non-viral	
predicted	211	189	400
	39	4561	4600
total	250	4750	5000

Fig 7 : confusion matrix for a logistic classifier trained on Tara Ocean assemblies and evaluated on RefSeq genomes grinded at 5000bp (250 viral examples and 4740 bacterial examples)

Viral sequences in microbiomes constitute often less than 10% of the total sequences. Thus the use of a tool like VirFinder will lead to more false positives in most conditions.

It is possible to improve the precision obtained in such conditions using an ensemble approach [Ali and Pazzani, 1996]. The ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. A simplest but powerful ensemble is the Bootstrap Aggregation technique (Bagging).



Conclusions

- We present a framework allowing the training of ecosystem-specific models directly on curated metagenomes
- This approach allows a better resilience toward eukaryotic sequences in the studied ecosystem
- Training simple ecosystem-specific models allows for easy implementation of an ensemble method, that improves the detection of viral contigs in metagenomes.

References

- Ali, M., & Pazzani, M.J. (1996). **Error reduction through learning multiple descriptions**. *Machine Learning*, 24(3), 173–202.
- Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., Sun, F. (2017) **VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data**, *Microbiome*, 5, p. 69
- Sunagawa S. *et al* (2015) **Structure and function of the global ocean microbiome**, *Science*, 348 (6237)