CRA-W
Computing Research Association
Women

the coalition
to diversify computing

CREU 2016-2017 Final Report:
Harvesting Viral Genomes from Sequences of Complex Communities

Andrea Garretto, Loyola University Chicago

I) Goals and Purpose

While metagnomics has facilitated the sequencing of viral communities inhabiting niches from across the globe, each study uncovering sequences with no recognizable homology to known proteins or genomes. Viral data repositories do not include sufficient representation of the diversity of viruses on Earth. Nevertheless, novel viral species genomes – particularly those in high abundance – have been able to be mined directly from complex community viral metagenomes. Discovery of such viral genomes often relies heavily on manual curation and prior studies have employed a variety of different criteria when sifting through sequencing data. In an effort to provide a comprehensive means of mining for complete viral genomes from complex sequence data sets, we developed the tool virMine. This tool provides an automated solution while maintaining flexibility, allowing researchers to select from a variety of filters.

II) Related Work

The most abundant organisms on Earth are viruses, of which one of the most notable are bacteriophages, or viruses that infect bacteria (Breitbart & Rohwer 2005). Bacteriophages (phages) play a critical role in the environment, structuring bacterial communities (Clokie et al. 2011; Jacquet et al. 2010) and mediating host mortality on a global scale (Berdjeb et al. 2011). As such, viral metagenomic studies have been conducted for numerous habitats on Earth (e.g. marine waters (Breitbart et al. 2002; Yooseph et al. 2007; Hurwitz & Sullivan 2013; Brum et al. 2015), soil (Williamson et al. 2007; Fierer et al. 2007; Zablocki et al. 2014), and freshwaters (López-Bueno et al. 2009, 2015; Roux et al. 2012)). In addition to their environmental importance, phages are vital components of the human microbiome (Virgin 2014). Studies of the phage communities within the gastrointestinal tract (Reyes et al. 2010; Minot et al. 2011, 2013; Norman et al. 2015; Manrique et al. 2016) have discovered novel ubiquitous phage species (Dutilh et al. 2014), a 'core' phage population amongst healthy individuals, and a correlation between disturbance of this core community and disease (Manrique et al. 2016). Regardless of the environment explored, the overwhelming majority of viral sequences produced – be it eukaryotic viruses or phages –exhibit no sequence homology to characterized viral species.

In stark contrast to eukaryotic and prokaryotic organisms, only a small fraction of viral – in particular phage – genomes have been sequenced and characterized. While metagenomics has been fruitful in unearthing gene markers and genomes of uncultivated eukaryotic and prokaryotic species (Hug et al. 2016), surveys of viromes face unique challenges (Bruder et al. 2016; Rose et al. 2016). First, unlike cellular organisms, there is no universally conserved gene in viruses. Viruses span a high degree of genetic diversity and are inherently mosaic (Hatfull 2008). Second, even when sequencing purified virions, sequencing data often includes non-viral (often host) DNA. This is further complicated by the fact that viral genomic DNA is often orders of magnitude less abundant than host cells or other organisms in the sample. Thus, tools have been developed to aid in distinguishing viral from non-viral sequences (Roux et al. 2015; Hatzopoulos et al. 2016; Yamashita et al. 2016). Third, extant viral data repositories do not include sufficient representation of viral species. Thus, bioinformatic tools for the interrogation of complex, bacterial communities have limited application in virome analyses.

The identification of viral genome sequences from relatively simple samples, i.e. samples containing a single or few viral species, is relatively straight-forward even in the presence of a large background of non-viral sequences. An example of such an inquiry would be the search for potential viral pathogens from clinical samples. Tools such as VIP (Li et al. 2016), VirAmp (Wan et al. 2015), and VirFind (Ho and Tzanetakis 2014) were designed specifically for such cases; they are, however, limited to the isolation of known viral taxa. Complex communities, particularly environmental samples present significantly greater challenges. While tools such as MetaVir (Roux et al. 2014) and VIROME (Wommack et al. 2012) have been developed to process (via homology to existing viral sequence databases) large and complex, virome data sets, they do not easily facilitate the isolation of genome sequences. To do this, alternative approaches have been employed. Typically, one of two approaches is taken. The first approach is founded on binning contigs from metagenomic datasets based upon their nucleotide usage profiles and/or contig coverage (see reviews Sharon and Banfield 2013; Garza and Dutilh 2015; Sangwan et al. 2016). The second, more frequently pursued method relies largely on manual curation. Several complete phage genomes have been mined from metagenomic data through inspection of sequences based upon their: size (Inskeep et al. 2013; Bellas et al. 2015; Ghai et al. 2016; Skvortsov et al. 2016), coverage (Dutilh et al. 2014; Bellas et al. 2015; Skvortsov et al. 2016); circularity (Dutilh et al. 2014; Bellas et al. 2015; Ghai et al. 2016), presence of sequence homology to annotated viral genes or genes of interest (e.g., terminases, structural proteins, portal proteins) (Sharon et al. 2011; Inskeep et al. 2013; Labonté & Suttle 2013a,b; Mizuno et al. 2013a,b; Nielsen et al. 2014; Bellas et al. 2015; Ghai et al. 2016; Paez-Espino et al. 2016; Skvortsov et al. 2016; Voorhies et al. 2016), and/or homology to CRISPR spacer sequences (Andersson & Banfield 2008; Inskeep et al. 2013). Similar searches have discovered novel eukaryotic viruses as well (e.g. Mokili et al. 2013; Schürch et al. 2014; Rosario et al. 2015; Liu et al. 2016). As highlighted in the recent report of the International Committee on Taxonomy of Viruses (ICTV) Executive Committee, genomes identified from metagenomic data will vastly expand our catalog of viral diversity (Simmonds et al. 2017).

Despite the bounty of sequence data ripe for mining, bioinformatics tools are presently a rate limiting step in analyses. For instance, assemblers vary in their efficiency in assembling viral genomes (Rihtman et al. 2016). Likewise analyses strategies can introduce unintentional effects (Smits et al. 2014, 2015).

III) Process

Herein we present the first software tool – virMine – for the automated analyses of viral metagenomic sequences for the discovery of viral genomes. This pipeline was specifically designed to accommodate the anticipated mosaicism and novelty present within viral communities in nature.

The pipeline integrates existing tools and new algorithms. Tools incorporated include Sickle (https://github.com/najoshi/sickle) for raw read quality control, SPAdes (Bankevich et al. 2012), metaSPAdes (Bankevich et al. 2012), and MEGAHIT (Li et al. 2015) for sequence assembly, GLIMMER (Delcher et al. 1999) for coding region prediction, and the BLAST+ suite (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). GLIMMER scripts have been modified to predict viral coding regions by increasing the threshold for gene overlaps and decreasing the threshold for the minimum gene length to 150 and 99 nucleotides respectively. These parameters are used for generating a training set; final predictions are made with a reduced overlap of 90 nucleotides. These parameter values were selected after testing GLIMMER's predictions for annotated phage genomes of various sizes (between 3Kbp and 300Kbp). Users can select to filter their assembly by a variety of parameters, including coverage which is evaluated using the BBTools package BBMap (https://sourceforge.net/projects/bbmap/) and the pileup function within SAMtools (Li et al. 2009). In addition to standard Python modules, the BioPython (Cock et al. 2009) library is also required. The pipeline classifies contigs, distinguishing non-viral, viral, and putative viral sequences. Development and testing was conducted on Ubuntu and MacOSX systems.

The pipeline includes two databases for distinguishing between non-viral and viral sequences. The first database includes amino acid sequences from non-viral sequences, collected from NCBI's RefSeq collection (ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Bacteria/). The second database includes sequences from the manually curated prokaryotic virus orthologous groups (Grazziotin et al. 2017) available at ftp://ftp.ncbi.nlm.nih.gov/pub/kristensen/pVOGs/home.html, and annotated RefSeq eukaryotic virus coding sequences, manually curated from ftp://ftp.ncbi.nlm.nih.gov/genomes/Viruses/.

IV) Results and Discussion

Figure 1 outlines the process employed by the virMine pipeline. A key aspect of the tool is its flexibility; it was designed to be modular, allowing users to access functionality individually or execute the full pipeline. Furthermore, to facilitate targeted analyses, filtration options and customization is available.

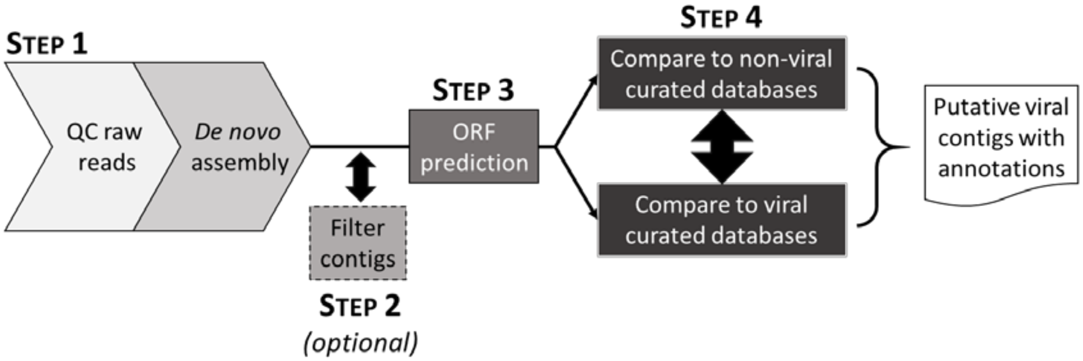
STEP 1
QC raw reads
De novo assembly
STEP 2
(optional)
Filter contigs
STEP 3
ORF prediction
STEP 4
Compare to non-viral curated databases
Compare to viral curated databases
Putative viral contigs with annotations

Fig. 1. Overview of virMine pipeline.

In the first step, raw sequencing data is evaluated using the QC tool Sickle. Reads are trimmed generating high quality data for assembly. Presently, the pipeline includes assembly by one of three methods: SPAdes, metaSPAdes, and MEGAHIT. These assemblers were selected as they represent a spectrum of rigor and expediency, including tools better equipped for assembly of low complexity samples (SPAdes) and those developed for complex metagenomes (metaSPAdes and MEGAHIT). In a prior study comparing tools for assembly of phage genomes from single or low complexity samples, the SPAdes assembler outperformed Velvet (Zerbino & Birney 2008) and Ray (Boisvert et al. 2010) and as such was selected for our pipeline. Because the assembly process is the most memory intensive step, the pipeline has been developed so that this first step can be executed independently of the remainder of the pipeline, which has modest resource needs. Thus Step 1 can be performed on a multi-core and/or large RAM system and exported to personal machines for execution of the remaining steps.

Step 2 is optional and permits the user to filter the contigs in the assembly based upon user-defined parameters. This can include some of the methods employed by previous studies mining for viral genomes, e.g. size, coverage, presence of genes or sequences (such as CRISPR spacer sequences) of interest. Users can specify a minimum and/or maximum contig length. Users can also specify a minimum coverage. In the case where a single viral isolate was sequenced, this option can easily distinguish between viral and non-viral (host and/or contaminant) sequences. Coverage is calculated by remapping the original reads to the contigs and the per contig coverage is calculated via BBMap. Coverage is not reported if this option is not selected. Alternatively, when SPAdes or metaSPAdes is used for assembly, users can select to use the SPAdes 'cov' value as a filter. Users can also filter contigs based upon sequences of interest by supplying these sequences of interest in a FASTA format file. Contigs are then BLASTed against this dataset. Results with a bitscore > 50 are considered 'real' hits and only contigs containing these hits will be considered further. It is worth noting that any of these filters can be selected by the user. Furthermore, the user can specify the order in which they are applied.

In Step 3, coding regions are predicted for each contig. Open reading frame (ORF) prediction is conducted using the tool GLIMMER (Delcher et al. 1999). Coding regions are predicted using GLIMMER trained to accommodate characteristics of viral genes, e.g. overlapping genes, short coding regions. (See the Methods for further details about parameters used for GLIMMER predictions.) The majority of viral genomes contain overlapping genes, a trend thought to have evolved to maximize protein production given physical constraints (Chirico et al. 2010). By relaxing the default values for the overlap and minimum gene length permitted during the GLIMMER search for coding regions, overlapping genes and small genes can be identified. It is important to note that this may, however, lead to an overestimation of ORFs within a genome as larger genes may be reported as multiple small or overlapping regions. This overestimation benefits downstream analyses.

In the final step, each predicted ORF is compared to two databases – a collection of non-viral sequences and known viral sequences. These two databases can be manually curated data collections, such as the ones supplied with the download, or can be user defined. Comparisons are facilitated via blastx. All hits are reported from both databases and the bitscores are compared to assess the likelihood that the ORF is viral. Here, the overestimation from Step 3 may assist in distinguishing between viral and non-viral origins as larger genes that are partitioned (e.g. a gene partitioned by the domains within its protein) can increase scores for viral components from, e.g., highly conserved protein domains. All ORFs predicted for a single contig are aggregated, producing a quantifiable score for each contig. Contigs predicted to be viral are written to file, as are their ORF predictions and blastx results. Contigs containing ORFs with no recognizable sequence homology to the viral database or non-viral database are also written to file, as these sequences may represent truly novel species.

V) Future Work

We have benchmarked our tool using synthetic datasets and will be presenting this work at the GLBio Conference May 15. We are in the process of testing this tool on real datasets and will be doing this during the summer. I submitted an abstract to the 2017 ISMB meeting (Prague, Czech Republic) and am waiting to hear if it's been accepted.

VI) Web Links

http://asquaredlab.weebly.com/

VII) Presentations and Publications

Garretto A, Putonti C. Identifying Viral Genomes within Complex Communities. 2017 Great Lakes Bioinformatics Conference (Chicago, IL): May 15-17. *Poster*

REFERENCES

1. Andersson AF, Banfield JF. Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. Science. 2008;320:1047–50.

2. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology. 2012;19:455–77.

3. Bellas CM, Anesio AM, Barker G. Analysis of virus genomes from glacial environments reveals novel virus groups with unusual host interactions. Frontiers in Microbiology [Internet]. 2015 [cited 2016 Oct 11];6. Available from: http://journal.frontiersin.org/Article/10.3389/fmicb.2015.00656/abstract

4. Berdjeb L, Pollet T, Domaizon I, Jacquet S. Effect of grazers and viruses on bacterial community structure and production in two contrasting trophic lakes. BMC Microbiol. 2011;11:88.

5. Boisvert S, Laviolette F, Corbeil J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. J. Comput. Biol. 2010;17:1519–33.

6. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. 2005;13:278–84.

7. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, et al. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. U.S.A. 2002;99:14250–5.

8. Bruder K, Malki K, Cooper A, Sible E, Shapiro J, Watkins S, et al. Freshwater Metaviromics and Bacteriophages: A Current Assessment of the State of the Art in Relation to Bioinformatic Challenges. Evolutionary Bioinformatics. 2016;25.

9. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, et al. Ocean plankton. Patterns and ecological drivers of ocean viral communities. Science. 2015;348:1261498.

10. Chirico N, Vianelli A, Belshaw R. Why genes overlap in viruses. Proceedings of the Royal Society B: Biological Sciences. 2010;277:3809–17.

11. Clokie MR, Millard AD, Letarov AV, Heaphy S. Phages in nature. Bacteriophage. 2011;1:31–45.

12. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.

13. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1999;27:4636–41.

14. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nature Communications [Internet]. 2014 [cited 2016 Aug 29];5. Available from: http://www.nature.com/doifinder/10.1038/ncomms5498

15. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, et al. Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. Appl. Environ. Microbiol. 2007;73:7059–66.

16. Ghai R, Mehrshad M, Megumi Mizuno C, Rodriguez-Valera F. Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. The ISME Journal [Internet]. 2016 [cited 2016 Oct 11]; Available from: http://www.nature.com/doifinder/10.1038/ismej.2016.110

17. Grazziotin AL, Koonin EV, Kristensen DM. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Research. 2017;45:D491–8.

18. Hatfull GF. Bacteriophage genomics. Current Opinion in Microbiology. 2008;11:447–53.

19. Hatzopoulos T, Watkins SC, Putonti C. PhagePhisher: a pipeline for the discovery of covert viral sequences in complex genomic datasets. Microbial Genomics [Internet]. 2016 [cited 2016 Aug 29]; Available from: http://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000053.v1

20. Ho T, Tzanetakis IE. Development of a virus detection and discovery pipeline using next generation sequencing. Virology. 2014;471–473:54–60.

21. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nature Microbiology. 2016;1:16048.

22. Hurwitz BL, Sullivan MB. The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. PLoS ONE. 2013;8:e57355.

23. Inskeep WP, Jay ZJ, Herrgard MJ, Kozubal MA, Rusch DB, Tringe SG, et al. Phylogenetic and Functional Analysis of Metagenome Sequence from High-Temperature Archaeal Habitats Demonstrate Linkages between Metabolic Potential and Geochemistry. Frontiers in Microbiology [Internet]. 2013 [cited 2016 Oct 11];4. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2013.00095/abstract

24. Jacquet S, Miki T, Noble R, Peduzzi P, Wilhelm S. Viruses in aquatic ecosystems: important advancements of the last 20 years and prospects for the future in the field of microbial oceanography and limnology. Advances in Oceanography and Limnology. 2010;1:97–141.

25. Labonté JM, Suttle CA. Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. Frontiers in Microbiology [Internet]. 2013 [cited 2016 Oct 11];4. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2013.00404/abstract

26. Labonté JM, Suttle CA. Previously unknown and highly divergent ssDNA viruses populate the oceans. The ISME Journal. 2013;7:2169–77.

27. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.

28. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

29. Li Y, Wang H, Nie K, Zhang C, Zhang Y, Wang J, et al. VIP: an integrated pipeline for metagenomics of virus identification and discovery. Scientific Reports. 2016;6:23774.

30. Liu Z, Yang S, Wang Y, Shen Q, Yang Y, Deng X, et al. Identification of a novel human papillomavirus by metagenomic analysis of vaginal swab samples from pregnant women. Virol. J. 2016;13:122.

31. López-Bueno A, Rastrojo A, Peiró R, Arenas M, Alcamí A. Ecological connectivity shapes quasispecies structure of RNA viruses in an Antarctic lake. Mol. Ecol. 2015;24:4812–25.

32. López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. High diversity of the viral community from an Antarctic lake. Science. 2009;326:858–61.

33. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. Healthy human gut phageome. Proc. Natl. Acad. Sci. U.S.A. 2016;113:10400–5.

34. Minot S, Bryson A, Chehoud C, Wu GD, Lewis JD, Bushman FD. Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. U.S.A. 2013;110:12450–5.

35. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, et al. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res. 2011;21:1616–25.

36. Mizuno CM, Rodriguez-Valera F, Kimes NE, Ghai R. Expanding the Marine Virosphere Using Metagenomics. Rocha EPC, editor. PLoS Genetics. 2013;9:e1003987.

37. Mizuno CM, Rodriguez-Valera F, Garcia-Heredia I, Martin-Cuadrado A-B, Ghai R. Reconstruction of Novel Cyanobacterial Siphovirus Genomes from Mediterranean Metagenomic Fosmids. Applied and Environmental Microbiology. 2013;79:688–95.

38. Mokili JL, Dutilh BE, Lim YW, Schneider BS, Taylor T, Haynes MR, et al. Identification of a novel human papillomavirus by metagenomic analysis of samples from patients with febrile respiratory illness. PLoS ONE. 2013;8:e58404.

39. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nature Biotechnology. 2014;32:822–8.

40. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, et al. Disease-specific alterations in the enteric virome in inflammatory bowel disease. Cell. 2015;160:447–60.

41. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth's virome. Nature. 2016;536:425–30.

42. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, et al. Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature. 2010;466:334–8.

43. Rihtman B, Meaden S, Clokie MRJ, Koskella B, Millard AD. Assessing Illumina technology for the high-throughput sequencing of bacteriophage genomes. PeerJ. 2016;4:e2055.

44. Rosario K, Schenck RO, Harbeitner RC, Lawler SN, Breitbart M. Novel circular single-stranded DNA viruses identified in marine invertebrates reveal high sequence diversity and consistent predicted intrinsic disorder patterns within putative structural proteins. Front Microbiol. 2015;6:696.

45. Rose R, Constantinides B, Tapinos A, Robertson DL, Prosperi M. Challenges in the analysis of viral metagenomes. Virus Evolution. 2016;2:vew022.

46. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ. 2015;3:e985.

47. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, et al. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS ONE. 2012;7:e33641.

48. Roux S, Tournayre J, Mahul A, Debroas D, Enault F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics. 2014;15:76.

49. Sharon I, Battchikova N, Aro E-M, Giglione C, Meinnel T, Glaser F, et al. Comparative metagenomics of microbial traits within oceanic viral communities. The ISME Journal. 2011;5:1178–90.

50. Skvortsov T, de Leeuwe C, Quinn JP, McGrath JW, Allen CCR, McElarney Y, et al. Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland. PLoS ONE. 2016;11:e0150361.

51. Smits SL, Bodewes R, Ruiz-Gonzalez A, Baumgärtner W, Koopmans MP, Osterhaus ADME, et al. Assembly of viral genomes from metagenomes. Front Microbiol. 2014;5:714.

52. Smits SL, Bodewes R, Ruiz-González A, Baumgärtner W, Koopmans MP, Osterhaus ADME, et al. Recovering full-length viral genomes from metagenomes. Frontiers in Microbiology [Internet]. 2015 [cited 2017 Feb 3];6. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2015.01069

53. Virgin HW. The virome in mammalian physiology and disease. Cell. 2014;157:142–50.

54. Voorhies AA, Eisenlord SD, Marcus DN, Duhaime MB, Biddanda BA, Cavalcoli JD, et al. Ecological and genetic interactions between cyanobacteria and viruses in a low-oxygen mat community inferred through metagenomics and metatranscriptomics: Cyanobacteria-virus interactions in a low-O 2 mat community. Environmental Microbiology. 2016;18:358–71.

55. Wan Y, Renner DW, Albert I, Szpara ML. VirAmp: a galaxy-based viral genome assembly pipeline. GigaScience [Internet]. 2015 [cited 2017 Feb 3];4. Available from: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/s13742-015-0060-y

56. Williamson KE, Radosevich M, Smith DW, Wommack KE. Incidence of lysogeny within temperate and extreme soil environments. Environ. Microbiol. 2007;9:2563–74.

57. Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, et al. VIROME: a standard operating procedure for analysis of viral metagenome sequences. Standards in Genomic Sciences. 2012;6:427–39.

58. Yamashita A, Sekizuka T, Kuroda M. VirusTAP: Viral Genome-Targeted Assembly Pipeline. Front Microbiol. 2016;7:32.

59. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol. 2007;5:e16.

60. Zablocki O, van Zyl L, Adriaenssens EM, Rubagotti E, Tuffin M, Cary SC, et al. High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. Appl. Environ. Microbiol. 2014;80:6888–97.

61. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.