Metatranscriptomic Data Analysis

Finding modules of co-expressed genes in *Crassvirales* bacteriophages

Bryan van den Brand [1], Daniel Carrillo Bautista [1] and Bas Dutilh [1]

[1] Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, 3584 CH Utrecht, The Netherlands

Abstract

Motivation: The *Crassvirales* are a viral order that includes an abundant group of bacteriophages in the human gut. First described in 2014, most *Crassvirales* phages have been found infecting members of the Bacteroidetes phylum. As with most other viruses, the functions of many proteins encoded in *Crassvirales* genomes are not described. Here we hypothesise that genes that co-express tend to participate in the same biological process and might have a related function. Through metatranscriptomic data analysis we described in detail co-expression of *Crassvirales* genes, grouping the gene expression patterns in modules of closely related genes. For this, we apply new methods and machine learning algorithms that are capable of analyzing the minute differences in gene expression found in the metatranscriptomic data originating from wastewater and human gut samples. We demonstrate that gene modules can be found, grouping genes spread over the *Crassvirales* genomes into putative functional clusters, with one excellent result for the *Crassvirales* genome OLQR01000043. The quality of the gene modules found is strongly influenced by the amount of samples usable for a genome. In the future applying the approach to a different viral order with a larger dataset would allow for verification of the approach and its results.
Availability: https://github.com/Bryan-vd-Brand/MetaTranscriptomic-Gene-Grouping
Contact: bryan.v.d.brand@gmail.com

1 Plain language summary

In this paper we wish to further explore one of the most abundant groups of viruses infecting bacteria in the human gut, the *Crassvirales* bacteriophages.As for many other viruses, the function of most of the genes encoded on the DNA of these phages remains unknown. After infection of bacteria, genes encoded in the viral DNA are transcribed into messenger-RNA (mRNA) which is then translated into proteins that will carry out different functions in the phages life cycle. By measuring the levels of *Crassvirales* mRNA present in human gut samples we can know what genes of these phages are actively transcribing. We hypothesize that genes with a similar pattern of transcription, for example gene x has a high level of transcription than gene y also has a higher level of transcription, tend to participate together in the same biological process and might have a related function. By taking many samples from different locations and measuring the mRNA levels for each of the *Crassvirales* genes we can analyze the differences between the samples and discover patterns, showing what genes tend to transcribe together. In order to find these patterns we apply a machine learning algorithm known as, Independent Component Analysis (ICA). ICA allows us to input all the measured mRNA levels and output values representing the strength of a gene in a pattern. Because of the random initialization of ICA, we ran this algorithm one thousand times to increase the confidence of the gene modules of co-transcribed genes. From the 260 human gut samples analyzed in this study, we found 33 gene modules in 6 different *Crassvirales* species. These modules are of different sizes and are distributed over the species genomes. The gene module with the highest confidence score was found in the genome OLQR0100043 and consists of 4 genes, 3 of them with functions already described and related to the structure of the virus particle, and one gene of unknown function. Considering its close position to the other genes in the module and the highly similar functions of the other 3 genes it is likely that the unknown gene function is involved in the same biological process. Altogether, we have developed a method that is able to relate genes based on their expression patterns using transcriptomic data and machine learning procedures. Noticeably, this method can be applied to any microorganism and benefits from a high number of samples.

2 Introduction

The *Crassvirales* order is among the most abundant and ubiquitous group of viruses in the human gut. The prototypical member of this order, the crAssphage was discovered in 2014 after cross-assembly of human fecal metaviromes (Dutilh *et al.*, 2014). In that study the phage was found in all samples and in some of them accounted for up to 90% of the viral load. Subsequent research found many new viral genomes related with the prototypical crAssphage, leading to the creation of the *Crassvirales* order in 2021. *Crassvirales* has been found widely present in human populations (Guerin *et al.*, 2018) (Edwards *et al.*, 2019). Phages form an important part of the human gut biome where they control bacterial subpopulations and vary strongly between individuals. The *Crassvirales* are double stranded DNA bacteriophages ranging from approximately 90 to 190 Kb, with their circular genome organised into two parts of roughly equal size with strictly opposite gene orientation and inverted GC skew (Shkoporov *et al.*, 2018). During its life cycle, the bacteriophage shows an unusual relationship with its host and delays lysis in favour of infecting progeny (Shkoporov *et al.*, 2021). The hosts of *Crassvirales* bacteriophages are predicted to be members of the Bacteroidetes phylum, which are difficult to culture in laboratory settings (Yutin *et al.*, 2021). However efforts in this direction led to the co-culture of ΦcrAss001 and *Bacteroides intestinalis* as the host (Shkoporov *et al.*, 2021), as well as characterizing additional *Crassvirales* bacteriophages (Guerin *et al.*, 2021). Comparison of the crAssphage proteins sequences to families of related proteins identified by using blastp did not establish any specific relationships with other known phages. From the constructed proteomic tree it is shown that crAssphage is only distantly related to other viruses (Dutilh *et al.*, 2014). Several proteins associated with a relatively conserved structural gene module in *Crassvirales* phages have been identified, namely the major capsid protein (MCP), portal protein and large terminase subunit (TerL) (Yutin *et al.*, 2018). Subsequent research allowed for the identification of nearly 600 diverse genomes of *Crassvirales* phages, with predicted sequence similarity to the *Crassvirales* large terminase subunit (Yutin *et al.*, 2021). These genomes show a significant amount of sequence diversity but a relatively conserved gene order.

Three readily discernable blocks of genes were predicted by Prodigal encoding respectively, virion components and proteins involved in virion assembly, components involved with replication machinery and components of transcription machinery (Yutin *et al.*, 2021). The structural gene module was found to be the most conserved over all *Crassvirales* phage species and encodes the MCP, TerL, portal protein, Integration Host Factor (IHF), tail stabilization protein, tail tubular protein as well as uncharacterized genes. This module could be seen as the genomic signature of *Crassvirales* phages (Yutin *et al.*, 2021).

The inspiration for the machine learning approach is gained primarily from the Imodulon publication (Rychel *et al.*, 2020). Their approach using Principal Component Analysis and Independent Component Analysis on transcriptomic data from lab cultivated samples allowed for the classification and analysis of gene co-expression in a diverse set of bacteria. Here we adopt their approach for use on metatranscriptomic data and viral genomes of the *Crassvirales* order. The metatranscriptomic data consists of human gut and wastewater samples, where we reuse data from other publications that were published on the MGnify database (Mitchell and Almeida, 2019). We hypothesize that genes that co-express tend to participate in the same biological process and might have a related function, allowing for the further specification of gene function in viral genomes.

3 Methods

3.1 Data collection from MGnify

To obtain gene expression data potentially containing *Crassvirales* bacteriophages, the MGnify database was screened for metatranscriptomic samples from the human gut or wastewater biomes (Mitchell and Almeida, 2019). For this, the MGnify API was used to download the sequencing reads in fastq format, which were further inspected for quality, adapters and rRNA content (Kopylova *et al.*, 2012). The terms "metatranscriptomic", "root:Host-associated:Human:Digestive system" and "root:Engineered:Wastewater" were used to query the MGnify database, yielding 325 samples. Sequencing reads of these samples were inspected with Fastqc and adapters were removed when necessary using Fastp with default paramters (Chen *et al.*, 2018). SortmeRNA was applied to the data to inspect the rRNA content of the data files (Kopylova *et al.*, 2012). Fastqc and Multiqc allowed for the manual inspection of some of the data, recognizing and removing some studies marked as metatranscriptomic but actually containing 16S amplified rRNA data.

3.2 Genome and gene annotations

A set of 276 *Crassvirales* complete genomes representing type species were used as reference data set in this study. The genomes were annotated using Prodigal v2.6.3 (Hyatt *et al.*, 2010) under three different translation tables, and selecting the most suitable based on coding density, number of Open Reading Frames (ORFs) predicted and mean length of the proteins. To functionally annotate the ORFs, these were compared to a set of conserved proteins in the *Crassvirales* order (Yutin *et al.*, 2021) (Yutin *et al.*, 2018).

3.3 Alignment against *Crassvirales* phages

To calculate the gene expression values, all metatranscriptomic samples were aligned against the reference set of *Crassvirales* phage genomes using the STAR aligner (Dobin *et al.*, 2013). Default settings were used except the 'alignIntronMax' setting which was set to 1 to account for *Crassvirales* phages not containing any introns, but rather mostly self-splicing elements (Yutin *et al.*, 2018). The resultant SAM files were sorted, converted to BAM and filtered for aligned reads by samtools (Danecek *et al.*, 2021). For the horizontal coverage statistic pileup from BBtools was applied to the aligned read BAM files (Bushnell, 2021). To ensure sufficient quality of samples for each *Crassvirales* phage genome quality thresholds were applied, where the sample must cover at least 50% of the length of the *Crassvirales* phage genome and should recruit at least 1000 uniquely aligning reads.

3.4 Calculating reads per kilobase per million values for metatranscriptomic data

To define the pattern of expression of *Crassvirales* phages the metatranscriptomic alignments are converted to Reads per Kilobase per Million (RPKM) values. *Crassvirales* species have been shown to have a relatively conversed sequence identity in some genes, which leads to reads mapping to multiple species with the same quality, even when one of these species is not in the sample (Guerin *et al.*, 2018)(Shkoporov *et al.*, 2018). With a significant amount of the aligned reads being multimappers, a custom algorithm was written to account for multimapping reads in the RPKM calculation. The algorithm is based on previously published work of separating the weight of a read in metagenomic data but adapted for the specifics of metatranscriptomic data (Iverson *et al.*, 2012). By calculating an estimation of the abundance of each *Crassvirales* phage in the sample, multimapping reads can be weighted so that the total weight of 1 is distributed over the *Crassvirales* phage genomes it aligns to. For each

species, abundance is estimated as the fraction of uniquely aligning reads divided by the length they cover in the genome. The multimapping read gets distributed over the genomes it aligns to, weighted by the abundance fraction. However since the abundance fraction accounts for all genomes the weights for the genomes the multimapping read aligns to are scaled to be a total of 100%. For example if a multimapping read aligns to two genomes with a abundance fraction of 5% and 10% these are scaled to 33% and 66%. The abundance fraction is updated based on previously processed multimapping reads (Iverson *et al.*, 2012). The abundance estimation for genomes is influenced by their relative sequence identity to other *Crassvirales* species. To address this the algorithm calculates the length of the unique area for each genome and divides the fraction of that genome by that area. This balances the abundance fraction to avoid favouring unique *Crassvirales* species. The percentage of the data in a sample aligning to *Crassvirales* phages is variant, influencing the relative range of RPKM values calculated for the sample. To address this variance the RPKM value is divided by the total number of *Crassvirales* aligning reads instead of the total number of sequencing reads in the sample. The final matrix of RPKM values across data files for genes in *Crassvirales* phage genomes is the input for the machine learning algorithm Independent Component Analysis (ICA).

3.5 PCA and ICA to find patterns of co-expression

In order to find the number of features in the RPKM matrix, PCA was applied. The PCA variance explained property shows the amount of variance explained by each of the generated axis. The number of axis required in order to reach 99 percent of the variance explained was then used as the number of signals to look for using ICA. This approach is similar to the iModulon publication on which this work is based (Rychel *et al.*, 2020). By using for the number of patterns that are required to explain almost all the variance the dimensionality of the problem is reduced, effectively discarding signals that would be based mostly on noise (Hyvärinen and Oja, 2000). ICA is a stochastic search algorithm capable of discerning features (signals) from noisy source data. The Scikit-learn implementation of ICA is used in the pipeline (Pedregosa *et al.*, 2011). By estimating the strength of the signals and then maximizing their non-gaussianity several source signals can be extracted from a mixed source. The mixed source is the RPKM matrix and the source signal is the pattern of expression changing over the different samples in this matrix. Before applying the ICA algorithm it is useful to do some preprocessing. By whitening the data, the observed vector x representing a samples expression is transformed so that its components are uncorrelated and their variances equal to unity. ICA is applied to this whitened data and generates for each signal a component S with values for each of the genes in the genome. These values have signs that are interchangeable (Hyvärinen and Oja, 2000). Due to the random initialization of weights inherent to ICA the algorithm is run a thousand times, for each run gene modules are selected.

3.6 Selecting Genes from ICA's S components

From the generated source signal of ICA we want to select modules of genes. This requires a way of thresholding genes that are part of the module from the genes that are part of the noise. The values in the S component represent the strength of the gene in the pattern of expression found by ICA. For all the genes in a genome a value is generated. ICA relies on the Central Limit theorem, the distribution of a sum of independent random variables tend towards a Gaussian distribution (Hyvärinen and Oja, 2000). Due to this the individual components must be non-gaussian. To consistently select genes belonging to the gene module in the component this property can be used. First sorting all the absolute values in the components. Then taking the largest outlier gene and testing if the remainder is a normal gaussian distribution using D'augistinos K-squared test. This is repeated until the remainder is a normal Gaussian distribution. Here it is assumed that due to the remaining gene values being Gaussian ICA is no longer able to discern a source signal from the distribution and as such they represent noise. All the genes selected during this process constitute the module of co-expressed genes represented by this component. In figure 9 a visualization of a single component S from ICA is provided.

3.7 Distance permutation statistics

The quality and biological significance of the gene modules is difficult to interpret especially as the size of the gene modules increases. To aid the evaluation a value, indicating the relative clustering of genes in a gene module, is generated by applying distance permutation to the set of genes in a genome. The mean of the distance between a random selection of genes of the same size as a gene module found by ICA is calculated a thousand times. The distance of the found gene module is compared to these randomly generated gene module distances. The number of times the random gene modules distance is smaller than the real gene modules distance represents the p-value indicating the relative clustering of the real gene module. The underlying assumption is that co-expressed genes tend to be together on a genome, however it is a strength of ICA to find genes that are co-expressed not located together.

4 Results

4.1 Aligning Metatranscriptomic samples

In order to calculate gene expression values first the metatranscriptomic data must be aligned against the *Crassvirales* genomes. From the 325 sequencing samples originally gathered from MGnify, 260 are used as the sample dataset in the analysis. A set of 276 *Crassvirales* complete genomes representing type species were used as reference dataset. Using the STAR aligner the samples were aligned against the references (Dobin *et al.*, 2013). In the 260 samples the number of reads mapping with the same quality to multiple positions was found to be a distribution between 30 and 80 percent of the total number of reads aligning to *Crassvirales* genomes. After applying the quality requirements the super majority of the sample and *Crassvirales* genome combinations were eliminated mostly due to the horizontal coverage being insufficient. The number of samples per genome is displayed in Table 1.

Table 1. Samples per genome

OLQR01000043	23
crAss001_MH675552	18
OLOB01000056	11
OJOV01000080	9
OCHF01000011	8
OMDD01000022	7

4.2 RPKM matrix

To define the expression of genes on *Crassvirales* phages the alignments for the sample genome combinations of Table 1 need to be normalized and converted to a matrix of RPKM values. Due to the large number of reads mapping to multiple locations in different *Crassvirales* genomes a custom algorithm is applied that distributes the weight of the multimapping reads across genomes. The resultant RPKM matrix contains expression values for genes encoded on the genomes mentioned in Table 1.

4.3 PCA and ICA

From the RPKM matrix defining what genes are co-expressed can be done with a variety of methods. The approach used is based on the IModulon publication (Rychel *et al.*, 2020). The RPKM matrix is preprocessed by whitening and centering the values. For each genome the number of gene modules, containing sets of genes with a similar expression pattern, can be found by using the number of axis required to explain 99% of the variance. Effectively applying dimensionality reduction, gene modules that would mostly be based on noise are discarded.

Analyzing the expression values to know what genes are strongly co-expressed is done by applying ICA. ICA is a stochastic search algorithm capable of discerning features (signals) from noisy source data. For each gene a value is generated that represents their dominance in the signal. In figure 9 these values are visualized with outliers being most dominant. By selecting genes from these values based on a threshold the final set of genes representing the gene modules is selected.

4.4 Gene Modules

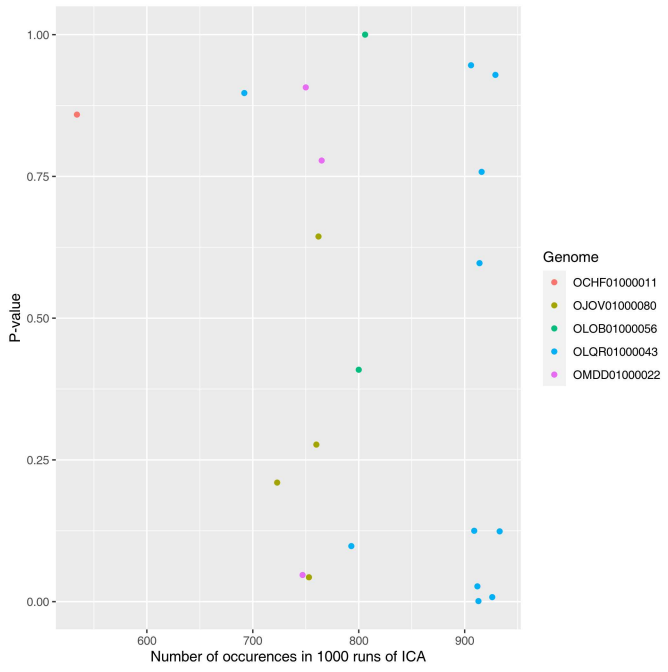

Fig. 1. Gene modules found from the top 5 genomes present in the samples are shown, with their number of occurrences in 1000 runs of ICA plotted along the x-axis and the relative density of genes inside the module compared to the density of genes in randomly generated gene modules expressed as a p-value along the y-axis.

Due to ICA randomly initializing its weights before convergence the results are slightly different every time the algorithm is ran. The gene modules that are relatively stable, generated at least 500 times when running ICA, are plotted on (Fig. 1) with the x-axis representing their stability. We have previously hypothesized that genes that co-express tend to participate in the same biological process, to further evaluate this hypothesis a p-value is calculated, indicating the relative distance between genes in the gene modules, with smaller values representing a high density of genes. The gene modules are color coded in the legend by this density value. By comparing the stability of a gene module to the relative density of the gene module a indication of reliability can be interpreted (Fig. 1).

The OLQR01000043 *Crassvirales* phage genome has the most gene modules with low p-values and high stability, most likely due to having the most samples available (Fig. 1. Whilst the other *Crassvirales* phage genomes do have gene modules with a higher relative stability in runs of ICA, a trend can be seen where genomes with lower numbers of samples are less stable.

To allow for visual interpretation of the found gene modules the gene modules are plotted on their respective genomes, with the genes placed on their positions in the genome alongside their function annotation if available (Fig. 3, Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8) OLQR01000043's gene modules show several interesting relationships (Fig. 3). There are some genes that appear dominant across gene modules, where several putative structural genes (gene75, gene77 and gene73) show up in most gene modules. The combination of TerL and portal proteins is found several times with varying gene density but high stability. This gene module has been described in literature as a highly conserved gene module used to identify contigs of *Crassvirales* phages (Yutin *et al.*, 2018) (Yutin *et al.*, 2021). Most of the gene modules have genes associated with phage packaging indicating those functions are relatively strongly represented in the expression data. The gene encoding the single stranded binding protein shows up several times in combination with the TerL and portal protein. In the 3rd gene module the single stranded binding protein is associated with a highly clustered set of genes without annotation. These genes might be associated with replication of the *Crassvirales* phages.

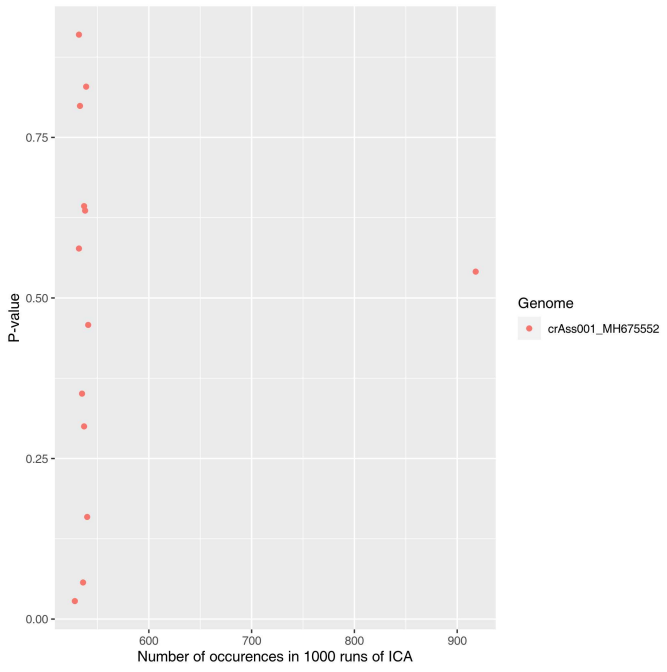

Fig. 2. Gene modules found from the cultivated ΦcrAss001 timeseries samples, with their number of occurrences in 1000 runs of ICA plotted along the x-axis and the relative density of genes from the gene module compared to the density of genes in randomly generated gene modules expressed as a p-value along the y-axis.

4.5 *Crassvirales* phage ΦcrAss001

The *Crassvirales* phage genome of crAss001 is a special case, where the pipeline has been rerun specifically on that genome due to the samples consisting of lab cultured infections of ΦcrAss001 in the host *Bacteroides Intestinalis*. The samples consist of a time series of metatranscriptomic

samples of ΦcrAss001 infections and is as such very different than the samples on which the other genomes gene modules are based. The stability of the gene modules is significantly lower (Fig. 2), which is expected due to the expression changing strongly across the time series as shown in the publication (Shkoporov *et al.*, 2021). Interestingly, the genes associated with *Crassvirales* phage packaging are still prevalent in the gene modules, with the Major Capsid Protein (MCP), gene77, portal protein and primase present in several gene modules (Fig. 4). The most stable gene module consisting of 5 genes does not share any genes with other gene modules. The last gene in this module is UDG, Uracil-DNA glycosylase, with no other functional annotations for the other genes.

5 Discussion

Whilst the approach generates interesting gene modules it does appear to be vulnerable to low sample counts. Rerunning the pipeline on a new dataset with many more samples would help to improve the stability of the gene modules. The number of gene modules generated appears to be related to the number of samples used for the genome, having more data should lead to more gene modules being found. Alternatively investigating alternative thresholds for data quality could allow for more samples being usable in the analysis of a *Crassvirales* genome. However, even with low sample counts and very noisy data, it was still possible to generate some gene modules with good stability and clustering of genes showing promise in the approach.

6 Conclusion

Inspired by the IModulon publication we have developed a method that recycles metatranscriptomic data to relate genes based on their expression pattern in the *Crassvirales* order. Principal Component Analysis and Independent Component Analysis are the algorithms at the core of the approach, transforming gene expression into gene modules consisting of genes that tend to co-express. We have hypothesized that these co-expressing genes tend to participate in the same biological process, allowing for further specification of gene function. We demonstrate that gene modules can be found, grouping genes spread over the *Crassvirales* genomes into putative functional clusters, with one excellent result for the *Crassvirales* genome OLQR01000043. Noticeably, this method can be applied to any microorganism mainly found in metatranscriptomic data and benefits strongly from a high number of samples.

7 Code Availability

A snakemake pipeline was written for this paper. The pipeline is published on github at https://github.com/Bryan-vd-Brand/MetaTranscriptomic-Gene-Grouping. The genome files, gene files and sample accessions used are also available alongside the pipeline.

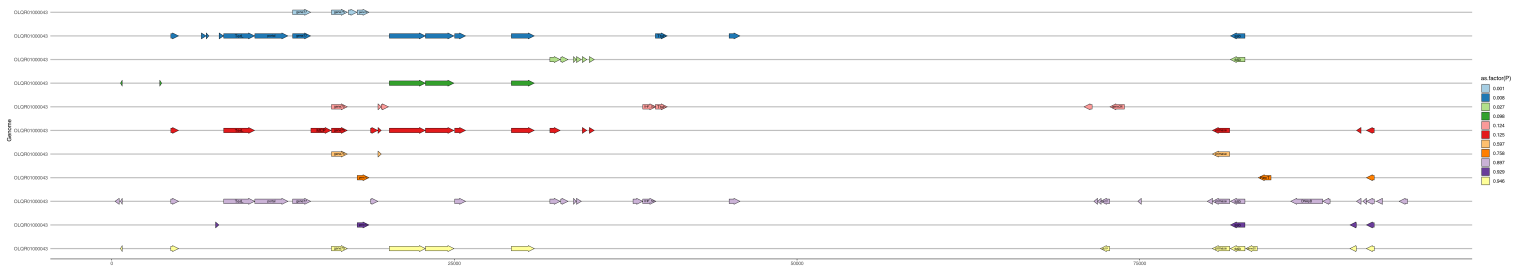

Fig. 3. Genes from gene modules found from samples with sufficient quality for the Crassvirales genome OLQR01000043 are plotted in their respective positions on the genome and annotated if available. The genes are color coded according to their relative gene density, expressed as a p-value in the legend.

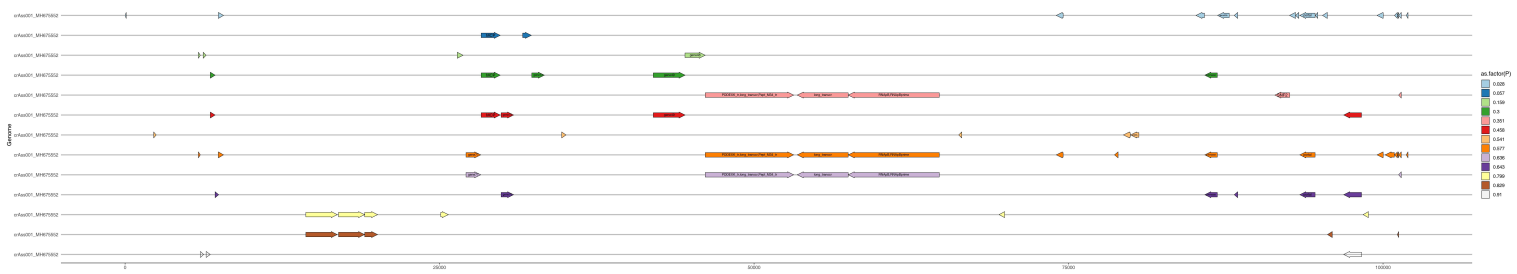

Fig. 4. Genes from gene modules found from cultivated ΦcrAss001 timeseries samples for the Crassvirales genome ΦcrAss001 are plotted in their respective positions on the genome and annotated if available. The genes are color coded according to their relative gene density, expressed as a p-value in the legend.

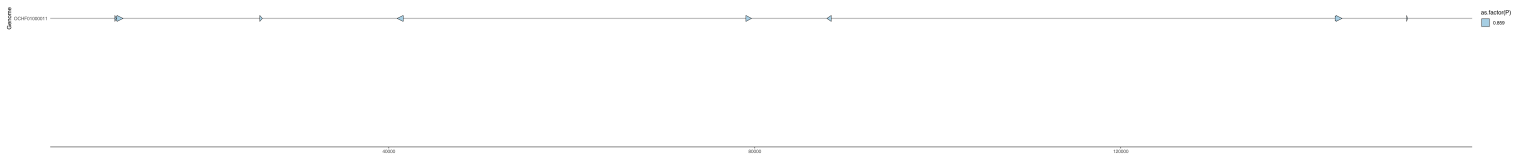

Fig. 5. Genes from gene modules found from samples with sufficient quality for the Crassvirales genome OCHF01000011 are plotted in their respective positions on the genome and annotated if available. The genes are color coded according to their relative gene density, expressed as a p-value in the legend.

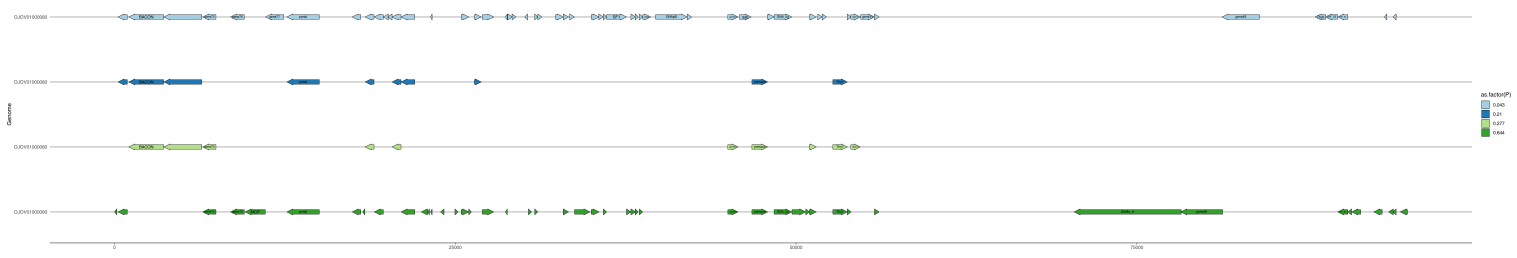

Fig. 6. Genes from gene modules found from samples with sufficient quality for the Crassvirales genome OJOV01000080 are plotted in their respective positions on the genome and annotated if available. The genes are color coded according to their relative gene density, expressed as a p-value in the legend.

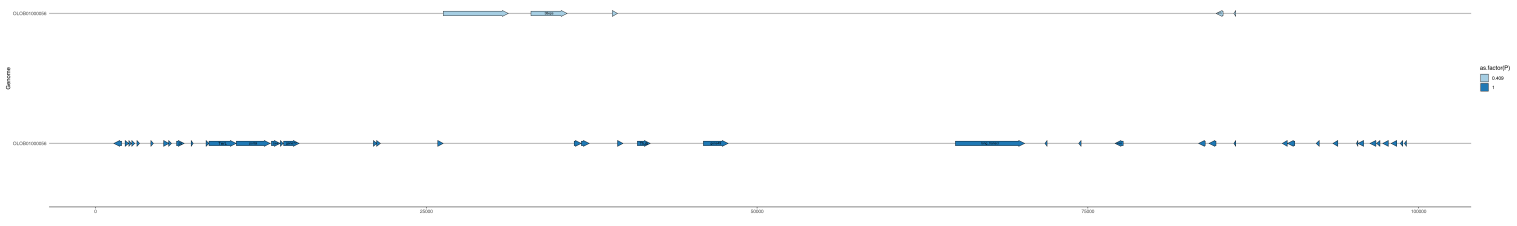

Fig. 7. Genes from gene modules found from samples with sufficient quality for the Crassvirales genome OLOB01000056 are plotted in their respective positions on the genome and annotated if available. The genes are color coded according to their relative gene density, expressed as a p-value in the legend.

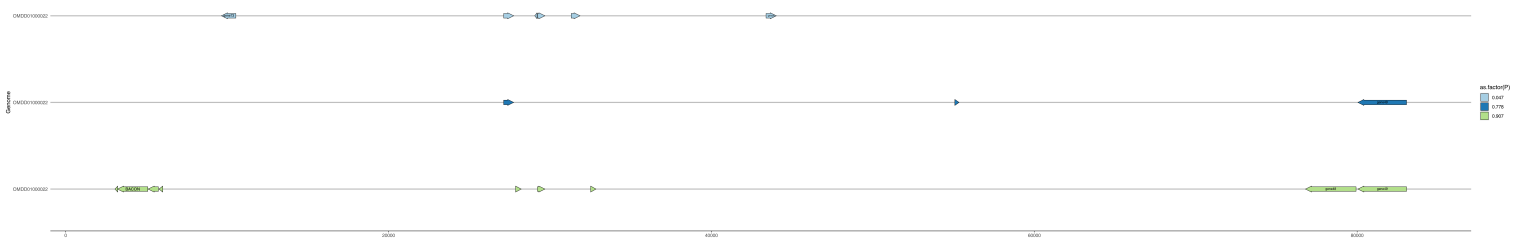

Fig. 8. Genes from gene modules found from samples with sufficient quality for the Crassvirales genome OMDD01000022 are plotted in their respective positions on the genome and annotated if available. The genes are color coded according to their relative gene density, expressed as a p-value in the legend.

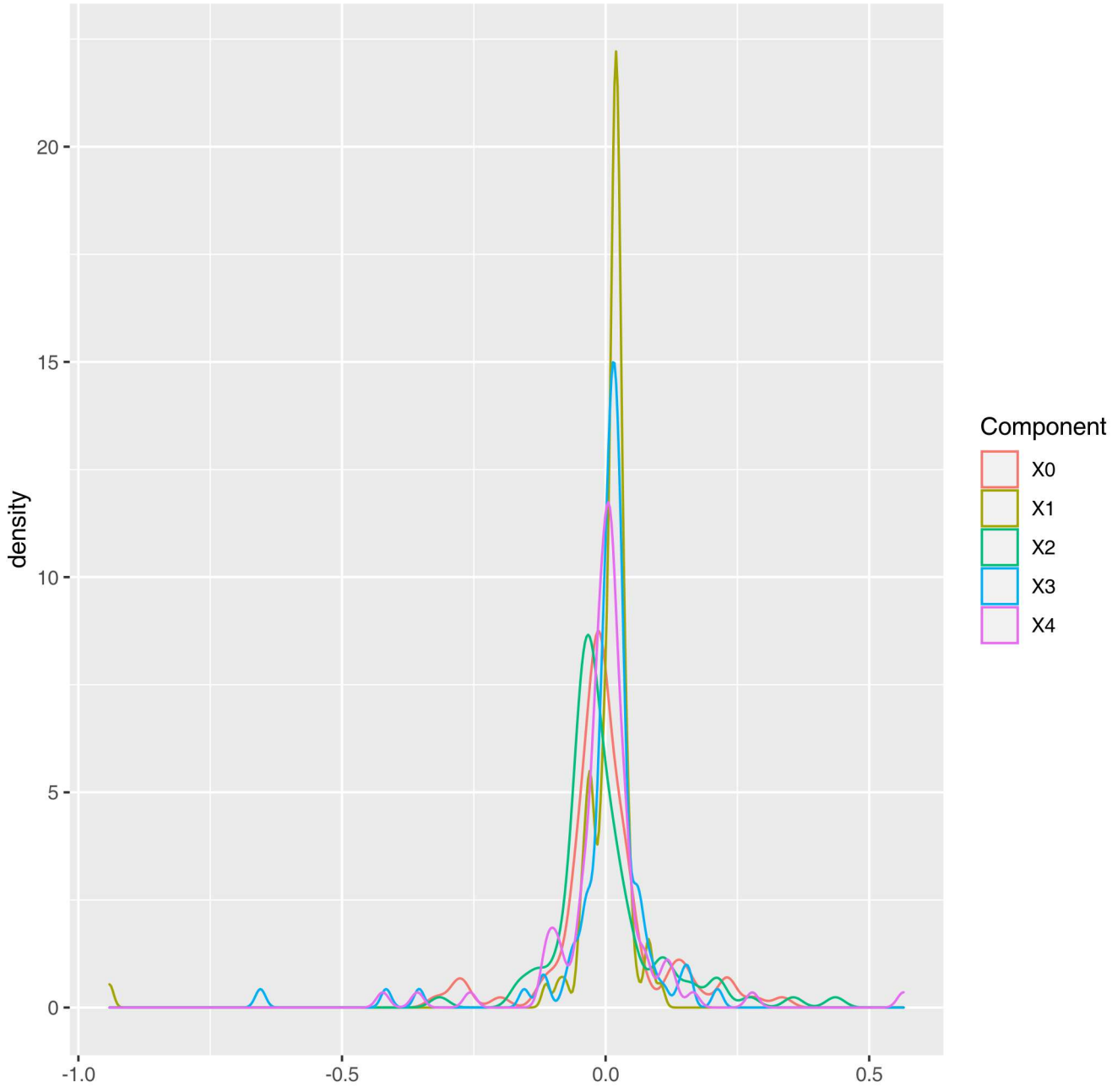

Fig. 9. The values indicating the relative dominance of a gene in a single gene module generated by ICA for the Crassvirales phage ΦcrAss001 are visualized as a density plot. The outliers on the x-axis of the plot represent genes that tend to be selected to be part of the gene module.

References

Bushnell, B. (2021). BBMap and BBTools. *sourceforge.net/projects/bbmap/*.

Chen, S. *et al.* (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Oxford Bioinformatics*, **34**(1), i884–i890.

Danecek, P. *et al.* (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, **10**.

Dobin, A. *et al.* (2013). STAR: ultrafast universal RNA-seq aligner. *Oxford Bioinformatics*, **29**, 15–21.

Dutilh, B. *et al.* (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun*, **5**(24), 4498.

Edwards, R. *et al.* (2019). Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiology*, **4**, 1727–1736.

Guerin, E. *et al.* (2018). Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host and Microbe*, **24**, 653–664.

Guerin, E. *et al.* (2021). Isolation and characterisation of ΦcrAss002, a crAss-like phage from the human gut that infects Bacteroides xylanisolvens. *Microbiome*, **9**, 1–89.

Hyatt, D. *et al.* (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**(1), 119.

Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Networks*, **13**(4), 411–430.

Iverson, V. *et al.* (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science*, **335**(6068), 587–590.

Kopylova, E. *et al.* (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Oxford Bioinformatics*, **28**, 3211–3217.

Mitchell, A. and Almeida, A. e. a. (2019). Mgnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*, **48**, D570–D578.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Rychel, K. *et al.* (2020). iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Research*, **49**(D1), D112–D120.

Shkoporov, A. *et al.* (2018). Φcrass001 represents the most abundant bacteriophage family in the human gut and infects bacteroides intestinalis. *Nature Communications*, **9**, 4781.

Shkoporov, A. *et al.* (2021). Long-term persistence of crass-like phage crass001 is associated with phase variation in bacteroides intestinalis. *BMC Biol*, **19**, 163.

Yutin, N. *et al.* (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiology*, **3**, 38–46.

Yutin, N. *et al.* (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nature Commun*, **12**(16), 1044.