Rough Diss Layout

Page sizes for the dissertation are to be A4. Dissertations must be typed, using font size 12, preferably in Times New Roman, and text is usually neater when 1.5 spaced. Note that only one side of a sheet should be used for text or illustrative material. The left margin should be 3.5 cm and a 2.5 cm right margin is recommended. All pages must be numbered. Preliminary pages (Title through to Lists of figures etc, see below) should be separately numbered using Roman numerals.

Blue = not sure if I am keeping.

Yellow = reword/fit better

Abstract - 400

Need some intro paragraph into the idea surrounding this research. Viral sequences note phage

Introduction – 1733

Hubert, et al (2009) discovered that there were 108 thermophilic bacterial spores m2, per year within arctic sea beds. This is unusual considering this environment differs extremely to their usual one. It is likely that the bacteria come from hydrothermal vent run off or “fluid flow”, in which they are carried from the vent into the ocean and remain dormant eventually entering the sea beds.

The bacteria from hydrothermal vents are considered extremophiles as the temperature can be in excess of 400°C are at depths in which hydrostatic pressure starts to take effect. These extremophiles thrive due to specialised features that they evolved over time, the viruses that depend on these bacteria for survival and propagation take advantage of these features. Recent studies have shown that these viruses are at high numbers at hydrothermal vents and also have high diversity however, their morphologies resemble that of archaeal viruses that have a lemon shape as well as a novel “spoon-shape” (Le Romancer, et al 2006). The viruses are also very host specific showing adaptations to help them infect their target hosts, one such adaptation is the ability to continue reproduction at lower temperatures then their hosts. However, the intricate details of these mechanisms are relatively unknown.

Oceanic viromes are very interesting as little is known about them compared to microbiomes and eukaryotic biomes, with every discovery leading to more questions and possibilities of research. One thing these viromes have in common are the two main proteins, endolysin and holin, which are used for cell lysis of their hosts. Meiring, et al (2012) sequenced the genome of the bacteriophage psymv2 that was taken from an Antarctic virome. Once sequenced it was found that the phage had no sequences similar to known endolysins however, did have a protein that showed the same characteristics. There where a N-terminal catalytic domain as well as a C-terminal cell wall binding protein domain. The holin proteins had a match suggesting that it would be a class II holin however, this protein is seen to be extremely diverse among phage and is likely to have a slightly different function or mechanism than what is suggested by the matched sequence. It is likely that these oceanic viromes and particularly deep-sea hydrothermal vent viromes differ greatly to surface viromes. In particular the taxonomy of viromes will differ as the viruses will have variations of proteins needed to overcome the extremes environments of their hosts as well as infect them.

The interplay between bacteria and viruses in the ocean is still being explored however, it is suggested that there is more than meets the eye. Bacteriophages follow a lytic or lysogenic life-cycle. The lytic cycle utilises host transcription mechanisms to produce viral proteins, these then assemble and are released causing the mortality of the host. The lysogenic cycle occurs when the phage genome is integrated into the host genome and replicated with the bacterial host. It is believed that lytic type phages are more common in oceanic waters (Hurwitz, Brum and Sullivan 2015) conversely, Williamson, et al (2008) showed that there were a number of lysogenic viruses present within hydrothermal vents which is unusual compared to surrounding waters. As previously mentioned, phage are able to reproduce at lower temperatures then the hosts optimum thus lysogeny can occur when suboptimal conditions for the host survival/growth are met (Erikson 2014; Williamson, et al 2008). It’s likely that novel genes found in viral genomes are advantageous to the virus due the selection pressures found within hydrothermal vents. If these genes help the survival of the virus then it is possible that they also benefit and increase the bacterial hosts survival.

It is thought that bacterial and viruses in hydrothermal vents have a near symbiotic or quasi-stable genetic relationship. One example of this is that when viruses infect hosts the viral genes can actually participate in the metabolism of the host in various ways. Within three different sediment samples both viruses and micro-organisms showed similar gene profiles in which metabolic genes for energy production and conversion where the most abundant. To see if viral genes participated within microbial metabolism the genes were aligned to the KEGG pathway. It was found that they did participate in microbial metabolism and actually compensated in six pathways with unique genes. On of these paths, two-component regulatory system, was of particular interest as it is responsible for responding to the change of conditions within the environment. This indicates that viruses infecting host cells not only increase the absorption of nutrients by the host by helping in metabolic pathways and increase overall survivability of their hosts within hydrothermal vents. This in turn benefits the virus as it can utilise the host for replication over a longer period of time (He, Li and Zhang 2017). It is likely that other mechanisms like this are present within viral genomes and need to be found via metagenomic studies.

Previously wet-lab approaches were the only way to investigate viruses and their genomics, these approaches are time consuming as well as expensive. The biggest disadvantage this approach has is that the information gained is based on inference or was implied from other sources. However, in 2001 an experiment conducted by Allander et al, gave way to the basis of current viral metagenomics. Then in 2005 a large-scale high-profile sequence of the Sargasso Sea was completed. This in turn paved the path for many viral experiments of marine ecosystems in water, sediments and hydrothermal vents (Greninger 2018). Due metagenomics having numerous successes such as being recognised by the International Committee for Taxonomy of Viruses (ICTV), it has gained traction within the scientific community as being the go-to resource for viral studies (Koonin and Dolja 2018).

As metagenomics gained traction the tools that allow the analysis of this type of data did as well. Before this however, most studies consisted of genomic tools and thus should be mentioned as they gave rise to metagenomic tools. Prophinder (Lima-Mendez, et al 2008) that detects viral sequences via comparison to known viral genes, tRNA and attachment site sequences recognition. Like Prophinder, PhiSpy (Akhter, Aziz and Edwards 2012) detects viral sequences in the same way but with the added complexity of recognising AT and GC skews in sequences as well as transcriptional direction. PHAST (Zhou, et al 2011) is arguably the simplest viral sequence finding tool using similar technology to Profinder however, what sets this tool apart is that the authors of the tool also created PHASTER (Arndt, et al 2016), a metagenomic variation of PHAST. PHASTER works in the same way as PHAST but on a different type of data, FragGeneScan is also built into the tool allowing recognition of partial genes (Hurwitz, et al 2018). In a study by Freitas, et al (2017) PHASTER was used to identify prophage sequences within a draft genome of the deep-sea bacteria *Moritella* sp. JT01. The tool was able to predict two intact prophages and one incomplete prophage. However, there was no significant different between the mean values of prophage sequences per kilobase of the tested groups. One interesting thing that PHASTER flagged in this study is an increased amount of incomplete and questionable sequences within JTo1 compared to other samples. PHASTER was able to help in not only identifying prophage sequence but also biotechnologically relevant genes.

Apart from PHASTER there are two other main tools used for metagenomic data analysis, these are VirSorter (Rox, et al 2015) and VirFinder (Ren, et al 2017). VirFinder utilises a machine learning approach to finding viral sequences which is based on alignment free kmer signatures. This a new approach to metagenomic data as training sets used undergo logistic regression modelling as well as lasso regularisation. This in turn allows for higher prediction accuracy of sequences (Ren, et al 2017). VirSorter is known as the “current state of the art” classification tool for viral sequence prediction from metagenomic data. This is due to the complexity and accuracy of the tool when comparing a given sequences to known genomic and metagenomic bacterial and viral genes. The prediction is generated based on factors such as viral hallmarks being present and the depletion of Pfam affiliated genes. Unknown viral contigs can also be analysed using this tool, if this occurs in a data set then enrichments values of those sequences with no homologs are compared (Hurwitz, et al 2018; Ren, et al 2017).

After 18 years finding novel phage sequences still remains a difficult task however, there is a multitude of tool that have been developed and continue to be improved in order to tackle this issue. There are many approaches taken to find viral sequences within metagenomes such as homology-based searches and kmer based approaches which are computationally faster compared to homology searches. Although there are many approaches the biggest task these tools face is how auxiliary metabolic genes (AMGs) blur the line between viral sequences and host sequences (Hurwitz, et al 2018). This occurs as the AMGs are viral sequences but originated in the bacterial genomes, these genes mediate the hosts metabolism during infection to ensure there is constant energy for replication. However, in these extreme environments these AMGs can benefit the host as well, one example of this is the protein NifA-Like coded by phage that takes part in nitrogen fixation and metabolism.

A microbial metagenome was constructed from samples collected from an experimental site in Svalbard, Norway. This location was chosen as it is one of the most northern point that is realistically reachable without needing specialised equipment such as an ice breaker vessel. Therefore, it is highly likely that hydrothermal bacteria from vent run off would be located in sediments in this region. The samples where sediments collected at three different depths to ensure that hydrothermal bacteria where collected but also to see if there are any differences between the depths. Once extracted the samples were heat treatmed to temperatures exceeding 50°C to ensure that only thermophilic microorganisms where present. This data was then sent off to [cannot for the life of me remember] and a shotgun metagenome was formed for each depth, these where named CoDL\_16, CoDL\_17 and CoDL\_19 with each one containing an R1 and R2 version for each read.

This paper aims to evaluate currently available metagenomic tools, explore the metagenomic samples from Svalbard for viral sequences using these tools and to identify possible AMGs leading to a better understanding of hydrothermal vent microbiomes and viromes.

Material and methods – 1232

As there are many metagenomic tools currently available with more continually being made to serve specific purposes and research, only a limited number could be chosen for evaluation. In order to choose what tools would be evaluated they needed to be able to detect phage sequences from microbial metagenomic data, be accurate and have the ability to be used in future research. All of the tools used meet these criteria, but each also has unique properties making it different to the other tools chosen. PHASTER (Arndt, et al 2016) was the first chosen, it is also one of the first tools created as its predecessor PHAST detected viral sequences using genomic data. VirSorter (Rox, et al 2015) was chosen as it is the go-to tool for these types of experiments as exploration. A tool known as MARVEL (Amgarten, et al 2018) was chosen due to some recent papers suggesting that it may have more accurate predictions and true positives then VirSorter. Finally, a relatively new tool PhageWeb (De Souse, et al 2018) was used as not much is known about it and thus could be very accurate in its predictions of viral sequences.

In order to evaluate these tools accurately and fairly a sample data set needed to be chosen and used on each of the tools. For this a subset of unaligned metagenomic data from Svalbard was used however, on testing this small random subset none of the tools detected any viral sequences. Therefore, a selection of samples from iMicrobe where used, samples was chosen from this source as they have been submitted from various published experiments and papers. The key factor in choosing these sources is that they contained metagenomic data from Arctic and Antarctic oceans as well as hydrothermal vents, meaning the data used for evaluation was similar to the data collected from Svalbard. The samples used for this evaluation came from a project named “Botany Bay Metagenomics” in which a metagenomic analysis of Cymbeastela concentrica sponge bacteria found novel functional properties. Three samples were randomly chosen (BOTANYBAY\_SMPL\_BBY01, BOTANYBAY\_SMPL\_BBY04 and BBAY33-4F-01-691) and trailed on each tool.

After considering the pros and cons of each of the tools from the preliminary evaluation as well as getting a feel for how each tool worked, it was decided that VirSorter would be the tool used for further analysis. To justify the use of this tool it was trailed on two more data sets from iMicrobe. The first data set was from a project that looked into an Antarctica aquatic microbial metagenome (AAMM), this was chosen as it would be a similar to the metagenome from Svalbard as it will contain many aquatic microbial genomes from extreme environments that may have viral sequences. A random selection of samples was chosen from this project, if the sample contained a specific depth all samples from that site where chosen. For example, if site 227 0.1 µm was selected samples site 227 0.8µm and site 227 3µm where also chosen. The second data set was from a project named “Moore marine phage/virus metagenomes” (MMP), this was chosen as it would directly look at viral genomes and how well VirSorter can identify these sequences. Unlike the first data set all of the samples were chosen as there was not a big sample size. Any potential viral sequences that VirSorter predicted where entered into BLAST (Altschul, et al 1990) in order to see if there were any similar known sequences or if these predicted sequences are novel. If the sequences did match similar sequences, the one with the highest overall score and similarity was noted, this sequence was then looked into further to see if there were any interesting proteins the viral sequences could potentially code for. Before the metagenomic data from Svalbard was used with VirSorter, it needed to be assembled. Prior to assembly taxonomic and functional analysis of the metagenome was conducted in order to grasp a better understanding of the data.

To complete the pre-assembly analysis of the Svalbard data a pipeline was created. To ensure the pipeline worked a subsample of the dataset was created using a simple sub-setting data command. Once these subsamples had been created, they were cleaned to ensure the highest possible quality of data. This was achieved by utilising Trim Galore! (Krueger 2012) A tool used to remove both low-quality bases and adapters from the end of both reads in each read pair. The quality score was set to 20 as it is the default score and stringency set to four to ensure quality was being kept but not a huge amount of data would be lost in the process.

Once the data had been cleaned a taxonomic analysis could be conducted. The first step in this analysis was to determine the species composition using Kraken (Wood and Salzberg 2014), which classifies short DNA molecules with taxonomic labels. Due to computational resources being limited and a restriction on what was possible during the time frame of the project a smaller database of kmers was used via MiniKraken 8GB. As Kraken can use both paired reads and single reads, the “—paired” command must be specified as the data contained paired reads, R1 and R2.

The output report from Kraken is then formatted using Krakens translate function, this formatted output is then converted to a file type that can be visualised via Krona (Ondov, Bergman, and Phillippy 2011). The most important command used in this conversion step is “uniq -c” as it finds every unique entry in the data and counts how many of each there are, giving an abundance of each species within the data. The visualisation that Krona produces is interactive allowing for in depth analysis as you can select specific areas of the multi-layered pie chart and enhance it to get a better understanding of that data. The full interactive Krona output can be found within the github repository linked in the appendix.

Finally, a functional profile of the metagenomic data can be built. There was two possible ways of completing this analysis the first comparing the metagenomic samples of different depths, the second looking at all the samples as a whole. For this project it was decided that comparing the depths would be the best option. A package known as HUMANn2 (Franzosa, et al 2018) is used for building this functional profile, as part of the process MetaPhlAn2 is run in order to identify what species are present so a tailor-made database of appropriate genes is generated to map against. Alignment against the gene database and the protein database will be conducted, leading to HUMANn2 computing which gene families are present to distinguish which pathways are present in the data.

Next a statistical comparison was conducted to compare the samples, as they are no very informative on their own. The outputs from HUMANn2 where combined to make one table where each column represents a sample. Before the data can be visualised, it needs to be renormalized as the abundance was only normalised *within* each sample not across all samples. During this stage all unmapped or unassigned values were also removed from the table. The data was then visualised to see if there is any statistical significance in the samples and if any correlations are present. To put together a more informative analysis of the data a statistical test can be done to find any differences that are significant statistically. This can be achieved via multiple methods, the best test for this data was via a Kruskal-Wallis H-test.

[all scripts can be found in the appendix].

Results – 791

Due to each tool using different prediction algorithms and being built on different technologies all have different capabilities and thus made the evaluation more difficult.

All of the samples used for evaluation showed at least one predicted viral sequence for each tool evaluated however these predicted sequences varied among all the tools. It’s unclear how accurate these predictions are without any other information alongside the predicted sequences. Due to this accuracy was not considered when evaluating these tools as no “true” wet lab genomes exist for the viruses within the samples as the viral sequences present are unknown. The exception to this was VirSorter as there are many published papers such as [XXX, XXX] that state how accurate this tool is and why it can be considered the go to resource.

***Table 1***. Tools used in an initial exploratory analysis in order to determine the pros and cons of each and which would be used for further analysis of the full metagenomic data set from Svalbard.

|  |  |  |
| --- | --- | --- |
|  | Advantages | Disadvantages |
| PHASTER | Fast.  Specific to prophage sequences.  Annotates the sequences in which it finds.  Terminal version available.  Allows multiple sequences to be run simultaneously.  Uses FragGeneScan for analysis of partial genomes. | Terminal version is built on old technology and can be hard to find.  Online version only allows a 40MB file upload.  File need to be in a specific format. |
| PhageWeb | Vast database to call upon.  Any file size accepted.  New meaning its fairly fast as not many use it reducing bottle necks.  Allow for terminal use.  Edit of parameters. | Time consuming during upload of data.  Finicky.  Only specific file types accepted.  Not widely used or implemented. |
| MARVEL | Arguable most specific tool in terms of findings due to use of Random Forest machine learning.  Test data included 1247 viral and 1029 bacterial genomes.  Nice output folder allowing for further analysis.  Updated fairly regularly.  Uses three features in prediction: density of genes, frequency of strand shifts and the fraction of genes that have a significant match to the pVOGs database. | Hard to set up as it needs multiple dependencies.  Barrier to entry as there can be issues with data entry and crashes. |
| VirSorter | Accurate, considered the gold standard by many papers.  Can be used via an online server Cyverse allowing for faster run times compared to local machines.  Finds numerous viral sequences.  Very user-friendly interface and logs are available to see what has occurred.  Shared data is findable, usable and often published.  Predictions based on viral hallmarks, depletion of Pfam affiliated genes and comparison to known genes. | Can find many novel sequences that do not lend themselves to further analysis.  Relies on a database.  In order to upload your own data another application is needed. |

When the data sets AAMM and MMP where input in to VirSorter, a combined total of 418 viral sequences where predicted. Just over 90% of these sequences came from AAMM which is a microbial. study (377 sequences) rather than MMP which is a viral study and accounted for under 10% of the total (41 sequences). Therefore, VirSorter was able to predict more viral sequences from samples that contained a mix of microbial genomes rather than samples that contain just viral genomes.

There was a total of 12 samples taken from the AAMM study to analyse with VirSorter. All of these samples except site 227 3µm contained at least one category 1 predicted sequence viral sequences however, most of the predicated sequences where category 2 (356 sequences). The difference between the categories is minimal but there is more room for error within category 2 and thus may contain some false positives. As previously stated, all of the samples from the MMP study where used due to there only being eight samples. Of these samples two, sample 720 and sample 834, had no predicted viral sequences, with the remaining samples containing at least one category 1 predicted sequence. The distribution between categories was reduced within this study with 15 category 1 sequences predicted and 26 category 2 sequences predicted.

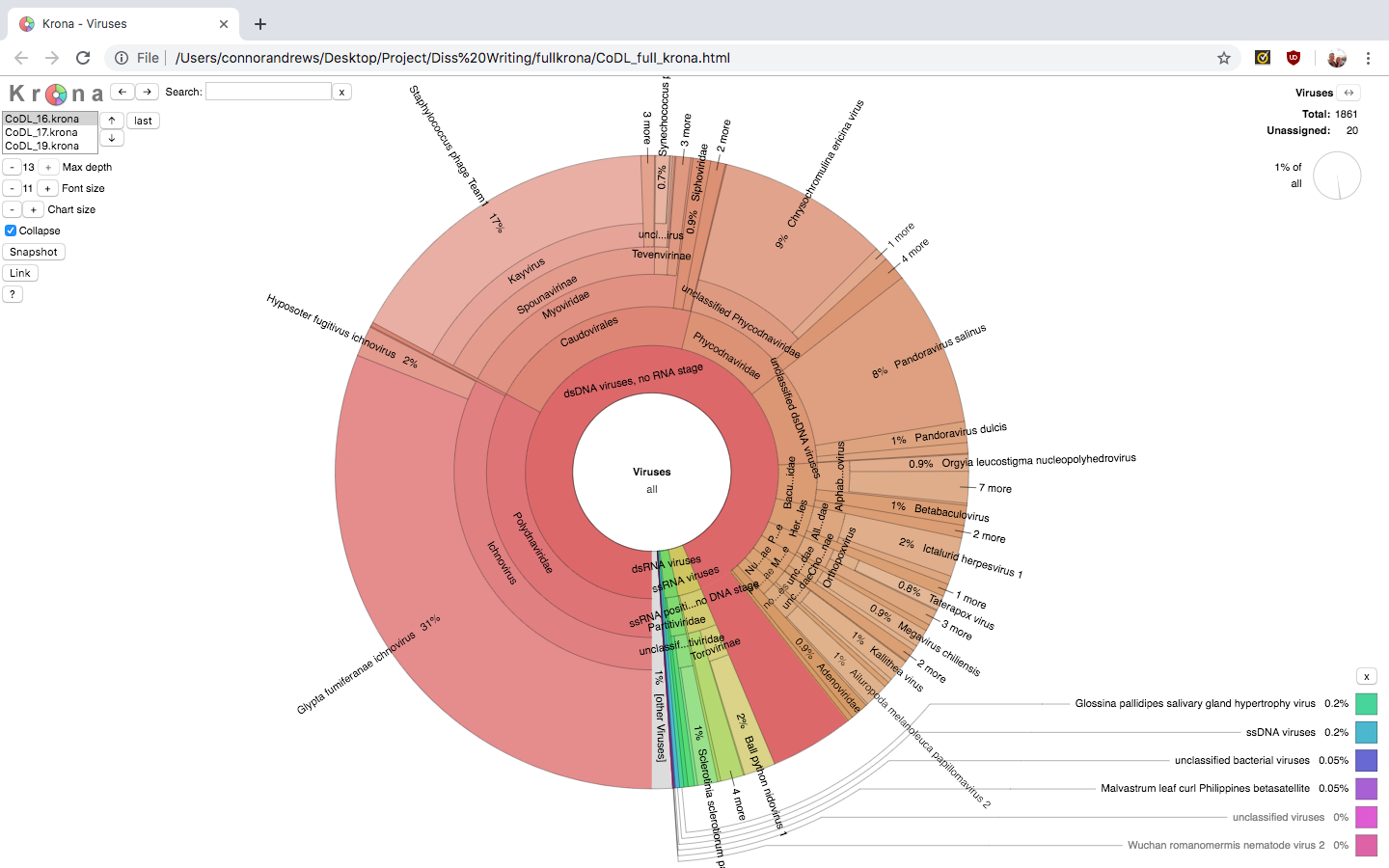
***Table 2***. Number of viral sequences predicted by VirSorter from the studies AAMM and MMP. Further broken down into the amount of novel sequences and sequences found to have a documented similar sequence from BLAST.

|  |  |  |
| --- | --- | --- |
| Study | AAMM | MMP |
| Total Predicted Sequences | 377 | 41 |
| Novel Sequences | 307 | 29 |
| Similar Sequences | 70 | 12 |

As seen in table 2. there is a much higher proportion of sequences that have no similar sequences compared to those that have similar sequences via a BLAST search. It is well known that viral genomes are researched and sequenced less then eukaryotic and bacterial genomes. This comparison is also true for marine viruses to those that would reside within eukaryotes, as there is a bias to research those as they can affect the health of humans [This might be more of statement for the discussion rather than at this point as it can link into why virsorter has a tendency to find novel genes rather than those that have already been discovered.]

The sequences that showed no similarity are likely to be novel as they have yet to be discovered via any other means. However, it is key to remember that these are prediction and would need further analysis via programmes such as NCBI’s Sequence Viewer (NCBI, 2019) or EMBL EBI’s Transeq (Maderia, et al 2019) to see if there are plausible proteins. VirSorter shows an interesting tendency to predict sequences that are novel this can be seen in table 2. where approximately 80% of the total predicted sequences had no similarity and are considered novel. Of the ~20% of sequences that did have similarity many coded for hypothetical proteins or resided in untranslated areas. However, those that coded for proteins had a tendency to be involved in metabolism, viral replication and cell signalling. It is likely that the sequences coding for metabolism proteins are AMGs as viruses do not have their own inherent metabolic pathways. Examples of some interesting proteins of non-viral origin and possible AMGs that where found are Rubrerythrin, Zinc binding protein deoxycytidylate deaminase domain, AAA+ family replicate ATPase and NifA-like protein.

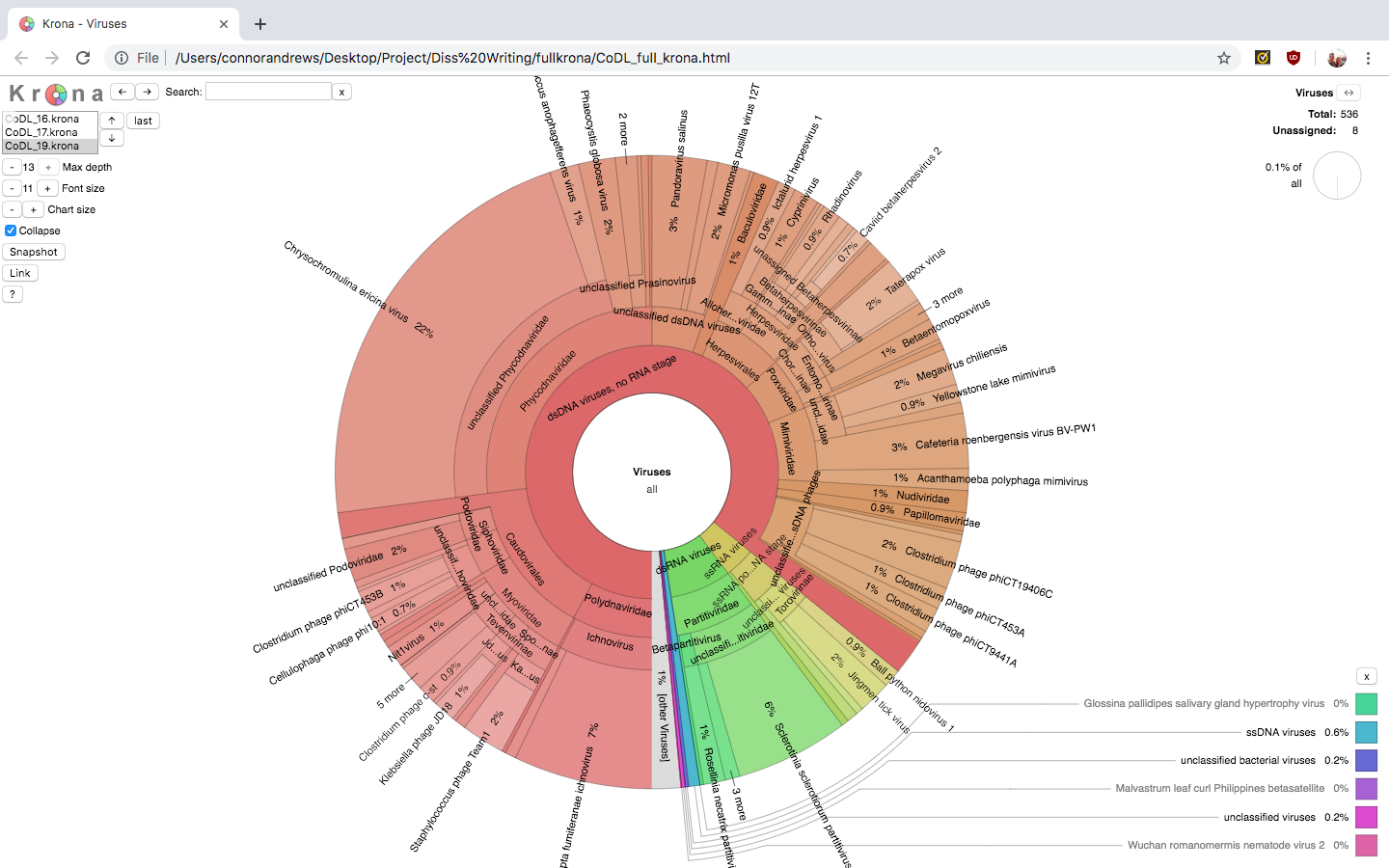
After analysing the species composition of the samples via Kraken, Krona gives an interactive output of taxonomic abundance as seen in figures 1-3. CoDL\_16 had 1778290 total assigned fragments, the lowest of all the samples. CoDL\_17 had 543200 total fragments which was the most of all samples and finally CoDL\_19 had 463051 total assigned fragments. Although bacteria are clearly more abundant in all of these samples, the variation in the bacteria is not as vast as what is seen within the viral taxonomic groups. For example the sample CoDL\_16 contains many double stranded DNA viruses that have no RNA stage (94% of all viruses) but also contains single strand RNA viruses (3%), double stranded rnda viruses (2%), single strand DNA viruses (0.2%), unclassified bacterial viral particles (0.05%) and interestingly a Glossina pallidipes salivary gland viruses (0.2%). Whereas the bacteria are mainly comprised of proteobacteria (41%), terrabacteria (30%) and unassigned bacteria (20%).



***Figure 1***: Viruses present within metagenomic data CoDL\_16. The viruses within this depth account for 1% (1861) of the total taxonomic abundance however, there is a wide array of different viral families present.



***Figure 2***: Viruses present within metagenomic data CoDL\_17. The viruses only account for 0.2% (1201) of the entire abundance but have a much higher variety then the bacteria present.



***Figure 3***: Viruses present within metagenomic data CoDL\_19. The viruses present account for 0.1% (536) of the total taxonomic abundance which is the lowest of metagenomic samples. However, this depth contains the most variation of viruses.

Functional profile

* Figure of the profiles either a PCA type graph of a heatmap.
* What genes are being shown to be active
* Difference between the depths indicate.

Discussion – 3000/4000

Discussion points

Tools

* Limitations of the tools, what tweaks can be made.
* My “story” and experiences with using the tools.
* Where I see tools leading in the future, linking to the thought that tools will continually be made for specific research goals/papers rather than centralising around one or a few tools and using those with continual updates and improvements.

VirSorter

* VirSorter has a skew to its predictions. The tool much more specialised to finding viral sequences within microbial/bacterial metagenomes rather than just viral sequences within a viral metagenome.
* This is due to the way the tool was built and the purposes for what it was built.
* Tends to also skew to cat 2 – less sure then 1 but still very accurate and high probability. Further explain this point and possibilities why.
* Want to find viral infections within a microbial genome. Allows us to pick out viral sequences. Use those for further analysis or removal from a draft genome of an organism that wants to be published?
* Build a viral genome based on the sequence extracted.
* References about the building of Viroster, microbial vs viral sequences found, paper with genome built using if possible?

AMGs

* Start with BLAST – reasoning (or should this be added to methods?) the sort of outputs it would give and what that meant for the predicted viral sequences.
* Quick statements on Hypothetical proteins and untranslated, what can be inferred by so many being found. What they actually are and what they may mean for the viral sequence and bacteria.
* Specific examples if possible.
* Those that I found particularly interesting.
* Rubrerythrin. What this protein is and does, likely to originate from virus or bacteria, AMG yes or no, how it is advantageous to code for, does it help the bacteria.
* Zinc binding protein. What this protein is and does, likely to originate from virus or bacteria, AMG yes or no, how it is advantageous to code for, does it help the bacteria.
* AAA+ family replicate ATPase. What this protein is and does, likely to originate from virus or bacteria, AMG yes or no, how it is advantageous to code for, does it help the bacteria.
* NifA-like. What this protein is and does, likely to originate from virus or bacteria, AMG yes or no, how it is advantageous to code for, does it help the bacteria.
* AMGs found in Svalbard data if there is any present or this stage is achieved.

Taxonomy

* Break down of the kinds of viruses present as a whole, are phages present or free-living viruses. Oceanic vs sediment vs elsewhere (e.g. eukaryotic infecting).
* Comparison of the depths and what this tells us.
* Any inference that can be made about possible viral sequences, infectious paths, plausible hydrothermal viral particles.
* Why there is a large proportion of dsDNA viruses compared to others. Same with families?

Functional profile

* What the functional profile is showing. Why are these particular genes more active, is there a pathway that links the increased function?
* What can be inferred from the differences seen in the profiles, are the depths important to viral activity, is there any plausible viral activity shown.
* Will add to this depending on the results that have been gathered.

Link into conclusion

* How what was found could possibly influence further research.
* Tools what can be done in the future, is there a go too, things to consider when picking one and any future tweaks that could be made.
* AMGs – what these mean for research, how to detect and possible implications.

Conclusion – 500

This is all just notes for now.

Although metagenomic tools are making strides in discovery of novel sequences and allowing us to better understand the interactions between phage and hosts there is still improvements that can be made. One of the main difficulties is the ability to distinguish between bacterial and viral sequences.

This is due to the interaction of genomes during infection especially if AMGs are present within the viral portion. These AMGs as well as other genes are often identified as bacterial rather than viral leading to a skew in the data that is presented although this is an issue when presenting data it can also have benefits, such as giving clues as to how hosts adapt to their environments, nutritional constraints they face and metabolic bottle necks in which phage proteins will help compensate.

I believe the main focus of metagenomic tool improvements is looking into computational capacity as this is the most constraining element on these tools. Another improvement currently being looked into is how to use multiple reference database to ensure the validity of results. Currently the best way to overcome these issues is to use multiple tools in conjunction choosing tools that meet the user’s requirements, when doing this it is preferable to use a pipeline mechanic to obtain the best results.

References

Akhter, S., Aziz, R. K., and Edwards, R. A., 2012. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity and composition based stratagies. *Nucleic Acid Research*, 40(16), pp. e126.

Allander, T., Emerson, S. U., Engle, R. E., Purcell, R. H. and Bukh, J., 2001. A virus discovery method incorporating DNase treatments and its application to the identification of two bovine parvovitus species. *Proc. Natl. Acad. Sci.*, 98, pp. 11609-11614.

Algren, N. A., Fuchsman, C. A., Rocap, G., and Fuhrman, J. A., 2019. Discovery of several novel, widespread and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. *ISME J*, 13(3), pp. 618-631.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J., 1990. Basic local alignment tool. *Journal of Molecular Biology*, 215, pp. 403-410.

Amgarten, D., Braga, L. P. P., da Silva, A. M., and Setubal, J. C., 2018. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front Genet*, (9) pp. 304.

Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D. S., 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acid Research*, 44(W1), pp. W16-W21.

Bellas, C. M., and Anesio, A. M., 2013. High diversity and potential origins of T4-type bacteriophages on the surface of Arctic glaciers. *Extremophiles*, 17(5), pp. 861-870.

Erikson, J., 2014. *Viruses hijack deep-sea bacteria at hydrothermal vents [online].* Michigan news: University of Michigan. Available at: <https://news.umich.edu/viruses-hijack-deep-sea-bacteria-at-hydrothermal-vents/> [Accessed 21st March 2019].

Franzosa, E. A., Mclver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., and Huttenhower, C., 2018. Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15, pp. 962-968.

Freitas, R. C., Odisi, E. J., Kato, C., de Silva, M. A. C., and Lima, A. O. S., 2017. Draft genome sequence of the deep-sea bacterium *Moritella* sp. JT01 and identification of biotechnology relevant genes. *Marine Biotechnology*, 19(5), pp. 480-487.

Greninger, A. L., 2018. A decade of RNA virus metagenomics is (not) enough. *Virus Res*, 244, pp. 218-229.

He, T., Li, H., and Zhang, X., 2017. Deep-Sea hydrothermal vent viruses compensate for microbial metabolism in virus-host interactions. *MBio*, 8(4).

Hubert, C., Loy, A., Nickel, M., Arnosti, C., Baranyi, C., Bruchert, V., Ferdelman, T., Finster, K., Christensen, F. M., Rosa de Rezende, J., Vandieken, V., and Jorgensen, B. B., 2009. A constant flux of diverse thermophilic bacteria into the cold Arctic seabed. *Science*, 325(5947), pp. 1541-1544.

Hurwitz, B. L., Brum, J. R., and Sullivan, M. B., 2015. Depth-stratified functional and taxonomic niche specialization in the ‘core’ and ‘flexible’ Pacific Ocean. *ISME J*, 9(2), pp. 472-484.

Hurwitz, B. L., Ponsero, A., Thornton, J., and U’Ren, J. M., 2018. Phage Hunters: Computational strategies for finding phages in large-scale ‘omics datasets. *Virus Res*, 244, pp. 110-115.

Jalasvuori, M., Jaatinen, S. T., Laurinavicius, S., Ahola-Iivarinen, E., Kalkkinen, N., Bamford, D. H., and Bamford, J. K., 2009. The closest relatives of icosahedral viruses of thermophilic bacteria are among viruses and plasmids of the halophilic archaea. *J Virol*, 83(18) pp. 9388-9897.

Koonin, E. V., and Dolja, V. V., 2018. Metaviromics: a tectonic shift in understanding virus evolution. *Virus Res*, 246, pp. A1-A3.

Krueger, F. 2012. *Trim Galore [online].* Babraham Bioinformatics. Available at: <http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/> [Accessed 18th May 2019].

Le Romancer, M., Gaillard, M., Geslin, C. and Prieur, D., 2006. Viruses in extreme environments. *Life in extreme environments*, pp. 99-113.

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R., 2008. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics,* 24(6), pp. 863-865.

Maderia, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., and Lopez, R., 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acid Research*, 47(W10 pp. W636-W641.

Meiring, T. L., Tuffin, I. M., Cary, C., and Cowan, D. A., 2012. Genome sequences of temperate bacteriophage Psymv2 from Antarctic Dry Valley soil isolate Psychrobacter sp. MV2. *Extremophiles*, 16(5), pp. 715-726.

NCBI, 2019. *Sequence Viewer [online].* National Center for Biotechnology Information. Available at: <https://www.ncbi.nlm.nih.gov/projects/sviewer/> [Accessed 15th June 2019].

Ondov, B. D., Bergman, N. H., and Phillippy, A. M., 2011. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12(385).

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

Rox, S., Enault, F., Hurwitz, B. L., and Sullivan, M. B., 2015. VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3.

Smeele, Z. E., Ainley, D. G., and Varsani, A., 2018. Viruses associated with Antarctic wildlife: From serology-based detection to identification of genomes using high throughput sequencing. *Virus Res*, 243, pp. 91-105.

de Souse, A. L., Maues, D., Lobato, A., Franco, E. F., Pinherio, K., Araujo, F., Pantoja, Y., da Costa da Silva, A. L., Morais, J., and Ramos, R. T. J., 2018. PhageWeb – web interface for rapid identification and characterisation of prophages in bacterial genomes. *Frontiers in Genetics*, 9.

Williamson, S. J., Cary, S. C., Williamson, K. E., Helton, R. R., Bench, S. R., Winget, D., and Wommack, K. E., 2008. Lysogenic virus-host interactions predominate at deep-sea diffuse-flow hydrothermal vents. *ISME J*, 2(11), pp. 1112-1121.

Wood, D. E., and Salzberg, S. L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(R46),

Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S., 2011. PHAST: A fast phase search tool. *Nucleic Acid Research*, 39(2), pp. W347-W352.

Appendix

All of the scripts will also be available on my github <https://github.com/C-And/Viral-Metagenomic-Thesis>

Place holder text