Lin Zhang

Recover and Analysis Virus genomes from Natural Microbial Communities

Abstract

Viruses are everywhere around us. They are the most abundant biological entities on Earth and play key roles in host ecology, evolution, and horizontal gene transfer. Viruses also has important implication to human health. Since its diversity and importance, more and more technologies are further exploring to this field, but its diversity and functionality are still hard to fully detect and predict. Lots of new viruses are still unknown or keeping mutated. Due to recent progress in viral metagenomics studies, there are some assembly-free computational pipeline approaches enabling direct recovery of high-quality viral genome sequences from metagenomic samples or uncovering draft genomes. If some of the tools can fully recover virus's gene sequences base on some reference data. This approach will be a breakthrough in the exploration of gene types and the analysis of functionality. This report is mainly focus on recover virus's gene sequences, testing programming efficiency and verify whether the recovered gene is reasonable.

Introduction

The program we are testing is FVE-novel and FastViromeExplorer. FastViromeExplorer is a program to identify the viruses/phages and their abundance in the viral metagenomics data. FVE-novel is be used to recover draft genomes of viruses and phages through reference-based mapping and iterative assembly using Spades. Base on some research, we found a paper which mentioned they use short-read assembly approaches but could not be fully recovered any more virus genome sequences from BioSample: SAMN11262775, SAMN11267325, and SAMN11267326. (Single reads: SRR8811960 (11GB), SRR8811961 (9.7), SRR8811964 (23GB). Double reads: SRR8811962 (8.9GB), SRR8811963 (11GB)).
Sources : https://www.biorxiv.org/content/10.1101/619684v1.full#F4

Therefore, we want to test our new approach if we can recover any virus sequences. In this specific case, the double read files will be our input for mappings and calculations. The goal is to test whether our FVE-novel program can recover some gene sequences that exists in the single read files. If we recover some high percentage sequences which can prove our scaffold approach is more complete than previous assembly-free, single-molecule nanopore approach. In this testing, the reference data we are using is GOV-viral-populations (6.7G). Theoretically, the larger biosphere reference data will be more proportion to recover some high percentage genes. But consider to efficiency, we will first use GOV as our reference data.

Method and Result

First, we run FastViromeExplorer using double read files and all runs are based on GOV reference data. This is the time cost:

SRR8811962 (8.9GB) takes 39m51.246s
SRR8811963 (11 GB) takes 59m19.743s.

This process does not take much time since it is only a procedure for identifying viruses and their abundance in viral metagenomic data. Then, use these results as an input for FVE-novel. FVE-novel will break down the sequences into many sequence scaffolds, do some cutting and mapping. FVE-novel cost:

SRR8811962 (8.9GB) takes 8362m5.626s
SRR8811963 (11 GB) takes 9746m8.261s

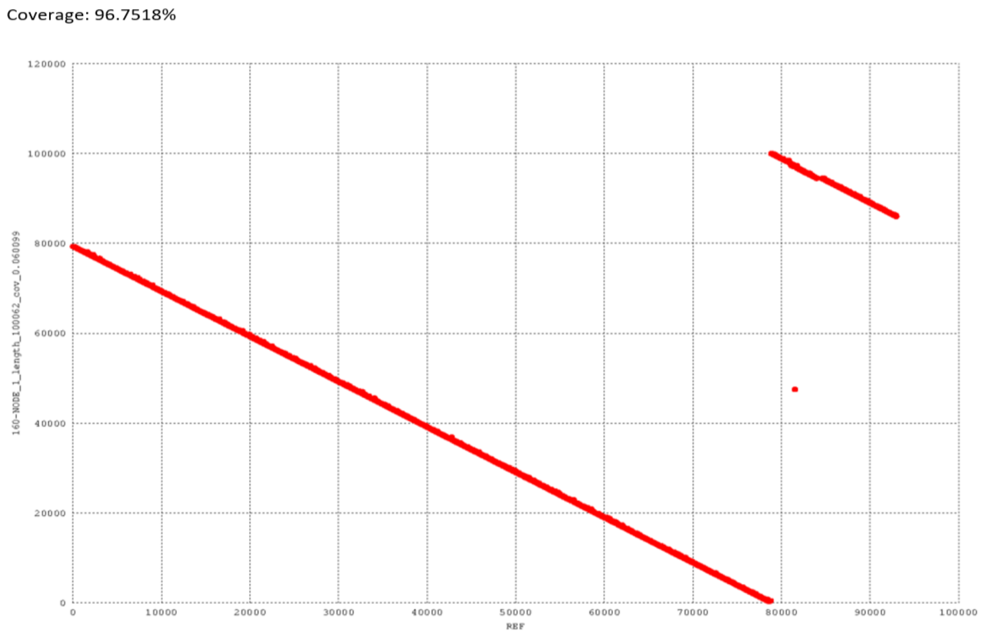
After running, we got two output files named *all-scaffolds-corrected.fasta(SRR8811962)* and *all-scaffolds-corrected.fasta(SRR8811963).* These files store all recovered genes sequences we produce. We process these data for further comparing. In this case, we extracted all sequence lengths form single read files and sort them. We are trying to find the longest 1000 sequences for next coverage test. In theory, the longer sequences are more contrastive and more likely to be recovered than shorter sequence. We have proven this in the later experiments. After extracting the longest 1000 gene segments from single read file and incorporated into a new fasta file. Next, we use blast to create a database (named "pr2"). Command:

makeblastdb -in Longest_1000_total.fasta -input_type fasta -dbtype nucl -out pr2

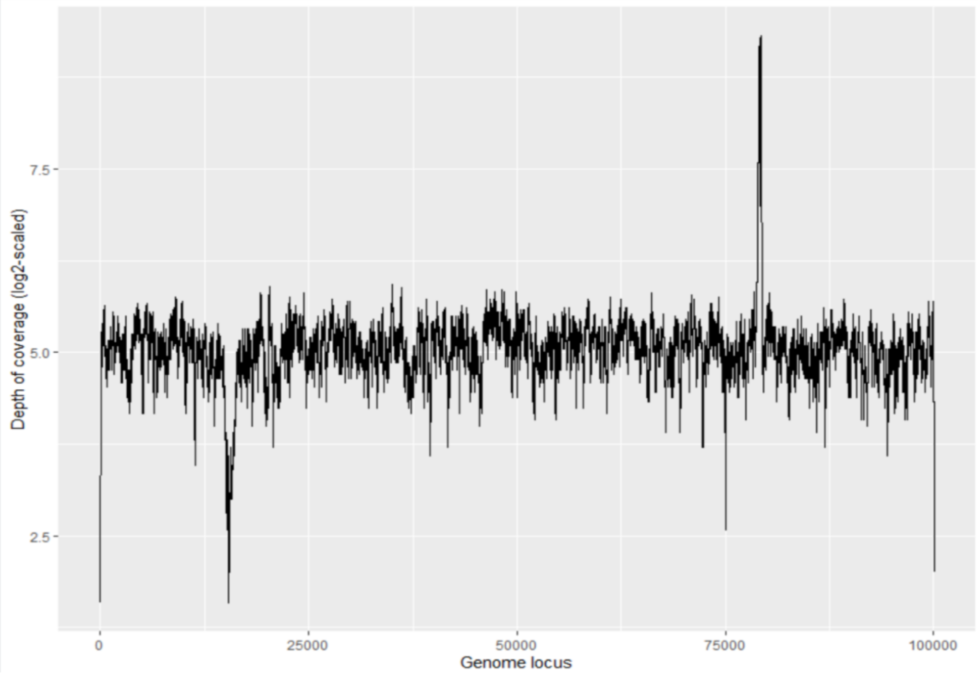
The pr2 will be used as our new database which come from single read file. Following, Run blastn between FVE- novel output and pr2 database (blastn is a build-in function from blast). Blastn is a quick comparison, it only cost few seconds. The results we got is what we want for comparing original genes and recovered sequences. The only issue we found is there are some overlapping in blastn result. Which means some of our FVE-novel nodes are mapping or sharing same positions in corresponding segments. We suppose to remove all overlapping and only count once when FVE-novel output sequences maps pr2 database. After rearranging the data, we calculate the total length for each recovered position in each sequence. Comparing these lengths with original sequences in database (longest 1000 sequences). We got the percentage that FVE-novel recovered. The highest one we produce is 96.7518% and following by 86.03854%, 85.29554%. The fourth one is dropping to 60% which we do not want to consider. Using same approach, we also extracted the 1000th to 2000th longest segments from single read file and do the same approach. The results are very much as same as we expected. The highest recovery percentage is 73.21821%, which is much lower than first 1000th longest sequences. This prove what we assume before, the longest sequences are more likely to be recovered since more contrast.

In order to analyze these high proportion of sequences more intuitively, we use Mummer and VirChecker. Mummer is a system for rapidly aligning entire genomes, whether incomplete or draft form. This program can plot all points according to the corresponding positions, so we can check which range of sequences were recovered and which pieces were missed. VirChecker is a program for recover error-free full genomes of viruses and phages by polishing and extending draft viral assemblies.

Mummer:



VirChecker:



For this highest 96.7518% coverage sequences. The original length is 93036. Total coverage length is 90014 which come from FVE-Novel output node length in 100062.

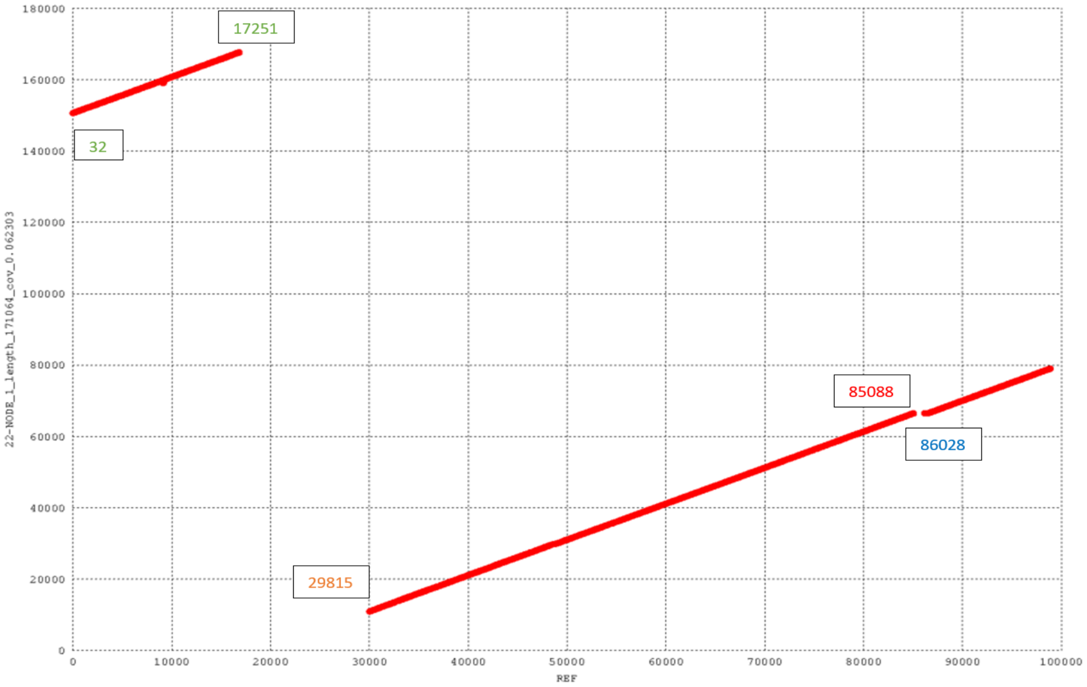
From these data, we know there are some unused length in FVE-Novel Node (100062-90014). There are two possibilities. 1.original sequence is not a complete gene. 2. There are some free-mapping in FVE-Novel approach. For Mummer graph, the x-axis is original sequences and y-axis is the FVE-novel Node. Almost all x-axis (original sequences) were covered by our result sequence. For y-axis, there is a gap between 80000 to 83000, which means this interval did not be used to recover any part of the original gene and we can also see this in VirChecker plot. There is a Huge fluctuation around this location (80000 to 83000). In meantime, we can speculate that this gene fragment may be a circular sequence. For example, if we assume the original gene was “ABCDEFGH”, so the recovered FVE-Node might be “FEDVBAHG“.

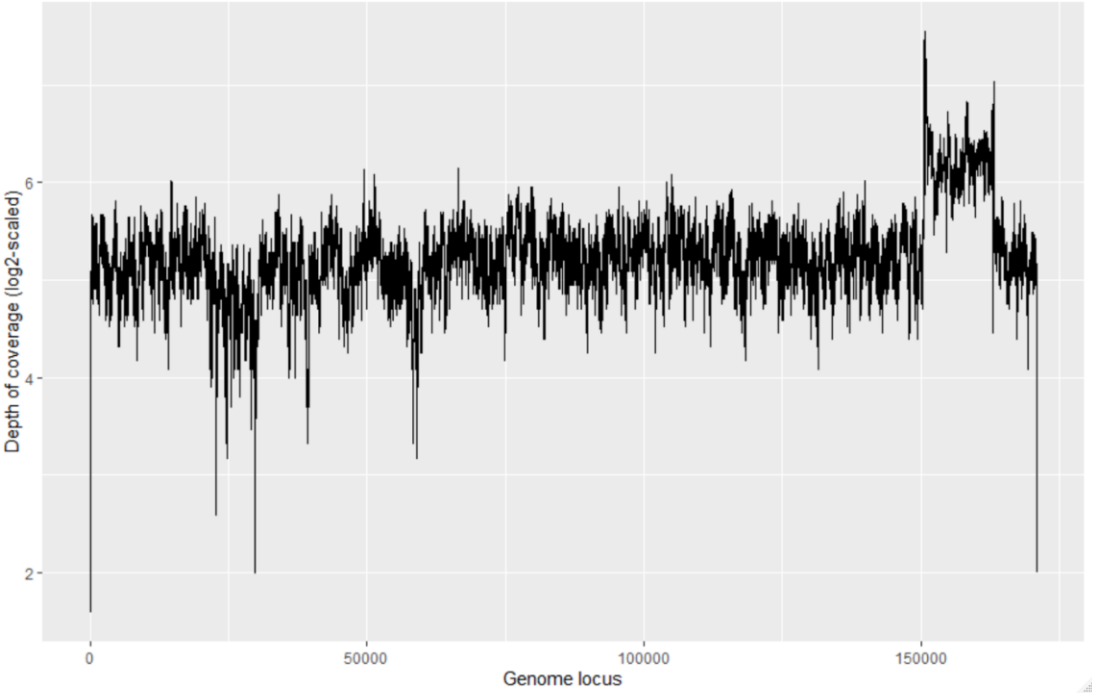
→
A B C D E F G H
| | | | | | | |
F E D V B A H G
←

Since Mummer does not remove overlapping, so we will see there is a point in the Mummer plot around position (81000, 50000). In vertical direction, there are two positions map to correspond x = 81000. In horizontal direction, that are two points map y = 50000 too. This means in the original gene sequence, there is a similar fragment so they can be mapped by the same y-axis twice. If we roughly calculate the horizontal range of the interval, it is about (81000-31000 = 50000). For vertical direction, it is around (98000-48000 = 50000). Since these ranges has similar length, they are more likely to be the corresponding region.

Next, we come to the second scale.

Coverage: 86.03854%





For this testing, the original length is 98908. Total coverage length is 85099 produced by FVE-Novel Node length 171064.

Position mapping(from Blastn):

FVE-Output			Original Sequence	
1	7535	mapping--	69065	76527
29868	66518	mapping--	48841	85088
10698	29889	mapping--	29815	48694
150656	168106	mapping--	32	17251
66296	79020	mapping--	86357	98906
66296	66504	mapping--	86028	86233

Obviously, there is a specific portion in VirChecker plot roughly from 150000 to 165000. The gene plot has a significant high band elevation. In general, this means this part of sequence are not coming from same sequence. But the really strange thing is if we look at the mapping interval from Blastn. There is a mapping from 150656 to 168106 in FVE-Output and it maps to 32 to 17251 in original sequence (marks in read). Which means 150000 to 165000 was used in comparing and mapping to original sequences. Meanwhile, even though there is an 86 percent of gene coverage but if we look at the cover position on FVE-novel Node sequence (y-axis). It is mainly located between 10000 to 80000 and 150000 to 165000. This kind of arrangement has a high probability of being a cycle sequence. Regarding to the rest of the sequences, if we look at the Virchecker we can see, the interval between 80000 to 150000 is smoother than 0 to 80000, which means they probably come from a same gene segment. But this interval was not be used for mapping, so we might infer that this is a continuous gene, but it is not complete.

Conclusion

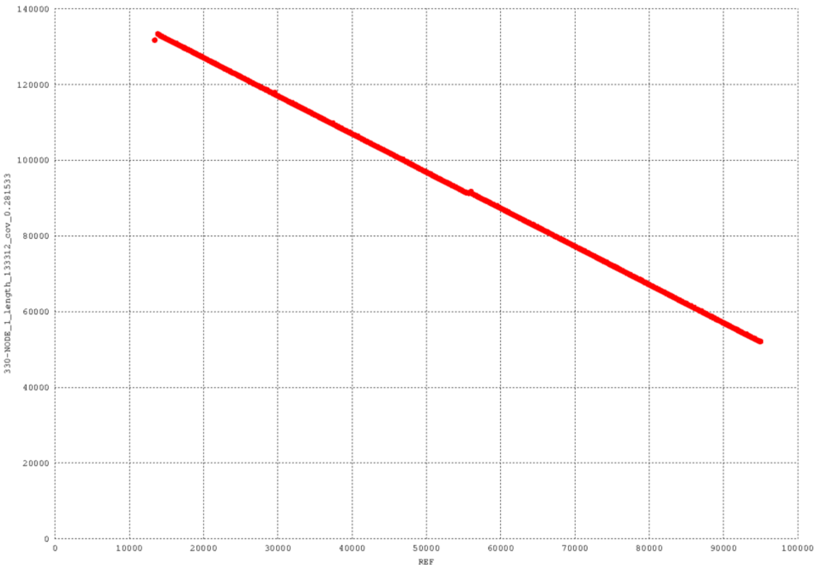
In this research, we tested FVE-novel, whether this new computational pipeline approach can fully recover some various gene base on sequence segments and GOV database. Some of these sequences are quite long, and have a high percentage of coverage, which can be potential near-complete viral genomes. In the first 2000 longest recovered genomes, four of the most complete viral genomes were artificially extracted to see if they matched the original genes. This allows us to accurately test how well the sequence be recovered by FVE-novel. In this experiment, we found one gene fragment is likely to be a cycle gene, and the other sequence in the biosample is more likely to be incomplete. Overall, this FVE-novel implements a novel strategy to recover viral genomes which can do more gene recover than before. It is a powerful tool for future genes detection and prediction.

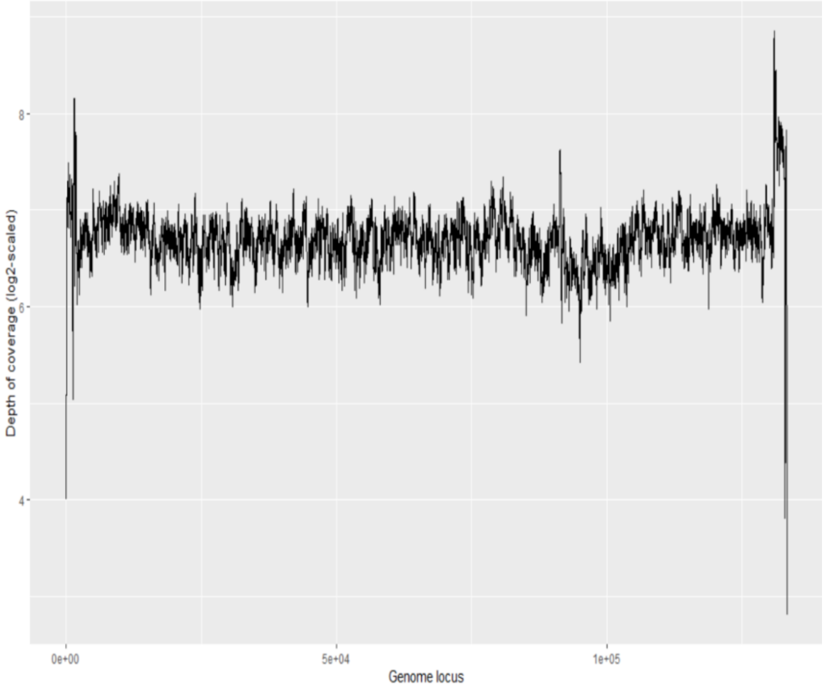
Other outputs

Coverage: 85.29554%

Original: SRR8811964.171749:	Length: 95046
FVE-novel Node:	Length: 133312
Tot_cov_length:	Length: 81070

Position:					
13825	26354	26366	55438	55562	95031



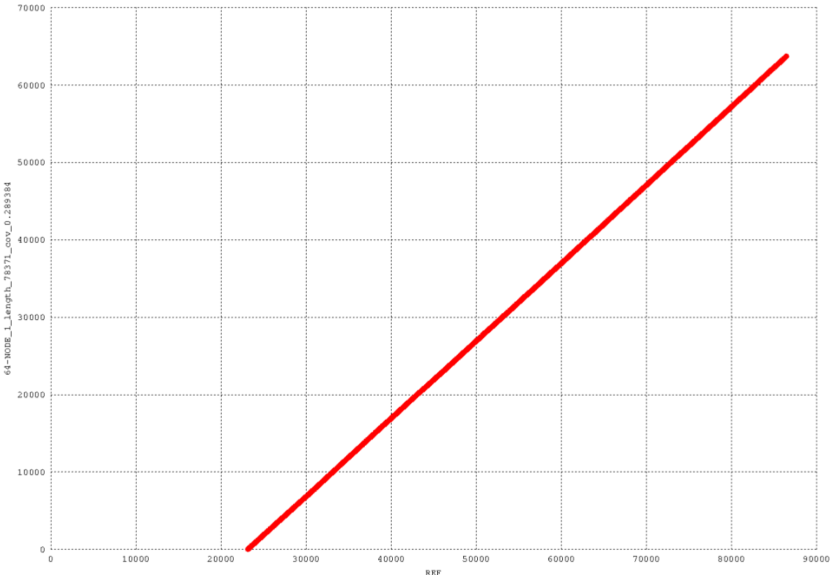


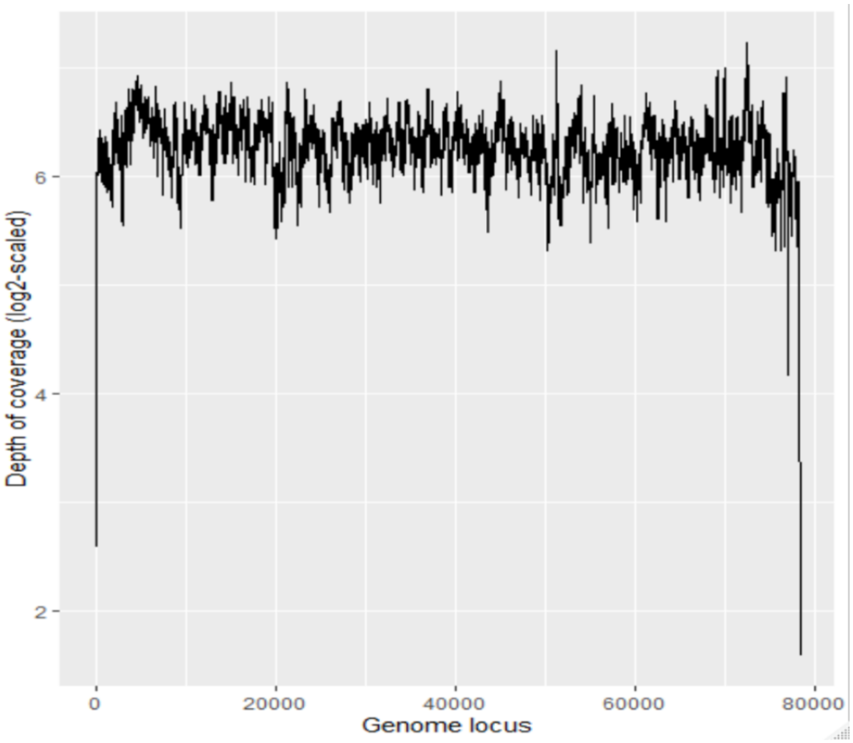
Coverage: 73.21821 %

Original: SRR8811964.245695:	Length: 86514
FVE-novel Node:	Length: 78371
Tot_cov_length:	Length: 63344

Position:

23167	86511


64-NODE_1_length_78371_cov_0.289384
REF


Depth of coverage (log2-scaled)
Genome locus