Computational Methods for Virus-Host Interactions Prediction

Liuyang Cai

First-year PhD student @ Prof. Zigui Chen's lab

Joint Graduate Student Seminar

2019/12/11

香港中文大學
The Chinese University of Hong Kong

CU Medicine HONG KONG

香港中文大學醫學院
Faculty of Medicine
The Chinese University of Hong Kong

Content

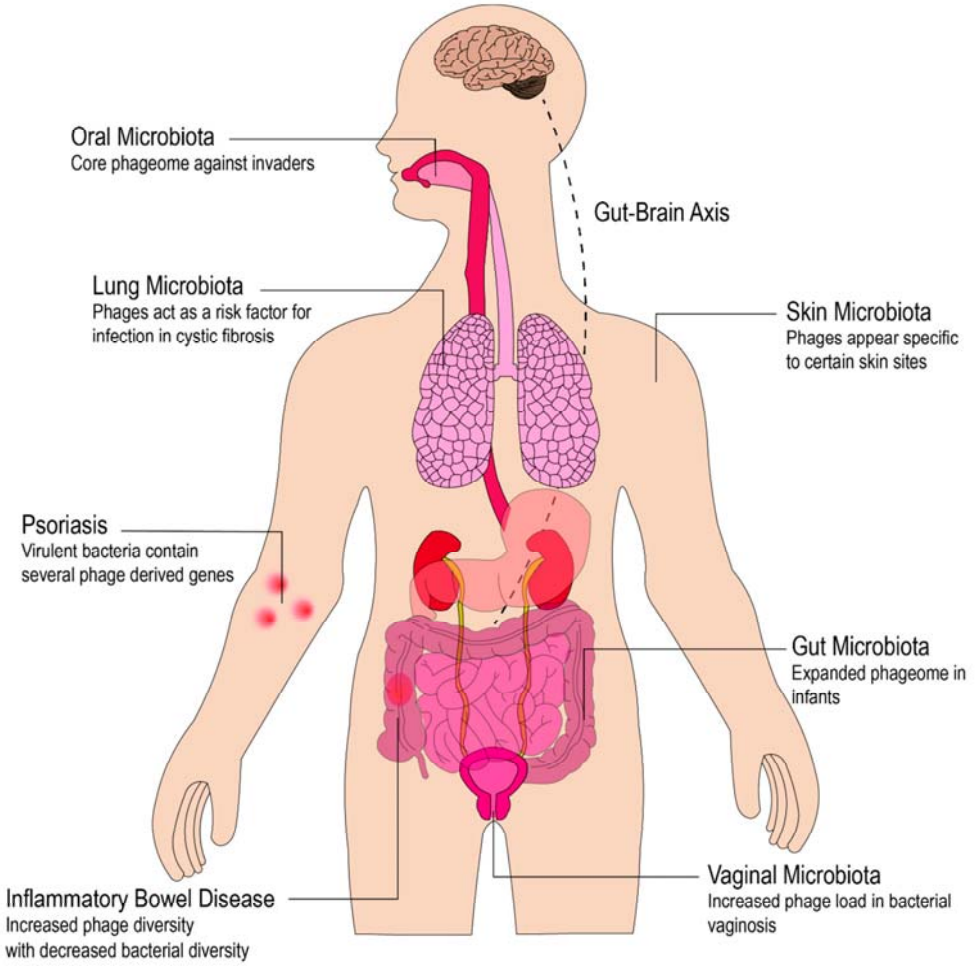
- Background
  - Bacteriophages and their clinical implications
  - Multi-levels phage-host interactions during phages' life cycle
  - Three computational strategies for phage-host interaction in DNA levels
- Software and algorithms implemented in the three strategies
- Comparisons of the above three strategies
- Summaries and prospective

Introduction of bacteriophage

- A bacteriophage (Phage) is a virus that infects and replicates within bacteria and archaea
- Phages are found throughout the biosphere, including in bodies of water and sewage, and in and on humans and animals, totaling an estimated 1x10^31 virions and outnumbering bacterial cells by ten-fold
- More than 6,000 phages have been characterized to date, with genome sizes 55 ranging from a few thousand to 480,000 nucleotides or more, most of them are as yet undescribed

Guttman B, et al. *Basic phage biology*. 2005
Comeau AM, et al. *Res Microbiol.* 2008

Clinical implications of phages



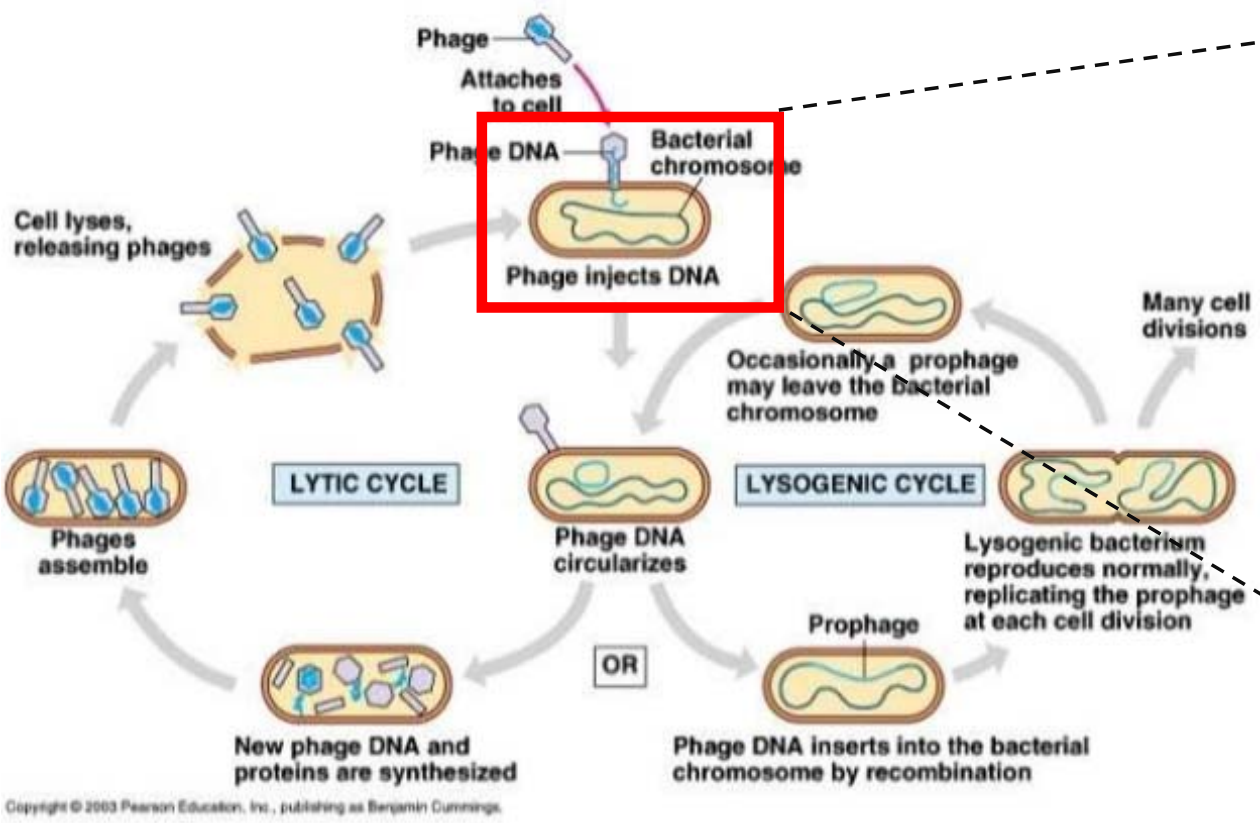
The use of phages to treat drug-resistant bacterial infections has stimulated interest in the use of phages to treat a variety of human diseases

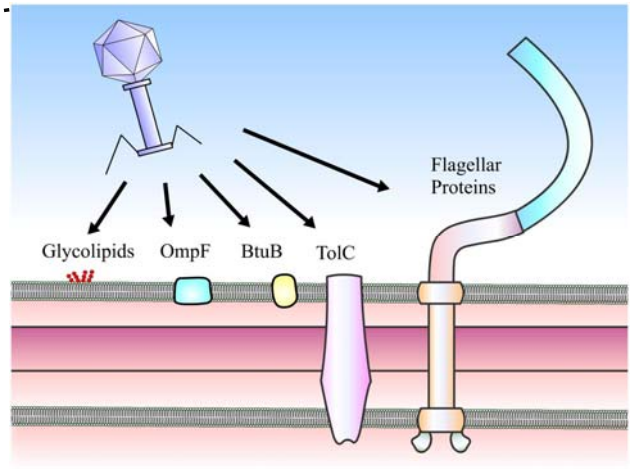
Phage lytic life cycle as the basis for conventional therapy. Strategies include using phage-derived enzymes and bioengineering of phages

Dylan Lawrence, et al. *Viruses*. 2019

Life cycle of bacteriophages contain multi-level phage-host interactions

Protein-protein

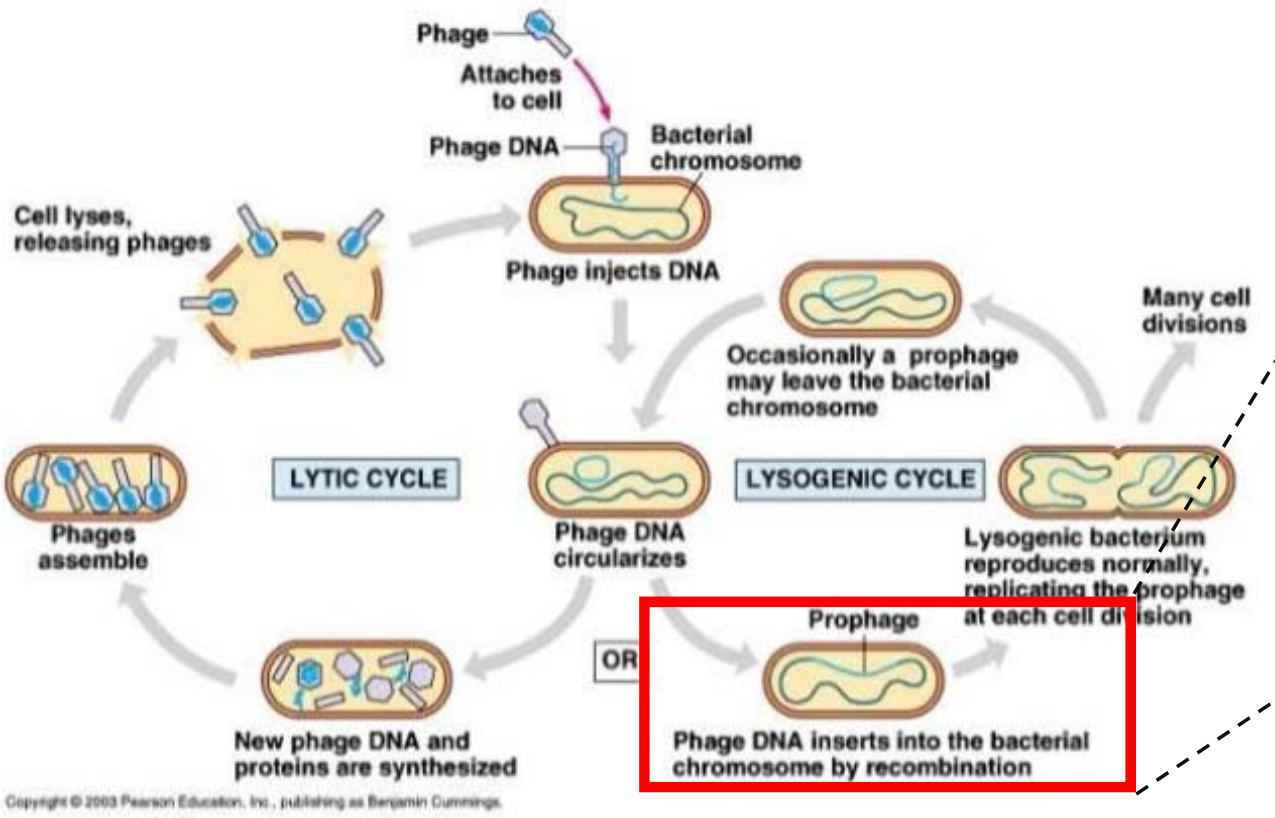




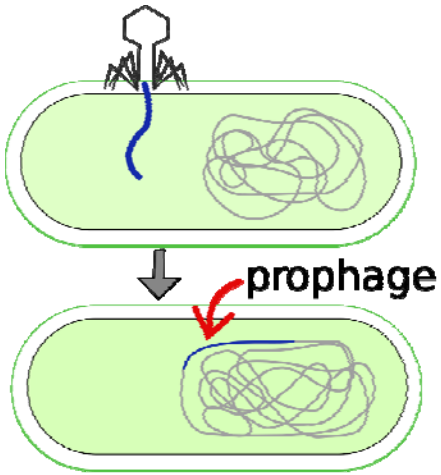
Tail proteins of these phages are capable of recognizing almost every host surface component including surface proteins, polysaccharides and lipopolysaccharides

Wenchen Song, et al. *Nucleic Acids Research. 2019*
Chaturongakul S. et al. *Front. Microbiol. 2014*
https://en.wikipedia.org/wiki/Prophage

Life cycle of bacteriophages contain multi-level phage-host interactions



DNA-DNA



Prophages are temperate phages integrated in bacterial genomes.

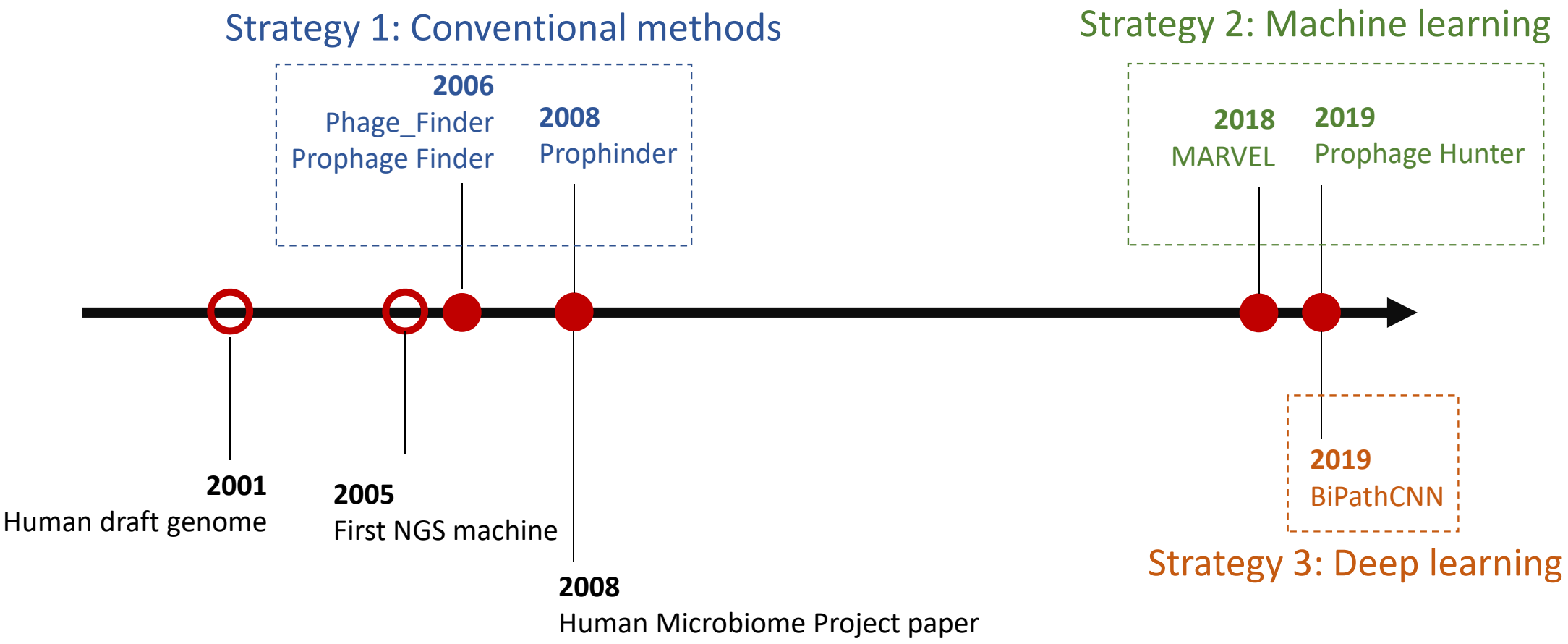
Wenchen Song, et al. *Nucleic Acids Research. 2019*
Chaturongakul S. et al. *Front. Microbiol. 2014*
https://en.wikipedia.org/wiki/Prophage

Prophages are important to the whole ecosystem

- Prophages are important agents of horizontal gene transfer
- Prophages can participate in a number of bacterial cellular processes, including antibiotic resistance, stress response, and virulence

Derrick E. Fouts. *Nucleic Acids Research*. 2006
Gipsi Lima-Mendez, et al. *Bioinformatics*, 2008

Three computational strategies for prophage identification

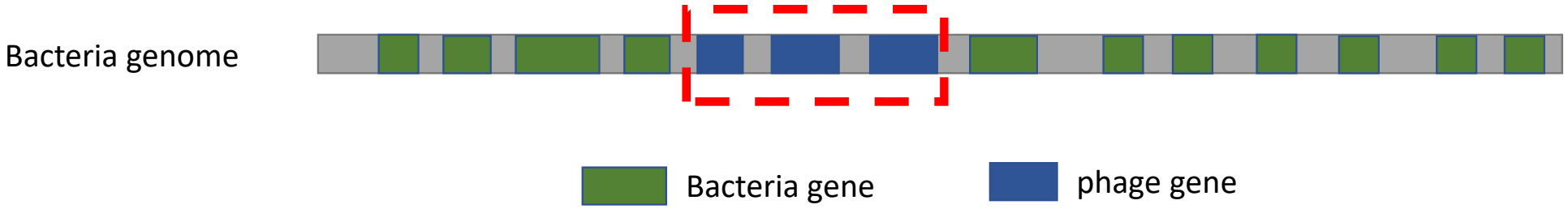


https://www.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-cost.html

Strategy 1: Identify clusters of phage-like genes within a bacterial genome

Software Name	Publishing year
Prophage Finder	2006
Phage_Finder	2006
Prophinder	2008

Identify clusters of phage-like genes within a bacterial genome



Detection of phage-like CDSs in the bacteria genome

- Blastp

Detecting phage-like gene clusters

- Sliding window

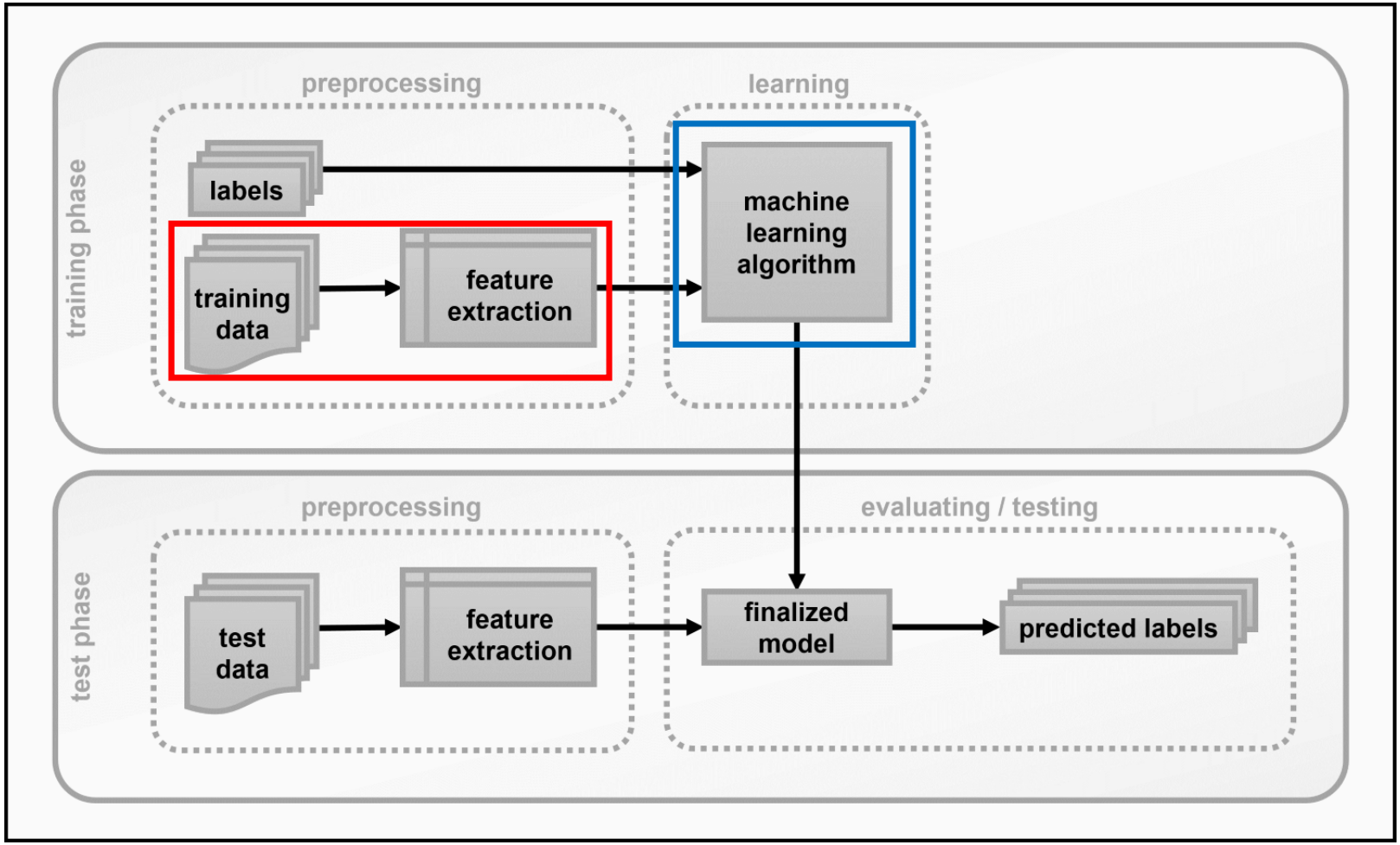
Selecting the putative prophages

- Statistical test

Strategy 2: Prophage identification is a CLASSIFICATION machine learning problem

Step 1

Step 2



https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781783980284/2/ch02lvl1sec14/understanding-the-machine-learning-workflow

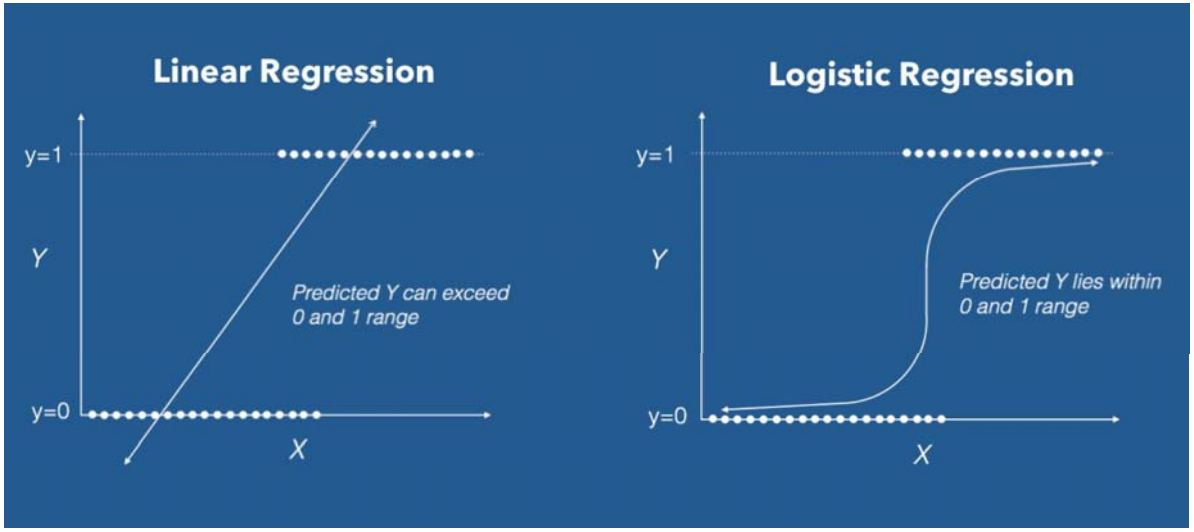
Selected features and implemented algorithms for MARVEL and PROPHAGE HUNTER

Software Name	Publishing Year	Features	Algorithms
MARVEL	2018	(1) Average gene length (2) Average spacing between genes, (3) Density of genes, (4) Frequency of strand shifts between neighboring genes, (5) ATG relative frequency, (6) Fraction of genes with significant hits against the pVOGs database	random forest
Prophage Hunter	2019	(1) Transcriptional orientation (2) protein length (3-22) Composition of 20 amino acids (23) Watson-Crick ratio (24) transcription strand switch	logistic regression

Training and testing Data preparation for prophage prediction

	Features					Labels
	average gene length	**average spacing between genes**	**density of genes**	**frequency of strand shifts between neighboring genes**	**…**	**Prophage**
Training set	1Kb	1.5Kb	0.5	0.5	…	Yes
	2Kb	2.8Kb	0.8	0.8	…	No
	3Kb	2Kb	0.9	0.9	…	Yes
Testing set	2.5Kb	5.6Kb	0.7	0.5	…	No
	3Kb	2.8Kb	0.6	0.4	…	Yes
	1.4Kb	2.6Kb	0.5	0.6	…	No

Introduction to Logistic Regression



Logistic regression equation :

Linear regression $Y = b_0+b_1\times X_1+b_2\times X_2.....+b_K\times X_K$

Sigmoid Function $P = \frac{1}{1+e^{-Y}}$

By putting Y in Sigmoid function, we get the following result.

$$\ln\left(\frac{P}{1-P}\right) = b_0+b_1\times X_1+b_2\times X_2.....+b_K\times X_K$$

Each X is a feature

https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

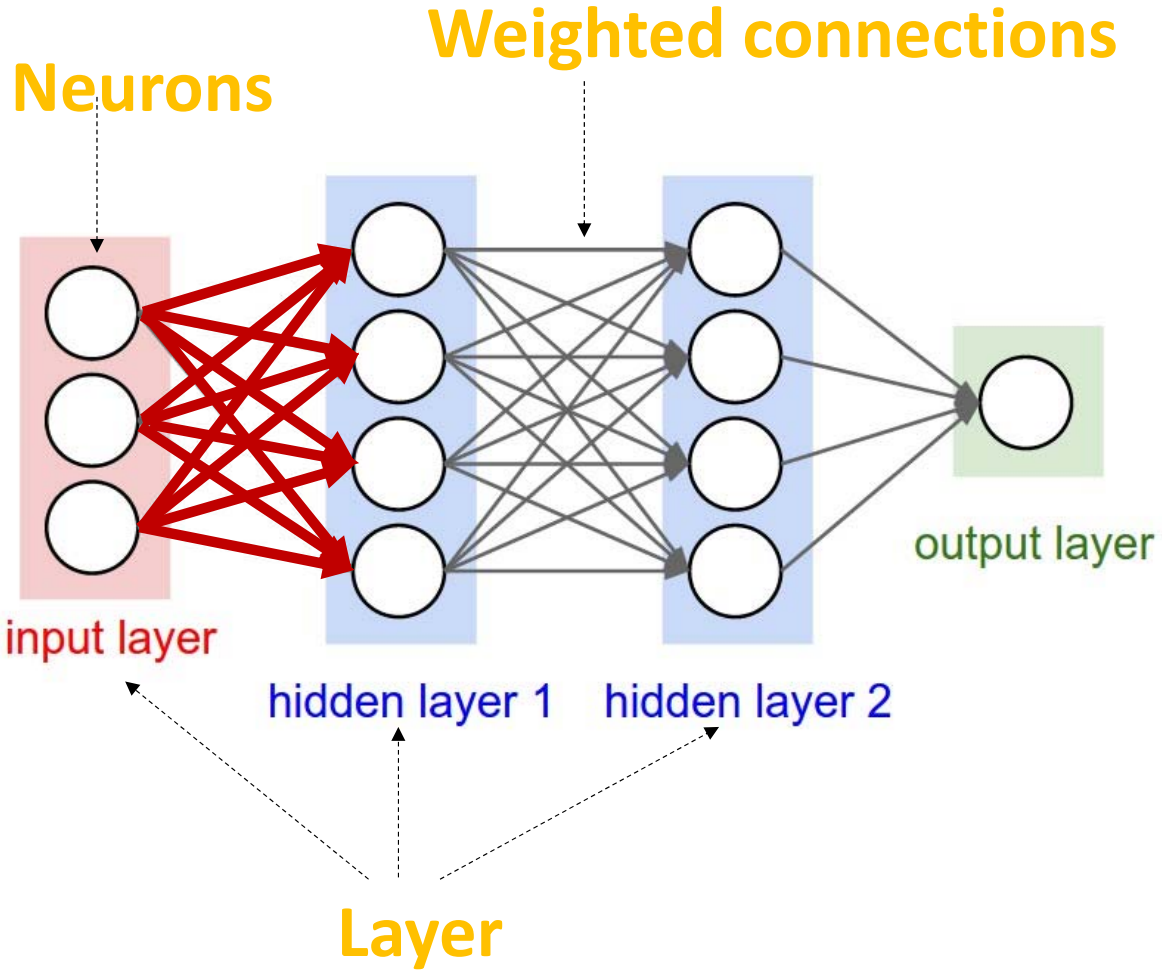
Machine learning methods need large sample size

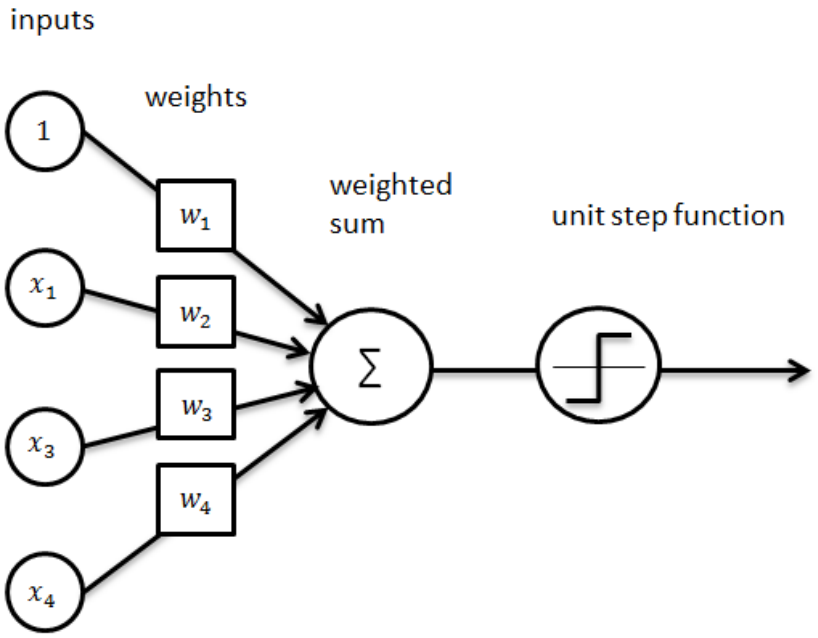
“taking a minimum sample size of 500 is necessary to derive the statistics that represent the parameters”

Software Name	Training set	Testing set
MARVEL	1247 phages + 1029 bacteria host	335 bacteria + 177 phage
Prophage Hunter	1031 phages + 21979 bacteria host	2509 phages + 5495 bacteria

Mohamad Adam Bujang, et al. *Malays J Med Sci.* 2018

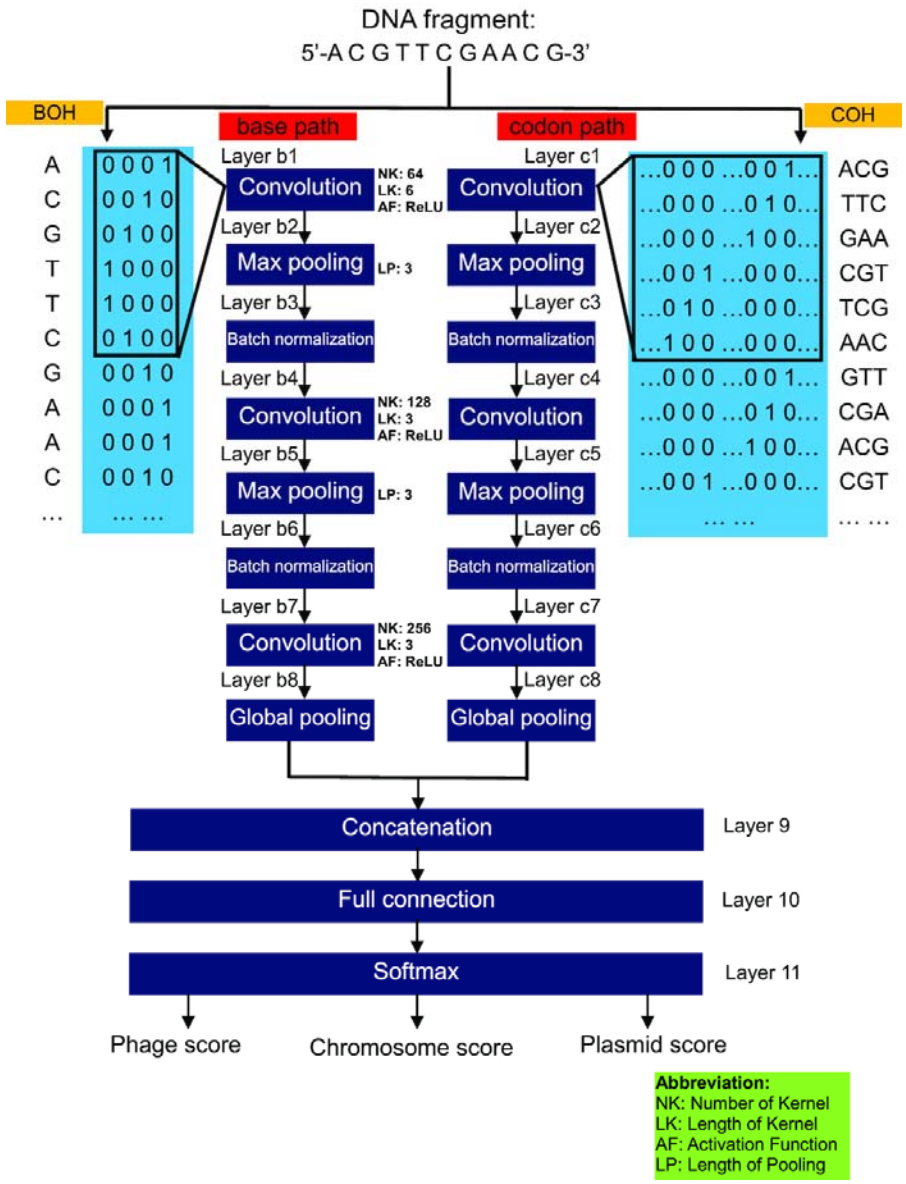
Strategy 3: deep learning methods using neural network





Structure of BiPathCNN

- BiPathCNN contains a “base path” and a “codon path,” which take BOH and COH as inputs, respectively. After multiple convolution operations, the data for the 2 paths are combined by a merge layer.
- 2,700,000 artificial contigs were generated to train PPR-Meta



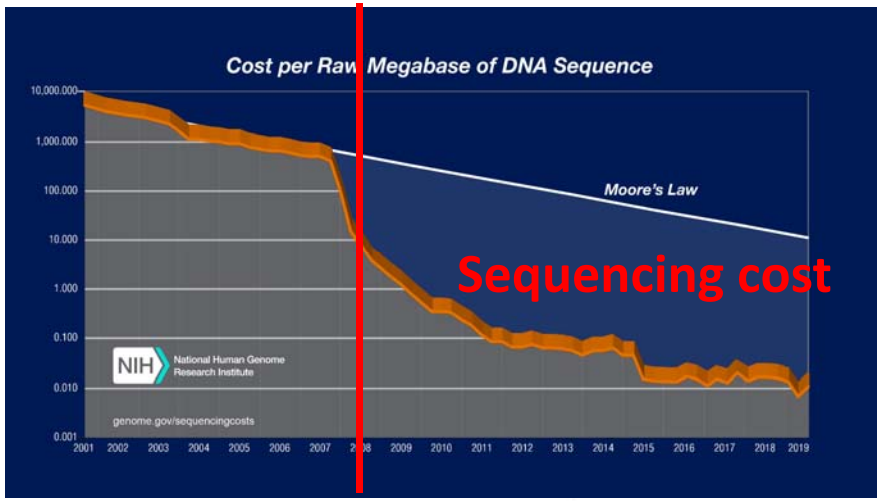
Summary I: Comparison among three strategies

		Strategy 1	Strategy 2	Strategy 3
Basic information	Used years	2005-present	2016-present	2018-present
	Methods	Conventional	Machine learning	Deep learning
	Selected features	None	Manually	Auto
	Explainable	Easy	Relatively easy	Hard
Required resources	Input data requirement	Medium	Large	Large
	Computational resources	Low	Middle	High

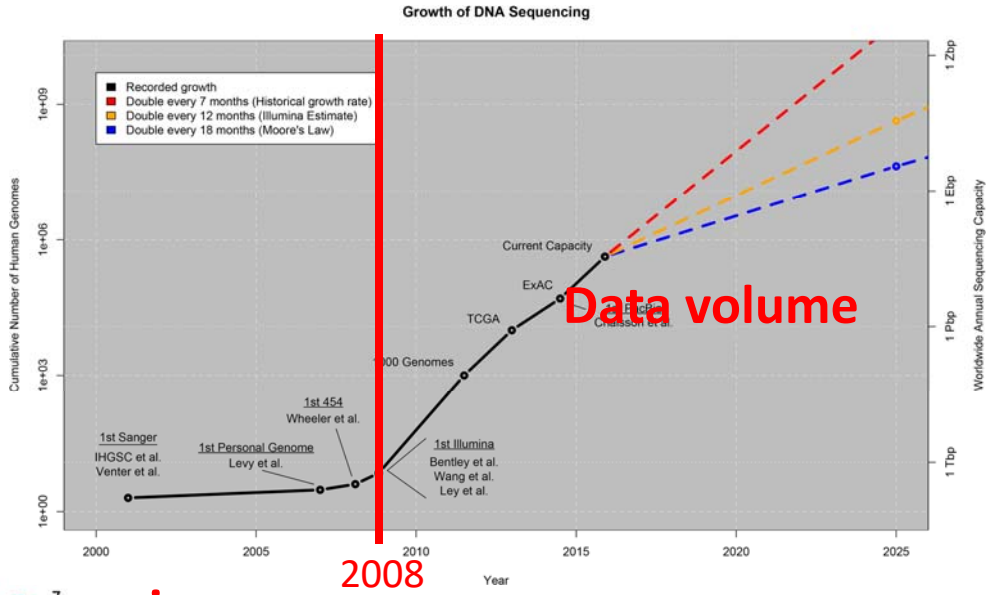
Summary II: Limitations and opportunities

- Machine learning methods use more information compared with conventional methods
- Complex models don’t always mean better models, especially when we have limited training datasets
  - Interactions between features
  - Overfitting
- More comprehensive nucleotide and protein reference database for virus is needed
- Combination of multiple methods is possible for better performance

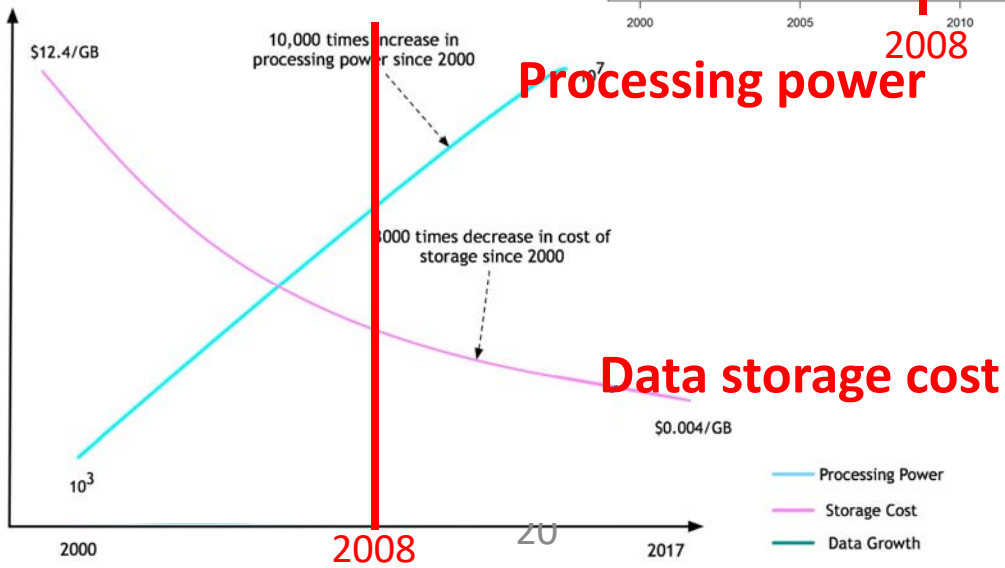
Future prospective I



2008



2008



2008

https://towardsdatascience.com/the-future-of-computation-for-machine-learning-and-data-science-fad7062bc27d
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data
https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195

Future prospective II

1. Data are accumulating and hardware is getting more powerful, which provide the foundations for more sophisticated models
2. Computational biology is becoming a cross-disciplinary area, in order to take advantage of the big data and architecture, applying those skills to leverage big data is vital in future research for computational biologists.

References

Dyer, M. D., Murali, T. M. & Sobral, B. W. Computational prediction of host-pathogen protein–protein interactions. *Bioinformatics* **23**, i159–i166 (2007).

Arango-Argoty, G. *et al.* DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).

Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front. Genet.* **9**, (2018).

Fouts, D. E. Phage_Finder: Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res* **34**, 5839–5851 (2006).

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: A Fast Phage Search Tool. *Nucleic Acids Res* **39**, W347–W352 (2011).

Fang, Z. *et al.* PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* **8**, (2019).

Bose, M. & Barber, R. D. Prophage Finder: A Prophage Loci Prediction Tool for Prokaryotic Genome Sequences. 5.

Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* **24**, 863–865 (2008).

Guttman B, Raya R, Kutter E. 2005. Basic phage biology, p 29-66. In Kutter E, 268 Sulakvelidze A (ed), Bacteriophages: biology and applications

Comeau AM, Hatfull GF, Krisch HM, Lindell D, Mann NH, Prangishvili D. 2008. Exploring 274 the prokaryotic virosphere. Res Microbiol 159:306-13.

THANK YOU

香港中文大學
The Chinese University of Hong Kong

香港中文大學醫學院
Faculty of Medicine
The Chinese University of Hong Kong