DISCOVERING RNA VIRUSES
in Metatranscriptomes

Why find RNA viruses?

For viral surveillance:
Preventing epidemics before they emerge requires identifying new, unseen viruses before they the cross the species barrier.

To understand complex environments:
RNA viruses play key roles in microbial turnover and gene transfer in aquatic environments and likely influence terrestrial systems too. Many animal and plant pathogens ranging from rhinovirus to ebolavirus encode their genomes with RNA but RNA viruses have not been sampled well in other environments. Most metagenomic studies targeting viruses have focused on DNA viruses, even though by one estimate, half the viral particles in the ocean are RNA viruses (Steward et al. 2013).

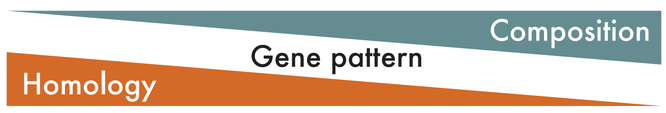
To identify new RNA phage:
Despite the fact that the first genome sequenced in 1976 was an RNA phage, only 10 more RNA phage genomes have been sequenced. There are only two known genera of RNA phage. Undiscovered RNA phage may play important roles in microbial communities.

For biotechnological applications:
RNA dependent RNA polymerases (RdRP) are not commonly used for in vitro RNA amplification because their error rates are high. Finding an RNA virus with a long genome or low diversity mutant spectrum could lead to the discovery of high-fidelity RdRPs that would allow a more direct and rapid method for the replication of RNA in vitro.

Methods

Feature selection is key
RNA viruses have high mutation rates and few identified environmental representatives. These factors make the identification of divergent RNA viruses difficult using *homology* based approaches. *Compositional* approaches like tetranucleotide frequency generalize better but have low information content. This is the variance bias tradeoff.

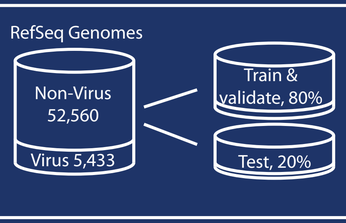
Gene pattern is a feature set representing the codon frequencies in the coding and two non-coding frames of each Open Reading Frame (ORF) and the non-coding regions. It is an intermediate complexity feature that also reduces noise by separating coding and non-coding regions. It is typically used for gene calling but we have applied it to classification.



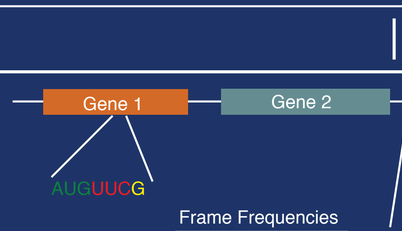
A machine learning approach

Steps (1) and (2) use **GeneLearn**
Experimental python software for learning from genomic data

(1) Select training data from reference genomes to split into training and test sets.



(2) Fragment genomes into 200 5KB segments, resulting in approximately 10 million training sequences.



(3) Feature selection is implemented on each sequence. Protein coding genes are identified using MetaGeneMark (Zhu et al. 2010). Frequencies are calculated for each codon in the coding and non-coding frames of the ORF and intergenic regions. This output results in a characteristic *gene pattern*.

Classification / training
Train a logistic classifier or multinomial logistic classifier:

$$f(\mathbf{w}) := \lambda R(\mathbf{w}) + \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{w}; \mathbf{x}_i, y_i)$$

Loss function: Logistic $\log\left(1 + e^{(y\mathbf{w}^T\mathbf{x})}\right)$

Regularization: L2 (Ringe) $\frac{1}{2} \parallel w \parallel_2^2$

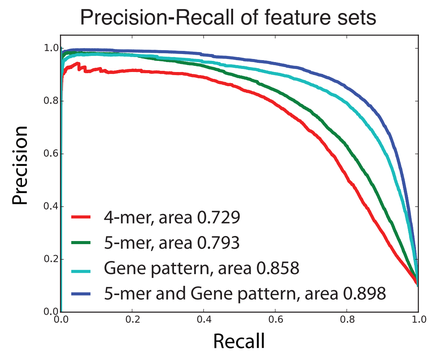
Optimization: L-BFGS

(4) Using *gene patterns* identified in previous step, a logistic classifier is trained using MLlib and Spark, resulting in a model for predicting viral sequences.

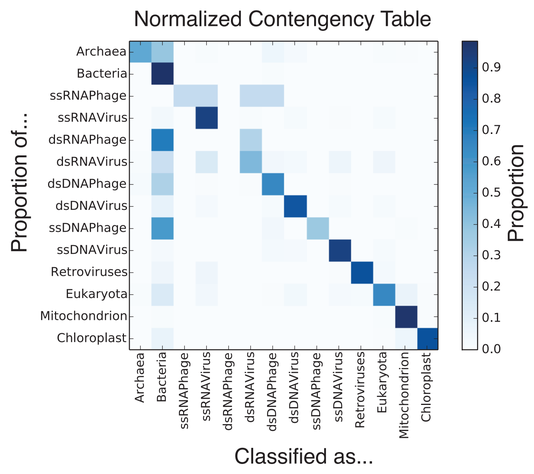
Spark MLlib
Scalable Machine learning framework (Meng et al. 2015)

Does it work?

Classification of virus fragments in genomes
When the precision and recall of logistic models from four different feature sets were compared, the full feature set identified viruses best. The classifier handled the imbalanced classes well. False positives often contained repeat motifs. Adding features that quantify repeat domains should minimize this.

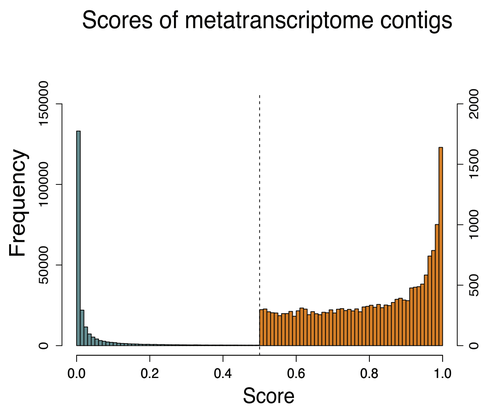


Multiclass classification
Sequences were classified into one of 14 categories using multinomial logistic regression. Performance was generally good, some viral classes with few training examples were misclassified at higher proportions.

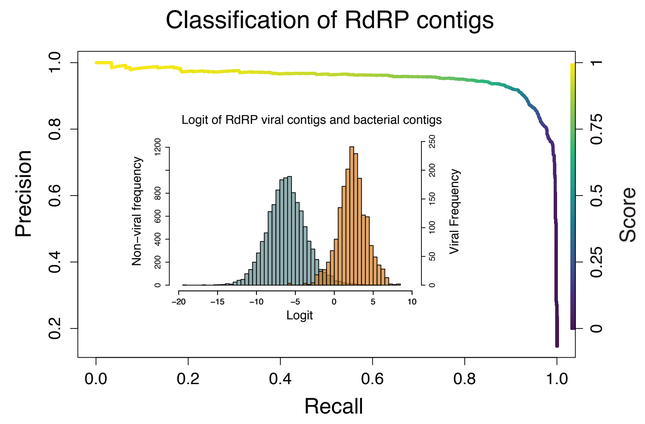


Validation with metatranscriptome data

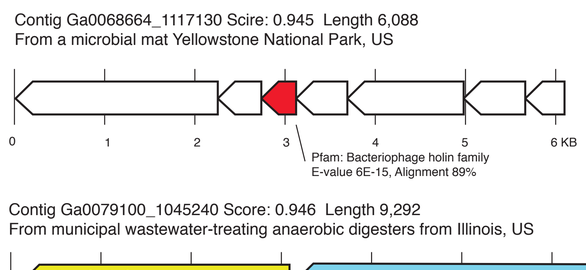
236,000 contigs longer than 4Kb were extracted from 560 public metatranscriptomes in the IMG database. The scores from the classifier show the distribution of classifications. Most reads are non-viral but about 2% of reads are classified as viral.

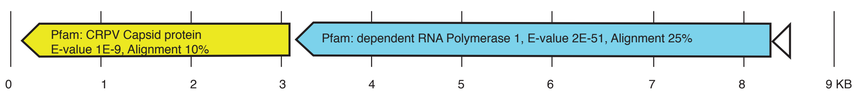


1865 contigs from metatransciptomes in IMG containing the viral marker gene RdRP (Eddy, 2011) were mixed with 10,833 randomly selected bacterial contigs and classified. Classification performance was similar to the performance on known genomes, indicating the model generalizes to real data.



Example contigs





92% of sequences classified as viral are viral — **Precision**

Recall — 91% of viruses are classified as viruses

Conclusions

Preparing RNA viral metagenomes is technically challenging but many RNA viruses are lurking in existing metatranscriptomes. Logistic classification can accurately identify the small RNA virus contigs at a precision and recall useful for identifying viruses in imbalanced metatranscriptome samples. The gene pattern can be combined with compositional information to accurately classify small RNA contigs as viral without relying on gene homology. This method is likely to generalize well to deeply divergent RNA viruses without homology to known viruses.

References

Besemer J, Lomsadze A, Borodovsky M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–18.

Eddy SR. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., et al. (2015). MLlib: Machine Learning in Apache Spark. preprint: arXiv:1505.06807 [cs.LG].

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. (2013). Are we missing half of the viruses in the ocean? ISME J 7:672–9.

Zhu W, Lomsadze A, Borodovsky M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 38:e132.

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

Adam R. Rivers
Qingpeng Zhang
Susannah G. Tringe
arrivers@lbl.gov • February 2016