Supplementary Materials

Enlarging training data with metavirome contigs

We added millions of viral sequences from metavirome samples for training to represent a more diverse viral population and boost the prediction accuracy of the deep learning algorithm at the same time. To make sure the sequences are purely from viruses, we carefully selected samples that had low contamination rate of host sequences. To measure the contamination rate of each sample, we mapped reads using bowtie2 (2.3.2) [8] in each sample to a reference database containing 7726 prokaryotic RefSeq used in VirSorter [22] and human RefSeq (GRCh38.p7), and used the mapping rate as the contamination rate. Since prokaryotic genomes may contain prophages that could overestimate the contamination rate, prophages were detected using VirSorter and removed from the prokaryotic RefSeq. Samples with contamination rate lower than 5% were considered relatively pure and were used in the downstream analysis. Samples from the same sequencing platform were combined and cross-assembled into contigs. Megahit [9] was used to assemble metavirome reads from Illumina sequencing, and Ray [1, 2] was used to assemble reads from Roche 454. To further reduce the non-viral contamination *in silico*, the resulting contigs were filtered through VirSorter and only the viral contigs with the highest confidence (Category I and II) were used for training. Samples from TOV were carefully cross-assembled and filtered in Roux et al. [21], so the resulting viral contigs were directly used for training in our study. The metavirome contigs were then fragmented into millions of fixed-length sequences. Table S3 lists the information of the metavirome datasets used in this study. The metavirome sequences were combined with sequences derived from viral RefSeq before May 2015 for training. The new model was evaluated and compared with the original model trained using only RefSeq, based on the test sequences from RefSeq after May 2015.

Simulation of metagenomic datasets

Metagenomic samples were simulated based on species abundance profiles derived from a real human gut metagenomic sample (accession ID SRR061166, Platform: Illumina) from the Human Microbiome Project (HMP) [17], commonly used for metagenomic data analysis [2, 12, 18, 3]. We first mapped reads from sample SRR061166 using bwa-0.7.15 [10] to virus and host genomes sequenced after May 2015 to generate the abundance profile. Here we only used RefSeq after May 2015 for evaluation to avoid any overlap with the training dataset, i.e. RefSeq before May 2015. Following a similar procedure as in VirFinder [19], reads from each sample were first mapped to viral RefSeq and the remaining unmapped reads were then mapped to host RefSeq using the command of `bwa mem`. About 2% of reads can be mapped to viral genomes, lower than the range of previously estimated viral fraction 4-17% for human gut metagenomics. This is largely due to the fact that only viral RefSeq after May 2015 were used for read mapping, which represent just a small subset of the total virus database. The abundance profiles can be found in the Supplementary Table 2.

We simulated metagenomic contigs based on the abundance profile of virus and host genomes. Given a total budget of 10 million base pairs for contigs, the number of base pairs for contigs from

each genome was computed proportionally. For each reference genome, contigs were sampled randomly and independently from the genome, where the contig length follows the same distribution as that in a real human metagenomics dataset for colorectal carcinoma patients, until the number of base pairs reaches 10 Mbp. Note that here we generated contigs directly from genomes, instead of directly simulating reads and then assembling contigs. This simualtion procedure avoided chimeric contigs that were artificially assembled using reads from different genomes. The R packages ROCR [23] and caTools [25] were used to compute AUROC and AUPRC, and the variation were evaluated using 30 bootstrap samples.

Viral analysis of human gut metagenomics from patients with colorectal cancer

Human gut metagenomics samples from patients with colorectal cancer and the control group were downloaded from European Nucleotide Archive (ENA) database (http://www.ebi.ac.uk/ena) with accession number ERP005534. The samples from 53 cancer patients and 61 normal patients were randomly split into 2/3 for training and 1/3 for testing. The patient ID and the disease status can be found in the Supplementary Table 3. The metagenomics samples from training were combined and cross-assembled using Megahit [9] in default settings. The majority 64% of the assembled contigs have the length ranging from 300-1000 bp (Figure 4a). We filtered contigs smaller than 500 bp to guarantee high accuracies in the downstream analysis including viral prediction and contig binning. We used DeepVirFinder to predict viral contigs. To fairly compare the prediction scores across sequences of different lengths, we normalized the prediction scores by computing a p-value for each score. The p-value was computed by comparing the score with the corresponding score distribution for host sequences in the same range of sequence length in the validation dataset. To control the false discovery rate, the predicted p-value for each contig was converted to a q-value using the R package qvalue [24]. The q-value is an estimation of the proportion of false prediction if the prediction is made at the level of the corresponding p-value. Contigs were sorted by q-values from the smallest to the largest, and the contigs having q-values $<$0.01 were predicted as viruses. The viral contigs predicted by DeepVirFinder were then grouped into contig bins using the software COCACOLA [11] in the default mode.

To study the association between the viruses and the cancer status, we mapped reads in each sample against the viral contigs in each bin using bowtie2 (2.3.2) [8]. The number of reads per kilobase of the contig per million mapped reads (RPKM) was used as a measure of contig abundance, and the average of the RPKM of contigs in each bin was defined as the contig bin abundance. Based on the abundance of viral bins in the training samples, a logistic regression classifier with L1 penalty was built to predict the cancer status using the R package glmnet [5]. The parameter lambda was determined using 5 fold cross-validation. PfamScan [4] and Blastn-2.6.0 (Evalue $<$1e-5) were used respectively to search proteins and DNA sequence against Pfam and NCBI nonredundant databases.

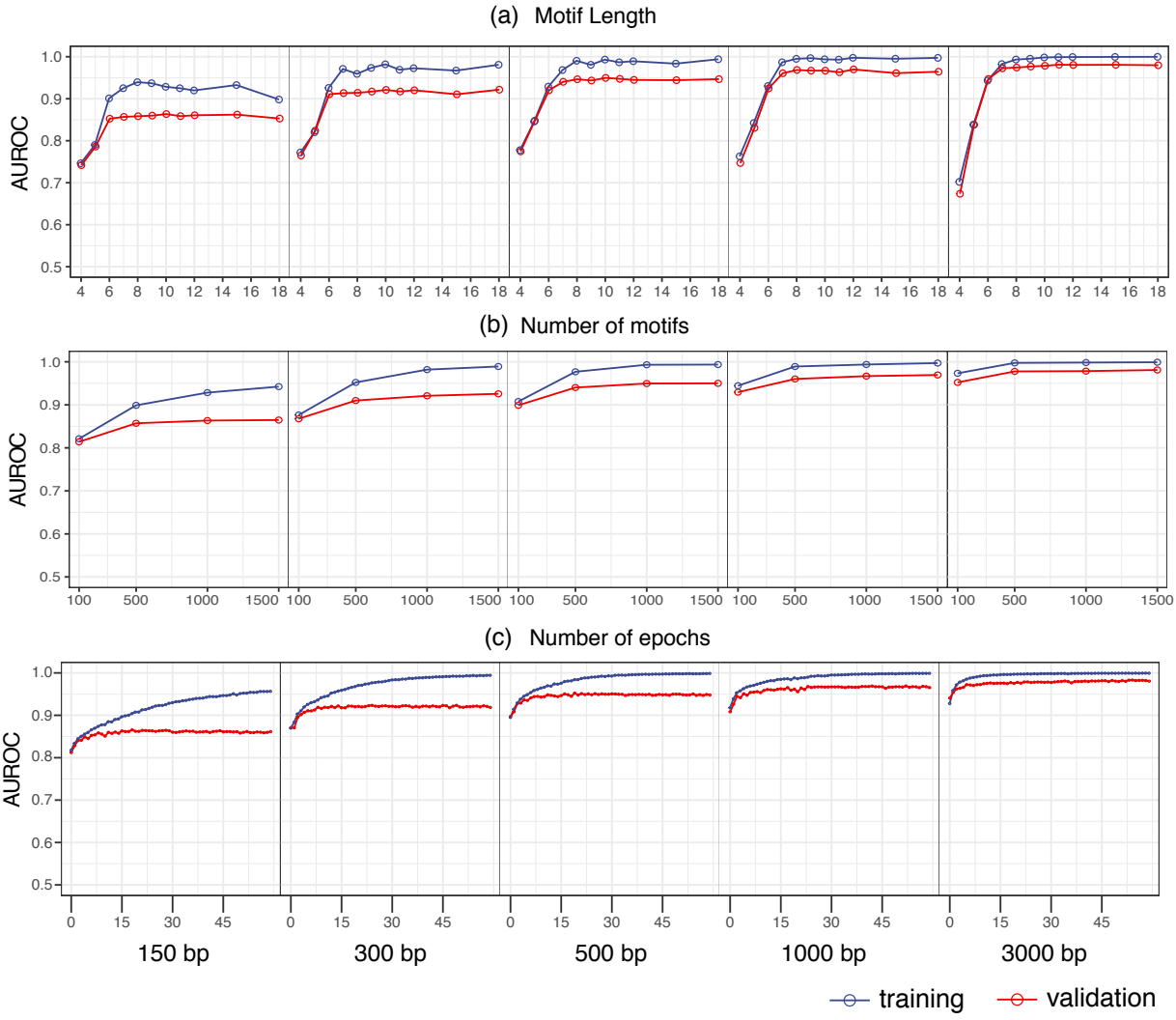

Figure S1: Determining the optimal paramter settings of DeepVirFinder. The effect of motif length, number of motifs, and the number of epochs on validation AUROC.

Table S1: AUROCs for VirFinder and DeepVirFinder when trained on sequences before May 2015, and tested on sequences after May 2015. The numbers in brackets are the standard errors estimated using 30 bootstrap samples.

Contig Length (bp)	VirFinder	DeepVirFinder
150	0.8101 (0.0007)	0.8766 (0.0007)
300	0.8771 (0.0012)	0.9272 (0.0005)
500	0.9163 (0.0012)	0.9494 (0.0011)
1000	0.9471 (0.0012)	0.9735 (0.0004)
3000	0.9770 (0.0008)	0.9847 (0.0009)

Table S2: The viral contig bins associated with the CRC. For each bin, the following information is shown: the coeficient of the bin, the number of contigs in the bin, the percentage of contigs containing proteins, the top Pfam hits of the proteins in the bin, and the top BLAST hits to the NCBI database.

Bin ID	Coeficient	# contigs	protein%	Pfam hits	Blastn hits
B7	-0.0193	957	61%	Phage related (tail, capsid, integrase, connector, holin, portal), Tail_P2_I, PhageMin_Tail, Podovirus_Gp16, RNA_lig_T4 Terminase, DUF	crAssphage, Unidentified phage
B19	0.1029	963	63.66%	Phage related (tail, capsid, integrase, connector, holin, portal), TMP-TENI, T4SS, Terminase, DUF	Escherichia virus, Enterobacteria phage, Salmonella phage, Stx1/Stx2 converting phage, Bacteriophage, Lambda genome, Unidentified phage
B20	-0.0743	254	31.10%	Phage related (tail, capsid, integrase, connector, holin, portal), TMP (Prophage tail length tape measure protein), DUF	Homo sapiens isolate endogenous virus, Bacteriophage, Enterobacteria phage, Stx1/Stx2 converting phage, Escherichia phage, Salmonella phage
B60	0.1666	384	53.65%	Phage related (tail, capsid, integrase, connector, holin, portal), Podovirus, Terminase, DUF	Staphylococcus phage, Unidentified phage, Aureococcus anophagefferen,
B61	0.0924	702	51.85%	Phage related (tail, capsid, integrase, connector, holin, portal), ADH_zinc_N, Terminase, DUF	Streptococcus phage, Unidentified phage clone, Faecalibacterium phage
B87	-0.0223	26	96.15%	DUF	NA
B110	-0.3475	72	94.44%	DUF	NA
B188	0.0174	56	35.71%	Phage_T4_gp19	NA
B218	0.0455	107	85.05%	Phage_sheath, DUF	NA
B227	0.1764	159	84.28%	DUF	Moraxella phage

Table S3: (a) Metavirome datasets used for generating more viral contigs for training. (b) Number of sequences at different lengths used for training for each type of metavirome dataset. PE: paired-end reads; # pure sample: number of samples have contamination rate <5%; Data size: the total size of the data.

* Roux et al. [21] used a customized cross-assemble pipeline with MOCAT [7] and Idba_ud [15, 16] to assemble the 104 samples from TOV project, and resulting contigs were further filtered by VirSorter to identify 298,383 viral contigs.

(a) Metavirome datasets

Human Gut Metavirome

Dataset	Platform	PE	Sample#	Pure Sample#	Purity Rate	Data Size	Assembler
IBD [14]	Illumina	Y	171	131	76.61%	174G	Megahit
SAM[20]	Roche 454	N	320	281	87.81%		
Healthy [13]	Roche 454	N	18	1	5.56%	8.82G	Ray
Healthy [6]	Roche 454	N	6	5	83.33%		

Tara Ocean Metavirome

Roux et al (TOV) [21]	Illumina	Y	104	NA*	NA*	925G	Customized*

(b) Number of sequences in metavirome datasets

Fragment length	Human Gut (Roche 454)	Human Gut (Illumina)	TOV	Total
150 bp	9,467	3,389	1,354,007	1,366,863
300 bp	4,688	1,658	671,331	677,677
500 bp	2,772	965	398,342	402,079
1000 bp	1,337	442	193,373	195,152
3000 bp	388	91	56,698	57,177

References

[1] S. Boisvert, F. Laviolette, and J. Corbeil. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, 17(11):1519–1533, 2010.

[2] S. Boisvert, F. Raymond, É. Godzaridis, F. Laviolette, and J. Corbeil. Ray meta: scalable de novo metagenome assembly and profiling. *Genome Biology*, 13(12):R122, 2012.

[3] M. J. Brittnacher, S. L. Heltshe, H. S. Hayden, M. C. Radey, E. J. Weiss, C. J. Damman, T. L. Zisman, D. L. Suskind, and S. I. Miller. Gutss: An alignment-free sequence comparison method for use in human intestinal microbiome and fecal microbiota transplantation analysis. *PloS One*, 11(7):e0158897, 2016.

[4] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, et al. The pfam protein families database in 2019. *Nucleic Acids Research*, 2018.

[5] J. Friedman, T. Hastie, and R. Tibshirani. glmnet: Lasso and elastic-net regularized generalized linear models. *R package version*, 1(4), 2009.

[6] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, 26(12):1721–1729, 2016.

[7] J. R. Kultima, S. Sunagawa, J. Li, W. Chen, H. Chen, D. R. Mende, M. Arumugam, Q. Pan, B. Liu, J. Qin, et al. Mocat: a metagenomics assembly and gene prediction toolkit. *PloS One*, 7(10):e47656, 2012.

[8] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357, 2012.

[9] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

[10] H. Li and R. Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[11] Y. Y. Lu, T. Chen, J. A. Fuhrman, and F. Sun. COCACOLA: binning metagenomic contigs using sequence composition, read coverage, co-alignment and paired-end read linkage. *Bioinformatics*, 33(6):791–798, 2017.

[12] C. Luo, L. M. Rodriguez-r, and K. T. Konstantinidis. Mytaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research*, 42(8):e73–e73, 2014.

[13] S. Minot, R. Sinha, J. Chen, H. Li, S. A. Keilbaugh, G. D. Wu, J. D. Lewis, and F. D. Bushman. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Research*, 21(10):1616–1625, 2011.

[14] J. M. Norman, S. A. Handley, M. T. Baldridge, L. Droit, C. Y. Liu, B. C. Keller, A. Kambal, C. L. Monaco, G. Zhao, P. Fleshner, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell*, 160(3):447–460, 2015.

[15] Y. Peng, H. C. Leung, S.-M. Yiu, and F. Y. Chin. Idba–a practical iterative de bruijn graph de novo assembler. In *Annual International Conference on Research in Computational Molecular Biology*, pages 426–440. Springer, 2010.

[16] Y. Peng, H. C. Leung, S.-M. Yiu, and F. Y. Chin. Idba-ud: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, 2012.

[17] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. A. Schloss, V. Bonazzi, J. E. McEwen, K. A. Wetterstrand, C. Deal, et al. The nih human microbiome project. *Genome Research*, 19(12):2317–2323, 2009.

[18] S. Rampelli, M. Soverini, S. Turroni, S. Quercia, E. Biagi, P. Brigidi, and M. Candela. Viromescan: a new tool for metagenomic viral community profiling. *BMC Genomics*, 17(1):165, 2016.

[19] J. Ren, N. A. Ahlgren, Y. Y. Lu, J. A. Fuhrman, and F. Sun. Virfinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, 5(1):69, 2017.

[20] A. Reyes, L. V. Blanton, S. Cao, G. Zhao, M. Manary, I. Trehan, M. I. Smith, D. Wang, H. W. Virgin, F. Rohwer, and others. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proceedings of the National Academy of Sciences*, 112(38):11941–11946, 2015.

[21] S. Roux, J. R. Brum, B. E. Dutilh, S. Sunagawa, M. B. Duhaime, A. Loy, B. T. Poulos, N. Solonenko, E. Lara, J. Poulain, et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature*, 537(7622):689–693, 2016.

[22] S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan. Virsorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, 2015.

[23] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, 2005.

[24] J. D. Storey et al. The positive false discovery rate: a bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

[25] J. Tuszynski. catools: Tools: moving window statistics, gif, base64, roc auc, etc. *R package version*, 1, 2008.