dsDNA Phage Andromeda with Quasispecies-like Genomic Hypervariability Establishes a Novel Phylogenetic Branch Within the *Autographivirinae* Subfamily

Damian J. Magill[1], Timofey A. Skvortsov[2], Leonid A. Kulakov[1*]

[1]Queen's University Belfast, School of Biological Sciences, Medical Biology Centre, 97 Lisburn Road, Belfast BT9 7BL, Northern Ireland, UK

[2]Queen's University Belfast, School of Pharmacy, Medical Biology Centre, 97 Lisburn Road, Belfast BT9 7BL, Northern Ireland, UK

*Correspondence: Leonid A. Kulakov, Queen's University Belfast, School of Biological Sciences, Medical Biology Centre, 97 Lisburn Road, Belfast, BT9 7BL, Northern Ireland, UK

Email: leonid.kulakov@gmail.com

Keywords: bacteriophage; genome variability; Autographivirinae.

Abstract

A new lytic phage Andromeda, specific to the economically important plant pathogen *Pseudomonas syringae* was isolated. It belongs to the *Podoviridae* family, *Autographivirinae* subfamily and possesses a linear dsDNA genome of 40,008 bp with four localized nicks. We reveal that Andromeda represents an atypical member of the *phiKMVlviruses* and establish a novel group of Andromeda-related phages within the *Autographivirinae* designated as the “ExophiKMVviruses”. The “ExophiKMVviruses” were revealed to share conservation of gene order and products with core *phiKMVviruses* despite a sequence-based relationship to T7 and SP6-related phages. Crucially Andromeda’s genome has no less than 80 hypervariable sites (SNPs) which show genome wide distribution resulting in heterogenous populations of this phage reminiscent of those of RNA virus quasispecies.

Introduction

The *phiKMVviruses* constitute a ubiquitously distributed group of bacteriophages closely related to *T7viruses* (Grymonprez et al. 2003). The major difference between these two groups of viruses is in the localisation of their single subunit RNA polymerase enzymes. Both groups fall under the *Autographivirinae* subfamily along with the *SP6viruses*. With respect to *phiKMVviruses* in particular, the majority of phages infect *Pseudomonas, Klebsiella,* and *Proteus* strains, and are clustered into two subgroups. The first subgroup includes a core set of phages displaying a high level of identity to *Pseudomonas* phage phiKMV (Grymonprez et al. 2003) and the second are the *KP34viruses* (Drulis-Kawa et al. 2011). Phages of the core subgroup infecting *P. aeruginosa* are especially abundant with first representatives were described in 1980s (Kulakov et al., 1985).
The analysis of the eponym phage phiKMV revealed a linear double stranded DNA genome of 42,105 bp containing 414 bp direct terminal repeats. *phiKMVviruses* are of interest for various reasons. Those infecting *Pseudomonas aeruginosa* exhibit a number of interesting biological phenomena including the presence of four localised single-strand interruptions (nicks) associated with the consensus sequence 5'-CGACNNNNNCCTACTCCGG-3' (Kulakov et al., 1985, 2009). All nicks found are located in the same DNA strand and could be "repaired" in vitro by DNA ligase treatment. It is interesting that DNA of phiKMV lacks one nick, located in the early genes region of other representatives of the group (Kulakov et al., 2009). The more recent study revealed that at least some phiKMV phages exhibit hypervariability associated with specific genomic regions (Magill et al. 2017). These hypervariable loci were found in genes associated with phage adsorption/host specificity and may therefore influence the evolutionary pathway of the phages (Magill et al. 2017).

This study describes the biological properties of a novel phage, Andromeda, which infects *Pseudomonas syringae*. The presence of a large number of hypervariable sites in the Andromeda's genome is investigated.

Materials and Methods

Phage Isolation and Purification

Bacteriophage was isolated following enrichment of 100 ml of water in cultures of *Pseudomonas syringae* pv *syringae* Van Hall followed by plating using the double agar overlay plaque assay.

To obtain high titre preparations of the phage cultures of *P. syringae* (OD600 = 0.3) were inoculated at a MOI = 0.01. Chloroform was added immediately after culture lysis (~4 hours) to maximise yield after insensitivity was determined. The lysate was concentrated by PEG precipitation, and purified using CsCl density gradient centrifugation as described by Sambrook and Russell (2001). Bands were extracted and dialysis ensued with modified SM buffer (100 mM NaCl, 50 mM Tris, 10 mM, 8 mM MgSO4, 10 mM CaCl2) using Amicon Ultra-15 100kDa filters. DNA extraction, phenol-chloroform extraction, and ethanol precipitation, was carried out using SDS and proteinase K as described previously (Sambrook and Russell, 2001). DNA was purified twice prior to sequencing using the MoBio DNA Purification Kit according to the manufacturer's instructions.

Electron Microscopy

Carbon-coated formovar grids were subjected to hydrophilic treatment with poly-l-lysine for 5 min. Upon drying, CsCl purified and dialysed phage suspensions were deposited on the grids and successively stained with 2% uranyl acetate (pH 4.5). Samples were observed using a Phillips CM100 transmission electron microscope at 100 KeV.

Determination of Adsorption Coefficient

Determination of adsorption efficiency of phages was carried out using the method described in Clokie et al. (2008).

All adsorption assays were conducted in triplicate with final adsorption constant calculated from the mean of the replicate constants.

One-step Growth Curves

Determination of bacteriophage one-step growth curves was conducted according to the established protocol. Bacterial cultures were grown until OD600 = 0.3 (10^8 cfu/ml) and inoculated with phage to give MOI $\geq$ 1. Adsorption was allowed for 5 min at 37^{o}C followed incubation with aeration at the same temperature. Duplicate 100 µl samples were taken every 5 minutes for 90 minutes. One of each samples were treated with 150 µl of chloroform and both were plated on the lawns of *P. syringae* using the double agar overlay plaque assay. Burst size was calculated by dividing the average of the values following the growth phase of the curve (the step) by the average of the baseline values before the first increase was observed in titre. The average of triplicates was then taken as the burst size. Latent periods were taken as the point in the graph of the untreated phage extraction whereby a rise in titre was observed, whereas the eclipse period was determined from the chloroform treated curve.

Visualisation of localised single-strand interruptions

Detection of localised nicks were conducted as described earlier (Kulakov et al. 2009). Phage DNA (3 µg) underwent overnight ligation at 4^{o}C in a final volume of 20 µl. NaOH was added to reactions to a final concentration of 50 mM. This was also done for 1.8 µg of unligated DNA. Both samples were incubated for 5 min on ice and then immediately analysis by gel electrophoresis (0.9% agarose).

Library Preparation and Sequencing

Sequencing libraries were prepared from 50 ng of phage genomic DNA using the Nextera DNA Sample Preparation Kit (Illumina, USA) at the University of Cambridge Sequencing Facility. A 1% PhiX v3 library spike-in was used as a quality control for cluster generation and sequencing. Sequencing of the resulting library was carried out from both ends (2x300 bp) with the 600-cycle MiSeq Reagent Kit v3 on MiSeq (Illumina, USA) and the adapters trimmed from the resulting reads at the facility.

Sequence Assembly and Bioinformatics

Sequencing reads underwent initial quality checking using FastQC (Andrews, 2010) followed by stringent trimming parameters. Reads with a Q-score < 20, containing N's, and with a final length < 40bp were discarded using Trimmomatic v0.32 (Bolger and Giorgi, 2014).

Assembly was carried out using Geneious R8 (Biomatters, New Zealand) and checked for potential misassembly by comparison with output from SPAdes v3.7.0 (Bankevich et al. 2012). Genome length contigs were input to the scaffolding tool SSPACE to extend these in order to ensure the resolution of genome ends (Boetzer et al. 2010). In addition, the sequence obtained was utilised as a reference for read mapping using bbmap v35.x (Bushnell, 2016) and subsequent manual inspection in IGV to check for potentially ambiguous regions (Thorvaldsdottir et al. 2013). ORFs were classified using the intrinsic Artemis ORF classification and GeneMark HmmS followed by manual verification of ribosomal binding sites (Lukashin and Borodovsky, 1998; Rutherford et al. 2000). Annotation was carried out using BLASTp, psiBLAST, and Hmmer algorithms, as well as enhanced domain classification using the Delta-BLAST algorithm (Finn et al. 2011). tRNAs were classified using tRNA-scanSE and Aragorn (Laslett and Canback, 2004; Lowe and Eddy, 1997), and rho-independent terminators were identified using Arnold with an energy cutoff of -10 employed, following by manual inspection of the output from subsequent Mfold analysis and adjustment as necessary (Zuker, 2003). To identify putative regulatory sequences, 100 bp upstream of every ORF was extracted using an in-house script and analysed using both neural network promoter prediction for prokaryotic sigma 70 promoters and Multiple EM for Motif Elicitation to identify phage specific promoters (Bailey et al. 2006).

Phylogenetic analysis

Initial localisation of putative *Autographivirinae* candidates was based on similarity in 3 key proteins: the major capsid protein, the terminase large subunit, and the presence of a single-subunit RNA polymerase. Upon obtaining sequences based on identity searches, genome size and organisation were utilised as points of further discrimination before their proposed inclusion as *Autographivirinae* members. Differentiation of *phiKMVviruses* from the T7/SP6 groups took place based on the downstream localisation of the RNA polymerase. For better characterised phages, descriptive data was also considered with respect to inclusion. Reannotation of phages was carried out using Prokka (Seemann, 2014).

Two conserved phage proteins were utilised in computing the phylogenetic relationships between putative *Autographivirinae* members: the single subunit RNA polymerase and the terminase large subunit. Alignments of the amino acid sequences were carried out in Clustal Omega (Sievers et al. 2011). The inference of phylogenies was then carried out using IQTree (Nguyen et al. 2014), which was allowed to compute the optimal protein model to be used. Visualisation of the resulting tree file was subsequently performed with the Interactive tree of life (IToL) (Letunic and Bork, 2016).
In order to permit analysis of gene localisation within the *Autographivirinae*, the GGgenes package was utilised within the R programming environment to generate a series of genome maps.

Molecular Modelling

Molecular models were constructed from corresponding amino acid sequences using the I-Tasser package and energy minimisation of the first, most favourable prediction carried out on the Yasara server (Krieger et al. 2009; Zhang, 2008). Fidelity of models was assessed using the validation tools of the Whatif server and checked for stereochemical clashes within Chimera (Petterson et al. 2004). Ramachandran analysis was subsequently carried out as an additional point of quality assessment. Superimposition and quantitative similarity of protein structures was provided by the TM-Align algorithm (Zhang and Skolnick, 2005).

Results and Discussion

Isolation and Biological Characteristics

Bacteriophage Andromeda was isolated from the River Lagan as described in Material and Methods. Upon infection of *Pseudomonas syringae,* phage Andromeda yields large, clear plaques of ~3mm in size. Electron microscopic analysis of Andromeda revealed a virion morphology meriting inclusion within the *Podoviridae* bfamily (Fig. 1). Specifically, Andromeda was found to possess a capsid of icosahedral symmetry with a length of 61 nm (+- 1.5 nm) and width of 62 nm (+- 1 nm). Andromeda was observed to possess a detectable tail structure.
Analysis of the growth characteristics revealed an adsorption coefficient of 1.48×10^{-9} $ml.min^{-1}$ (Fig. 2a), eclipse period of 20 – 25 min, and latent period of 25 – 30 min (Fig. 2b). Upon completion of its infection cycle, Andromeda yields on average 70 particles per cell.

Genome Structure and Gene Regulation

Phage DNA was sequenced as described in Material and Methods to a per base average coverage of 2149. The assembly and subsequent analysis were conducted and these revealed a linear dsDNA genome of 40,008 bp.

The GC content of the genome is 58.23% and using cumulative GC skew analysis, the putative origin of replication was found to be at 3001 bp. 46 ORFs were predicted within the Andromeda genome resulting in a total coding potential of 94.3%. Four ORFs utilise TTG, and 1 utilises GTG as start codons with the rest beginning with ATG. All ORFs are encoded off the same strand with the exception of gp20. Analysis of the Andromeda genome by both the PhageTerm and Li's methods revealed that this phage is terminally redundant and likely utilises a headful mechanism of packaging similar to that of phage P1 (+ve strand analysis (p = 1.17e-04), -ve strand analysis (p = 2.81e-06)). Based on nucleotide homology, the closest relative to Andromeda is the *P. tolaasii* phage Bf7 (Sajben-Nagy et al. 2012) with which it shares 83% identity across 85% of its genome (Fig. 3). Of the 46 ORFs, 19 were functionally assigned and are arranged within a strict modular architecture. Seven ORFs encode products associated with DNA replication, recombination, and repair. In particular, the presence of a single subunit RNA polymerase was detected (T3/T7 type: E value = 0) (gp26) and is located downstream of the DNA polymerase. This polymerase shows homology to those from a plethora of other phages unclassified beyond the family level, whose phylogenetic affiliations are considered below. One can observe some distant homology with this polymerase when compared to that of phiKMV, particularly in the N and C termini, the specificity loop, and the palm and finger domains. Indeed, molecular modelling of both polymerases and their superimposition revealed remarkably similar structures (Fig. S1) and a 73.5% level of similarity according to the TM-Align algorithm. This, along with the characteristic position of the RNA polymerase, initially suggests inclusion of Andromeda within the phiKMV supergroup. Eight ORFs in Andromeda encode structural proteins including typical gene products observed in members of the *Autographivirinae* such as tail tubular proteins. One fundamental difference that was observed was that in *phiKMVviruses* the tail fibre complex appears to be represented by four gene products whereas in Andromeda we only find one. The remaining 4 functionally assigned ORFs encode the terminase small and large subunits as well as the lysis machinery, with the lysozyme belonging to the GH24 family of proteins.

With respect to gene regulation in Andromeda, 9 sigma70 promoters were predicted of which 3 precede the early gene region. This appears to represent a strong signal permitting the rapid transcription of early genes. Due to the presence of an RNA polymerase, we attempted to predict whether phage specific promoters were present. No statistically significant elements were discovered and of the obtained weak predictions, these were largely found on the non-coding strand. It is possible that Andromeda utilises sigma70 elements in the generation of large polycistronic mRNAs. In regard to transcriptional termination, only one rho-independent terminator was found, downstream of gp5. No tRNA genes were discovered in this phage, a feature typical of the *Autographivirinae* group of phages.

Localized single strand interruptions have been reported in *phiKMVviruses* infecting *P. aeruginosa* (Kulakov et al. 2009) and *P. putida* phage tf (Glukhov et al. 2012). A search of the Andromeda genome failed to find the consensus associated with nicks in *P. aeruginosa* phages or *P. putida* phage. Upon subjecting Andromeda genomic DNA to NaOH denaturation, faint but distinguishable bands were observable on agarose gels which disappear when repeating this reaction following ligation (Fig. 4). These results prove that localised nicks containing adjacent 5'-phosphate and 3'-hydroxyl groups are present in the Andromeda genome. The low intensity of the bands observed may suggest that only a certain proportion of phage DNA molecules is nicked. This was previously demonstrated for T5 phage of E. coli where a number of major (presence in over 50% of molecules) and minor (lower presence) canonical nicks are known (Bujard, 1969; Hayward and Smith, 1972; Jonston et al. 1977). Together, these results prove that this enigmatic phenomenon is a widely present feature in the *phiKMVvirus* group.

Phylogenetic Analysis of Andromeda and Establishment of the "ExophiKMVviruses"

Andromeda's genomic architecture and the similarity in a great many of its gene products such as the RNA/DNA polymerases, tail tubular proteins, major capsid protein, terminase subunits, and lysis machinery, implicates this phage as a member of the *phiKMVviruses*. However, one cannot ignore the fact that this phage stands apart from other members in the group. This highlights a taxonomic gap and in order to truly establish the phylogenetic standing of Andromeda, a comprehensive analysis and reclassification of the phiKMV supergroup was necessary.

Autographivirinae consist of the *T7viruses, SP6viruses,* and *phiKMVviruses,* of which *Enterobacteria* phage T7, *Salmonella* phage SP6, and *Pseudomonas* phage phiKMV form the type members respectively. From the time of writing, the International Committee for Taxonomy of Viruses (ICTV) currently recognises thirteen *phiKMVviruses* (including nine *KP34viruses*), five *SP6viruses*, and three *T7viruses*, which along with three other genera and unassigned members, result in a total of 40 phages under the *Autographivirinae* umbrella.

We conducted an in-depth analysis of the NCBI database in order to provide a comprehensive characterisation of the *Autographivirinae* through the classification of poorly or unanalysed viral sequences that may populate the database.

Iterative searches of the NCBI database resulted in the discovery of 176 putative *Autographivirinae*, including those 40 recognised by the ICTV. Of these 66, 45, and 65 represent species of the *phiKMVvirus*, *SP6virus,* and *T7virus* genera respectively. These phages specific to a diversity of hosts from *Cyanobacteria* to *Pseudomonas.* Analysis of the genome size distribution and GC content revealed that *T7viruses* have on average the shortest genomes (39,327 bp) and a central GC content (Fig. 5). We reannotated all 166 phage genomes, revealing that *T7viruses, phiKMVviruses,* and *SP6viruses* encode on average 46, 53, and 57 ORFs respectively. With respect to *T7viruses*, virtually all ORFs are encoded on one strand as is known for these phages however, two exceptions were found in *Ralstonia* phage phiITL-1 and *Pseudomonas* phage phi-S1 which possess 2 ORFs each encoded on the opposite strand (KP343639.1; Sillankorva et al. 2012). Another interesting point with respect to phage phi-S1 is that it is unique in encoding a predicted tmRNA element. This feature is not reported and it is not simply unique amongst the *T7viruses,* but there are no reports of such an element existing within phage genomes to date.

The *Autographivirinae* are somewhat know for lacking tRNA elements, which naturally are in limited abundance in smaller viral genomes. Annotation also, has found singular exceptions in all major genera. *Rhizobium* phage RHEph01 is erroneously described as a *Myovirus* on NCBI, however homology clearly indicates that it is a *T7virus* (Santamaria et al. 2014). Within the genome of RHEph01, 2 tRNA genes were found (tRNA Asn and TRNA Leu) at separate points in the genome. As for the *Sp6viruses* and *phiKMVviruses,* tRNA genes were found within Cyanophage NATL1A-7 (tRNA Gly) and *Xanthomonas* phage XAJ24 (tRNA Trp) respectively (NC_016658.1; KU197013.1).

Phylogenetic analysis of the putative *Autographivirinae* using the major capsid protein and

terminase large subunit as markers, those phages infecting *P. aeruginosa* cluster together in a tight core group (Fig. 6a and Fig. 6b). This is also true of the *KP34viruses* which are found in close association with *P. aeruginosa* infecting *phiKMVviruses* albeit with slight topological differences in the tree. In both cases, we find that the *T7viruses* form a large and expanded portion of the tree, but yet are found in close association with one another. Where we see a significant difference is with the *SP6viruses*. Here, we find that this group occupies two distinct clusters in both trees. The first is split by the *T7viruses* and the second by the final group of the tree, a putative cluster described here as the "ExophiKMVviruses". When we analyse the gene order of representatives of each genus (Fig. 7), we find that the "ExophiKMVviruses", despite clustering with *T7viruses* and *SP6viruses* based on sequence identity, show the downstream localisation of the single-subunit RNA polymerase as is typical for "true" *phiKMVviruses.* This is a feature conserved across the group (Supplementary Fig. 8). This poses a conundrum as to the true phylogenetic affiliation of this *phiKMVviruses*. Analysis of genome size distribution of the putative *Autographivirinae* (Fig. 5) show that "ExophiKMVviruses" possess a genome size range similar to that of other *phiKMVviruses*.

We have shown previously that *phiKMVviruses* lack the thioredoxin binding domain in their DNA polymerases (Magill et al. 2017) that confers heightened processivity and fidelity upon their T7 brethren (Magill et al. 2018; Tabor et al. 1987). Molecular modelling and superimposition of the Andromeda and T7 DNA polymerases shows the absence of this domain and highlights another difference between these groups (Fig. 9).

Taken together, it seems that the "ExophiKMVviruses" occupy an interface group between the *phiKMVviruses*/*Kp34viruses* and the *T7viruses*/*SP6viruses,* with sequence based classification alone being insufficient to adequately resolve the relationships of the *Autographivirinae* as a whole.

Genetic Heterogeneity in the Andromeda Genome

It has previously been discovered that several members of the *phiKMVviruses* display a high level of genome variability that is localised to specific regions (Magill et al. 2017). This is no less true of Andromeda, which was found to possess the highest number of such variants. In-depth analysis of this genomic feature was conducted in order to gain insights into the nature of the changes and the potential mechanisms at play.

Analysis of deep sequencing data of the Andromeda genome revealed that there are at

least 80 positions (SNP) where nucleotide substitutions are detected with frequencies ranging ranging from 1% to 12.48%. Detailed information with respect to each variant is provided in Table 1. Cumulative analysis of these frequencies shows that lower frequency variants predominate, with a high level of tapering observed above 3% (Fig 10). When one analyses the localisation of these variants, a genome wide distribution can be seen (Fig 10 and 11). This is unlike that reported previously in other *phiKMVviruses* where a distinct bias of SNPs distribution (e.g. in genes responsible for phage adsorption) was observed (Magill et al. 2017). On a cumulative basis however, it is clear that regions of high variability are observable which suggests either the consecutive introduction of errors in a single event such as polymerase slippage, or that these represent hotspots for mutation (Fig 10).

Further analysis of the nature of variants revealed that 73/80 (91.3%) lie within coding regions which is unsurprising given that the coding potential of the genome is 94.3% (Fig 12.). With respect to those variants lying within coding regions, 55 of these result in an amino acid change. Of the 18 silent mutations, 17 were found to exist within the third codon position. This position is overall the predominant location of variants with the second codon position having a similar abundance (Fig 12.). Interestingly, these 17 silent mutations constitute only half of the total variants observed at the third position. The predominance of missense mutations in general is extremely interesting, particularly in the third position. Analysis of the proportion of transversion and transition mutations revealed an unexpected abundance of the former, with 78.75% of all variants being of this type (including those in non-coding regions). Most transversions are observed in the second codon position, closely followed by the third. In general, whilst there are more potential pathways to transversion mutations, they are in reality less common relative to transitions. The exchange of a single ring base for a double ring and vice-versa is a significant change and is more likely to result in an alteration at the amino acid level and indeed, one that is less likely to be conservative. The predominance of transversions here explains the high level of non-silent mutations observed across all codon positions and is a unique occurrence not previously reported in other biological systems. The additional abundance of transversion mutations may also further implicate the DNA polymerase as a responsible factor in variant generation through some form of biochemical bias. This is additionally supported by the presence of an insertion mutation in the Andromeda genome within ORF14 (Table 1.). This variant lies within an 8 base poly-G tract that may have given rise to a slippage event during replication and results in a +1 frameshift and truncated gp14. The possible involvement of the phage DNA polymerases in generation of hypervariability

in genomes of phiKMVviruses was discussed previously (Magill et al. 2017). In respect of Andromeda it shares the lack of a crucial thioredoxin binding domain with the rest of the phiKMV phages, however the genome wide distribution of variable sites in its genome makes this phage a very special case.

Finally, our analysis of the Andromeda phage has once again highlighted the existence of a novel phenomenon within *Pseudomonas* phages that results in the production of highly heterogeneous populations. Whilst error-prone polymerase activity is almost certainly to be involved in the generation of genomic hypervariability, it unlikely to be the only constituent at play, with intriguing patterns of codon position distribution and localisation suggesting the involvement of some post-replication mechanisms.

References

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.

Bailey, T.L., Williams, N., Misleh, C. and Li, W.W., 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research,* **34**(suppl_2), pp. W369-W373.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S. and Prjibelski, A.D., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology,* **19**(5), pp. 455-477.

Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. and Pirovano, W., 2010. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics,* **27**(4), pp. 578-579.

Bolger, A. and Giorgi, F., 2014. Trimmomatic: a flexible read trimming tool for illumina NGS data. *URL http://www.usadellab.org/cms/index.php,* .

Bujard, H. 1969. Location of single-strand interruptions in the DNA of bacteriophage T5+. Proc Natl Acad Sci USA, 62: 1167 – 1174.

Bushnell, B., 2016. BBMap short read aligner. *University of California, Berkeley, California.URL http://sourceforge.net/projects/bbmap,* .

Clokie, M.R.J. & Kropinski, A.M. 2008, "Bacteriophages: methods and protocols, volume 1: isolation, characterization, and interaction", Humana Press2008, vol. 28

Drulis-Kawa, Z., Mackiewicz, P., Kęsik-Szeloch, A., Maciaszczyk-Dziubinska, E., Weber-Dąbrowska, B., Dorotkiewicz-Jach, A., Augustyniak, D., Majkowska-Skrobek, G., Bocer, T. and Empel, J., 2011. Isolation and characterisation of KP34—a novel φKMV-like bacteriophage for *Klebsiella pneumoniae*. *Applied Microbiology and Biotechnology,* **90**(4), pp. 1333-1345.

Finn, R.D., Clements, J. and Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. *Nucleic acids research,* **39**(suppl_2), pp. W29-W37.

Glukhov, A.S., Krutilina, A.I., Shlyapnikov, M.G., Severinov, K., Lavysh, D., Kochetkov, V.V., McGrath, J.W., de Leeuwe, C., Shaburova, O.V., Krylov, V.N., Akulenko, N.V., Kulakov L.A. 2012. Genomic analysis of *Pseudomonas putida* phage tf with localized single-strand DNA interruptions. PLoS One, 7(12):e51163. doi: 10.1371/journal.pone.0051163.

Hayward, G.S. and Smith M.G. 1972. The chromosome of bacteriophage T5. I. Analysis of single-stranded DNA fragments by agarose gel electrophoresis. J Mol Biol, 63: 383 – 396.

Jonston J. V., Nichols, B.P. and Donelson J.E.1977. Distribution of "Minor" nicks in bacteriophage T5 DNA. Journal of Virology. 22: 510 – 519.

Joseph, S. and David, W.R., 2001. Molecular cloning: a laboratory manual. *The Quarterly review of biology,* **76**(3), pp. 348-349.

Krieger, E., Joo, K., Lee, J., Lee, J., Raman, S., Thompson, J., Tyka, M., Baker, D. and Karplus, K., 2009. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: four approaches that performed well in CASP8. *Proteins: Structure, Function, and Bioinformatics,* **77**(S9), pp. 114-122.

Kulakov, L.A., Ksenzenko, V.N., Kochetkov V.V., Mazepa V.N. and Boronin A.M. 1985. Homology and adsorbtion specificity of P_seudomonas aeruginosa bacteriophages. Mol. Gen. Genetics, 200, pp. 123 – 127.

Kulakov, L.A., Ksenzenko, V.N., Shlyapnikov, M.G., Kochetkov, V.V., Del Casale, A., Allen, C.C., Larkin, M.J., Ceyssens, P. and Lavigne, R., 2009. Genomes of "phiKMV-like viruses" of Pseudomonas aeruginosa contain localized single-strand interruptions. *Virology,* **391**(1), pp. 1-4.

Laslett, D. and Canback, B., 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic acids research,* **32**(1), pp. 11-16.

Lavigne, R., Burkal'tseva, M.V., Robben, J., Sykilinda, N.N., Kurochkina, L.P., Grymonprez, B., Jonckx, B., Krylov, V.N., Mesyanzhinov, V.V. and Volckaert, G., 2003. The genome of bacteriophage φKMV, a T7-like virus infecting *Pseudomonas aeruginosa*. *Virology,* **312**(1), pp. 49-59.

Letunic, I. and Bork, P., 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research,* **44**(W1), pp. W242-W245.

Lowe, T.M. and Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research,* **25**(5), pp. 955-964.
Lukashin, A.V. and Borodovsky, M., 1998. GeneMark. hmm: new solutions for gene finding. *Nucleic acids research,* **26**(4), pp. 1107-1115.

Magill, D.J., Kucher, P.A., Krylov, V.N., Pleteneva, E.A., Quinn, J.P. and Kulakov, L.A., 2017. Localised genetic heterogeneity provides a novel mode of evolution in dsDNA phages. *Scientific reports,* **7**(1), pp. 13731.

Nguyen, L., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q., 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution,* **32**(1), pp. 268-274.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E., 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry,* **25**(13), pp. 1605-1612.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. and Barrell, B., 2000. Artemis: sequence visualization and annotation. *Bioinformatics,* **16**(10), pp. 944-945.

Sajben-Nagy, E., Maróti, G., Kredics, L., Horváth, B., Párducz, Á., Vágvölgyi, C. and Manczinger, L., 2012. Isolation of new Pseudomonas tolaasii bacteriophages and genomic investigation of the lytic phage BF7. *FEMS microbiology letters,* **332**(2), pp. 162-169.
Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* **30**(14), pp. 2068-2069.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D. and Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology,* **7**, pp. 539.

Sillankorva, S., Kropinski, A.M. and Azeredo, J., 2012. Genome sequence of the broad-host-range Pseudomonas phage Phi-S1. *Journal of virology,* **86**(18), pp. 10239-12.

Tabor, S., Huber, H.E. and Richardson, C.C., 1987. Escherichia coli thioredoxin confers processivity on the DNA polymerase activity of the gene 5 protein of bacteriophage T7. *The Journal of biological chemistry,* **262**(33), pp. 16212-16223.

Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P., 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics,* **14**(2), pp. 178-192.

Walker, J., Clokie, M. and Kropinski, A., 2009. Bacteriophages: Methods and protocols, volume 1: Isolation, characterization, and interactions.

Zhang, Y., 2008. I-TASSER server for protein 3D structure prediction. *BMC bioinformatics,* **9**(1), pp. 40.

Zhang, Y. and Skolnick, J., 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research,* **33**(7), pp. 2302-2309.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research,* **31**(13), pp. 3406-3415.

Legends to Figures

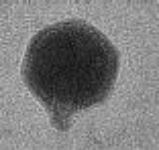
Fig 1. Electron microscopic image of bacteriophage Andromeda.

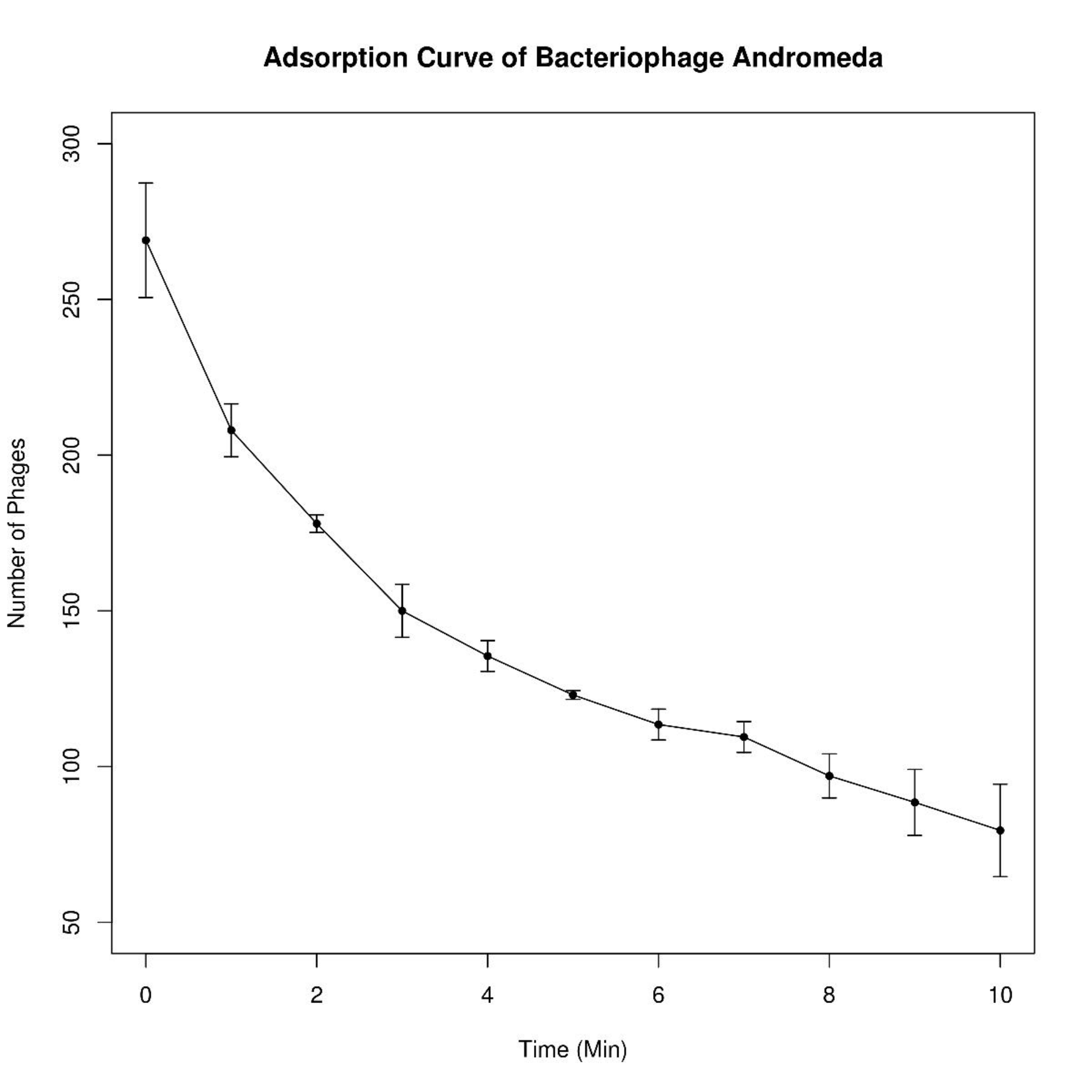
Fig 2. Growth characteristics of bacteriophage Andromeda. (a) Determination of Adsorption coefficient; (b) One-step growth curves: blue line – infected P. syringae cells treated with chloroform (intracellular phage); red line – untreated infected P. syringae cells (extracellular phage). All experiments were performed in triplicates.

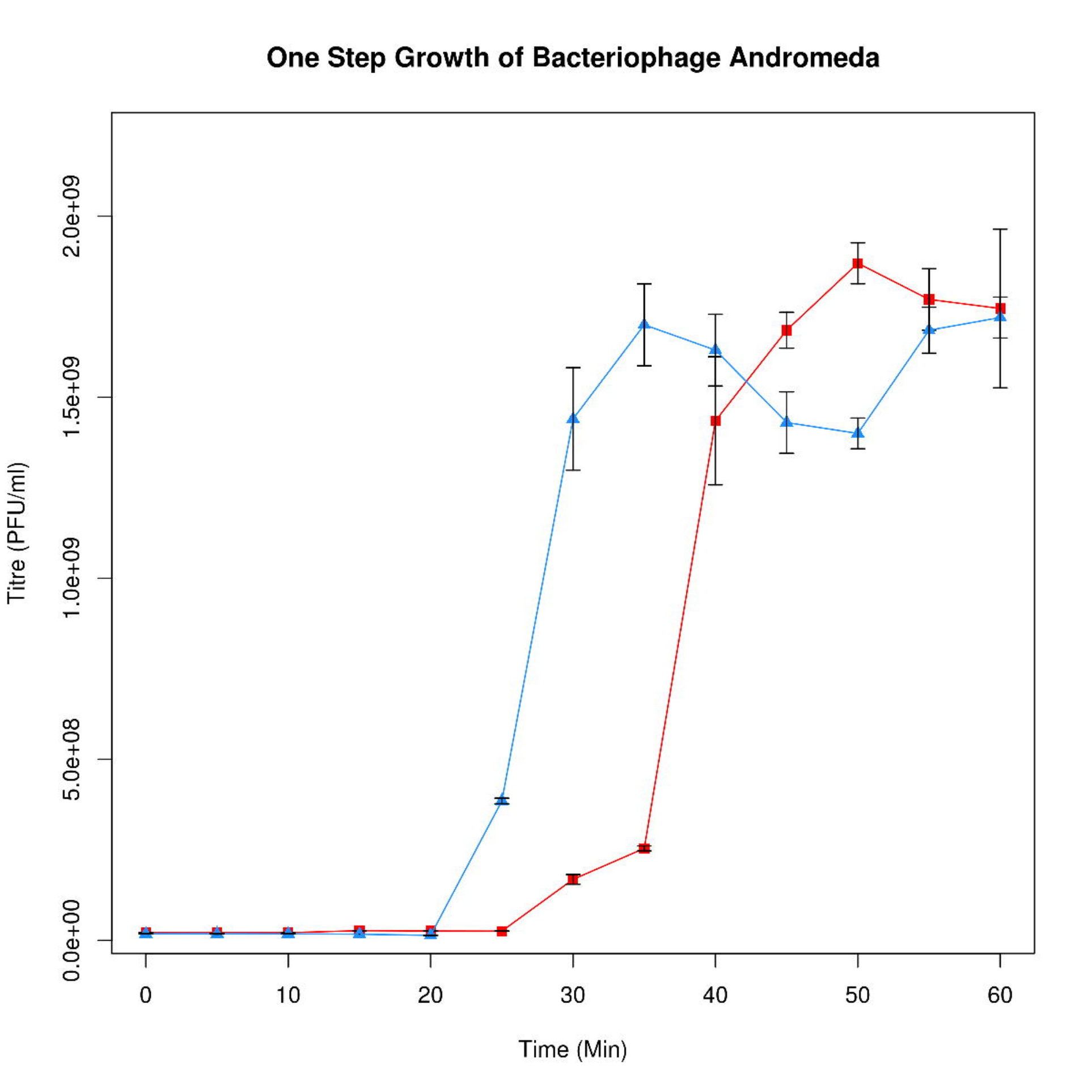
Fig 3. Genetic organisation of Andromeda and its homology to bacteriophage Bf7. A nucleotide homology levels between two phages are shown by the Mauve graphs across two collinear blocks at the top of the figure. Arrows on the genome map represent identified genes and show the direction of their transcription. They are coloured to reflect common functions of the encoded products. Putative genes with unidentified function are shown in grey colour. Modular and specific annotations of the predicted products are shown.

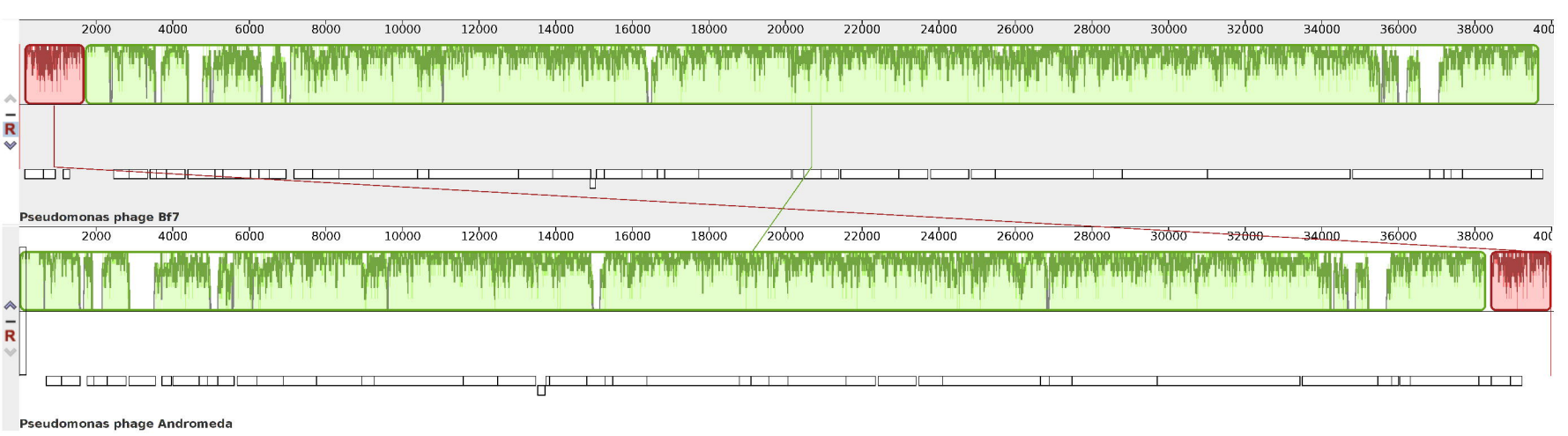
Fig 4. Visualisation of the localised single stranded interruptions in the Andromeda's genome. (1) 1 Kb ladder; (2) Untreated phage DNA; (3) NaOH treated DNA; (4) NaOH treated DNA which was preliminarily ligated by T4 DNA ligase.

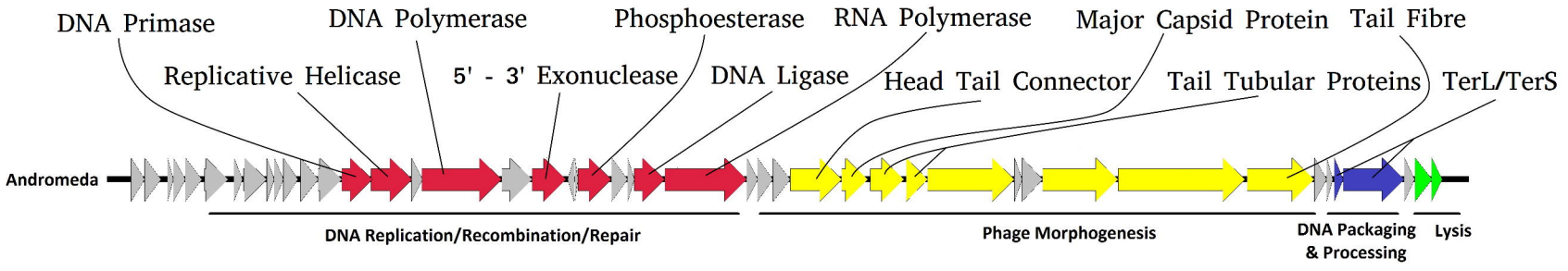
Fig 5. Grouping of Autographivirinae according genome sizes. Size distribution of the *T7viruses*, *SP6viruses* and subgroups of *phiKMVviruses* are shown and differentially coloured. Genomes of the proposed group of *ExophiKMVviruses* are shown in green.

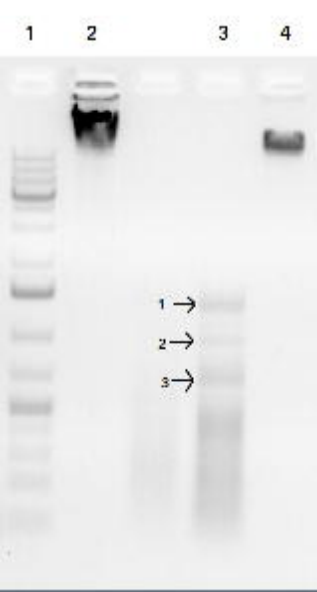


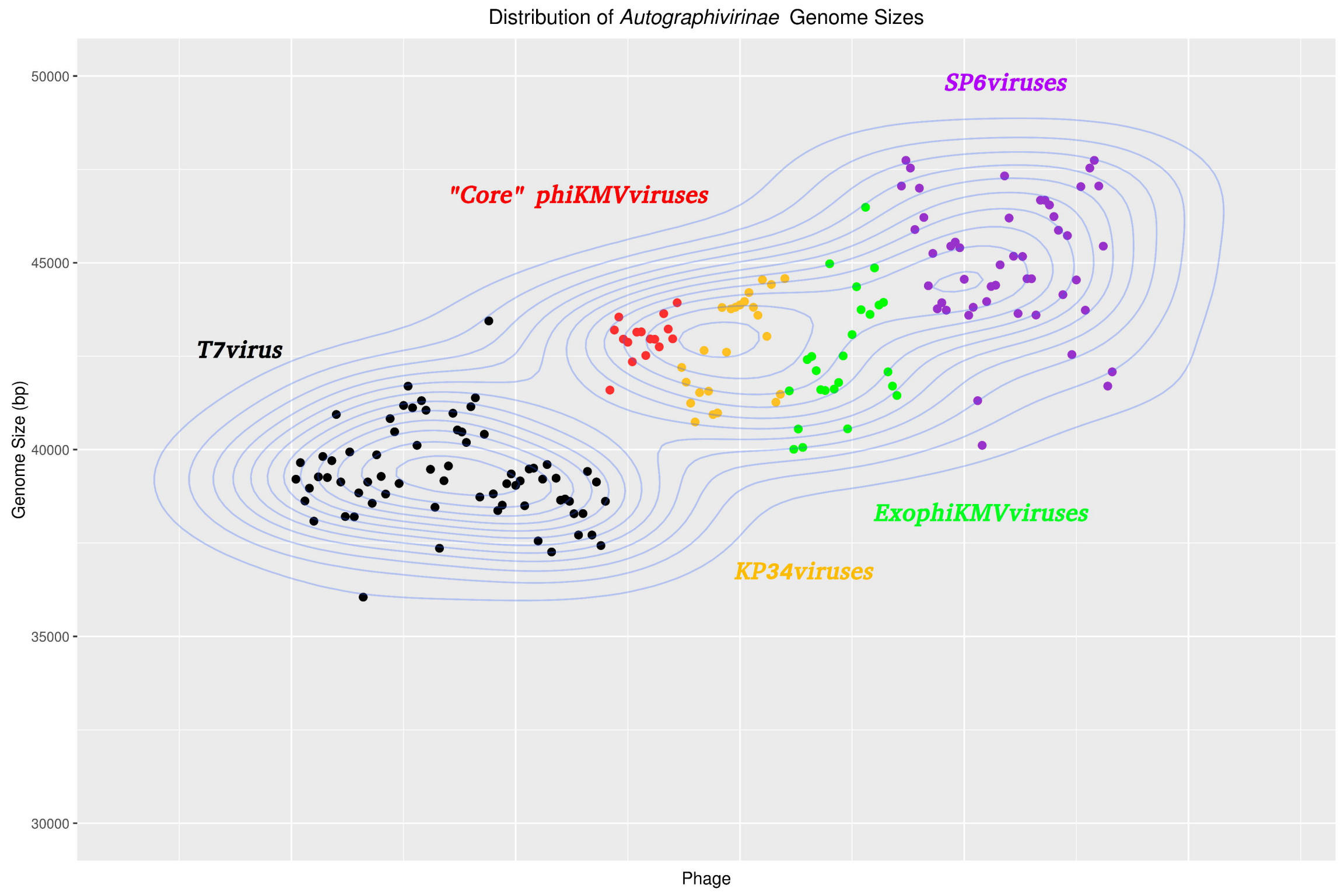

Adsorption Curve of Bacteriophage Andromeda
Number of Phages
300
250
200
150
100
50
0
2
4
6
8
10
Time (Min)


One Step Growth of Bacteriophage Andromeda
Titre (PFU/ml)
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
0
10
20
30
40
50
60
Time (Min)


2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
32000
34000
36000
38000
Pseudomonas phage Bf7
Pseudomonas phage Andromeda

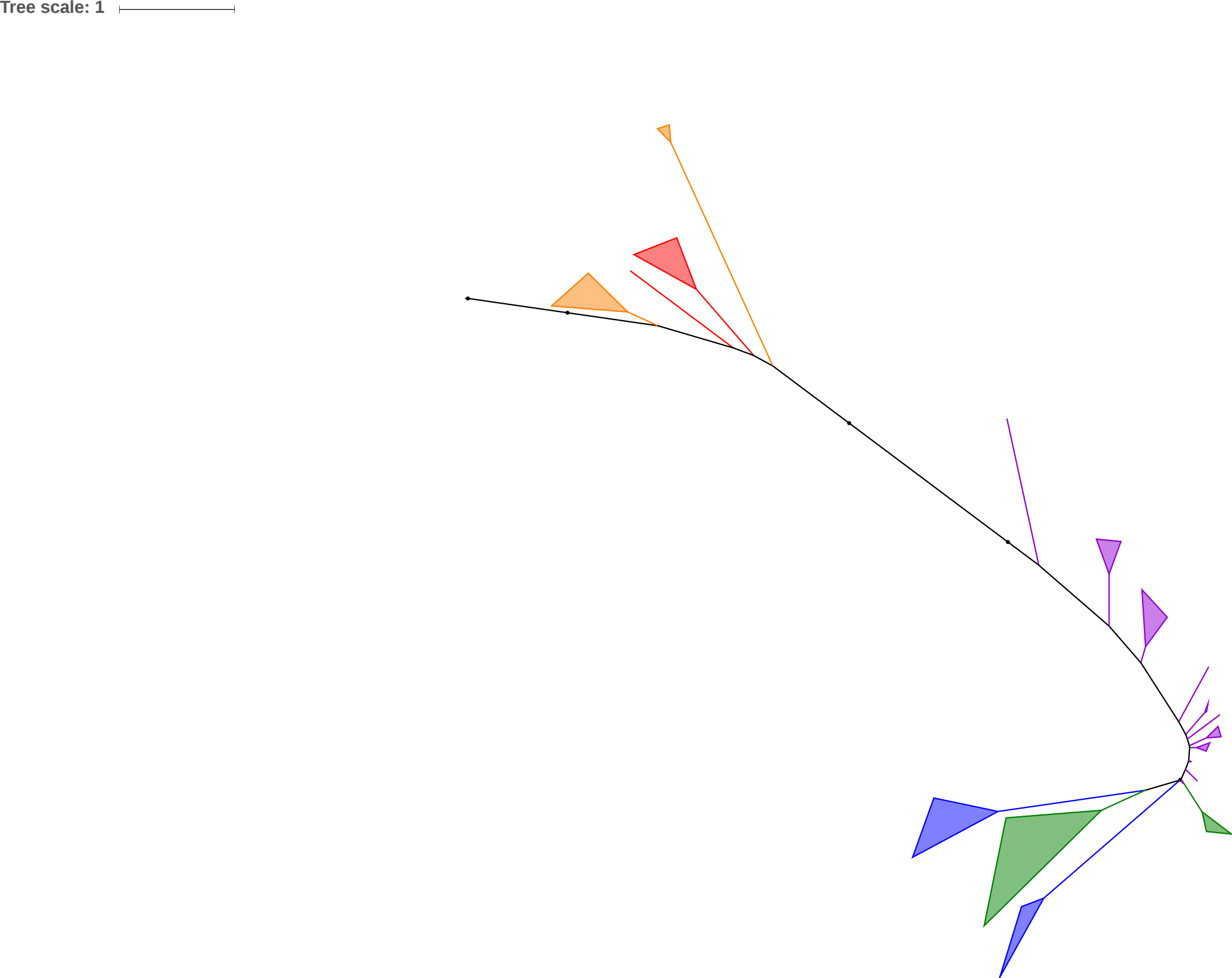

DNA Primase
DNA Polymerase
Phosphoesterase
RNA Polymerase
Major Capsid Protein
Tail Fibre
Replicative Helicase
5' - 3' Exonuclease
DNA Ligase
Head Tail Connector
Tail Tubular Proteins
TerL/TerS
Andromeda
DNA Replication/Recombination/Repair
Phage Morphogenesis
DNA Packaging & Processing
Lysis

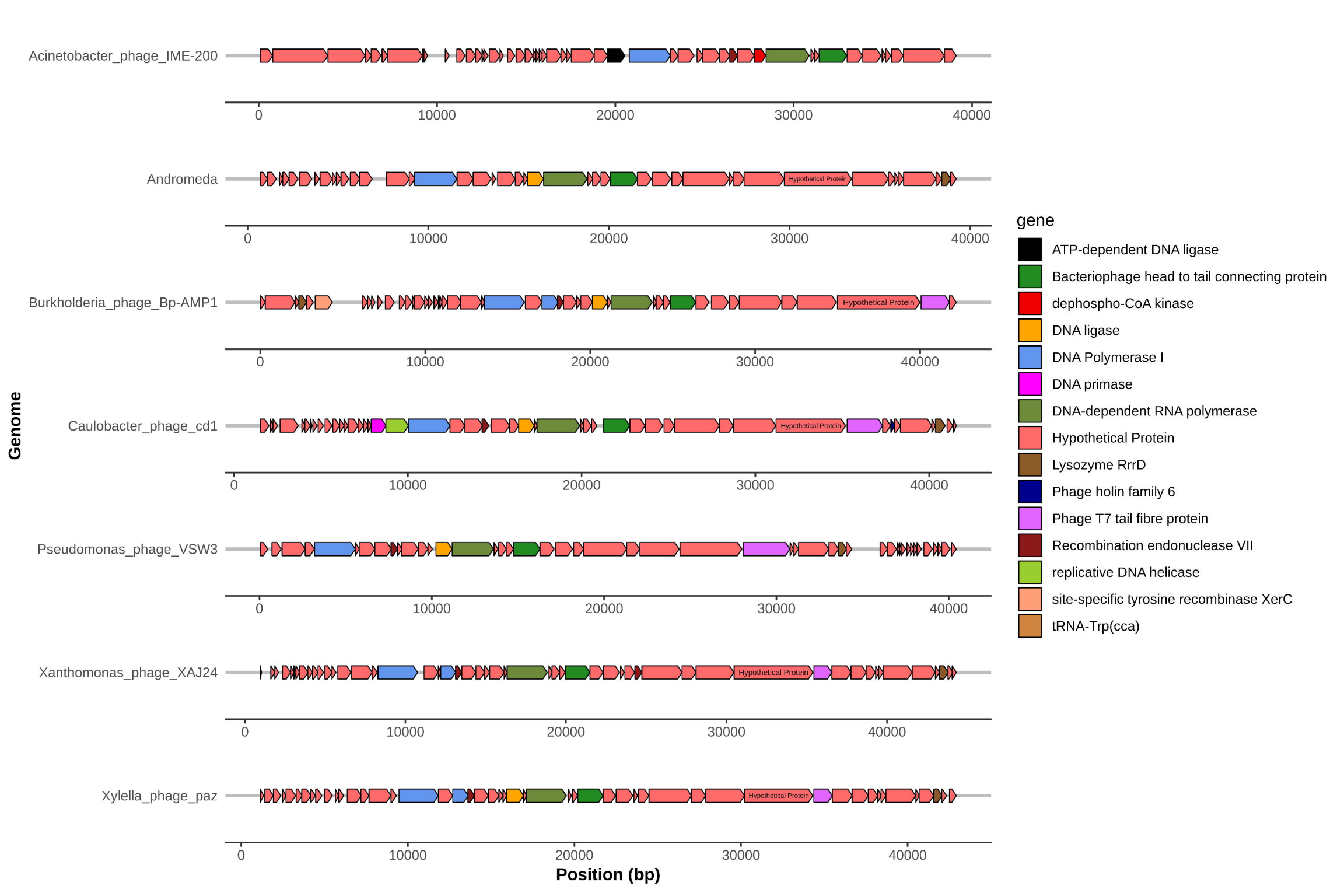

1
2
3
4
1
2
3

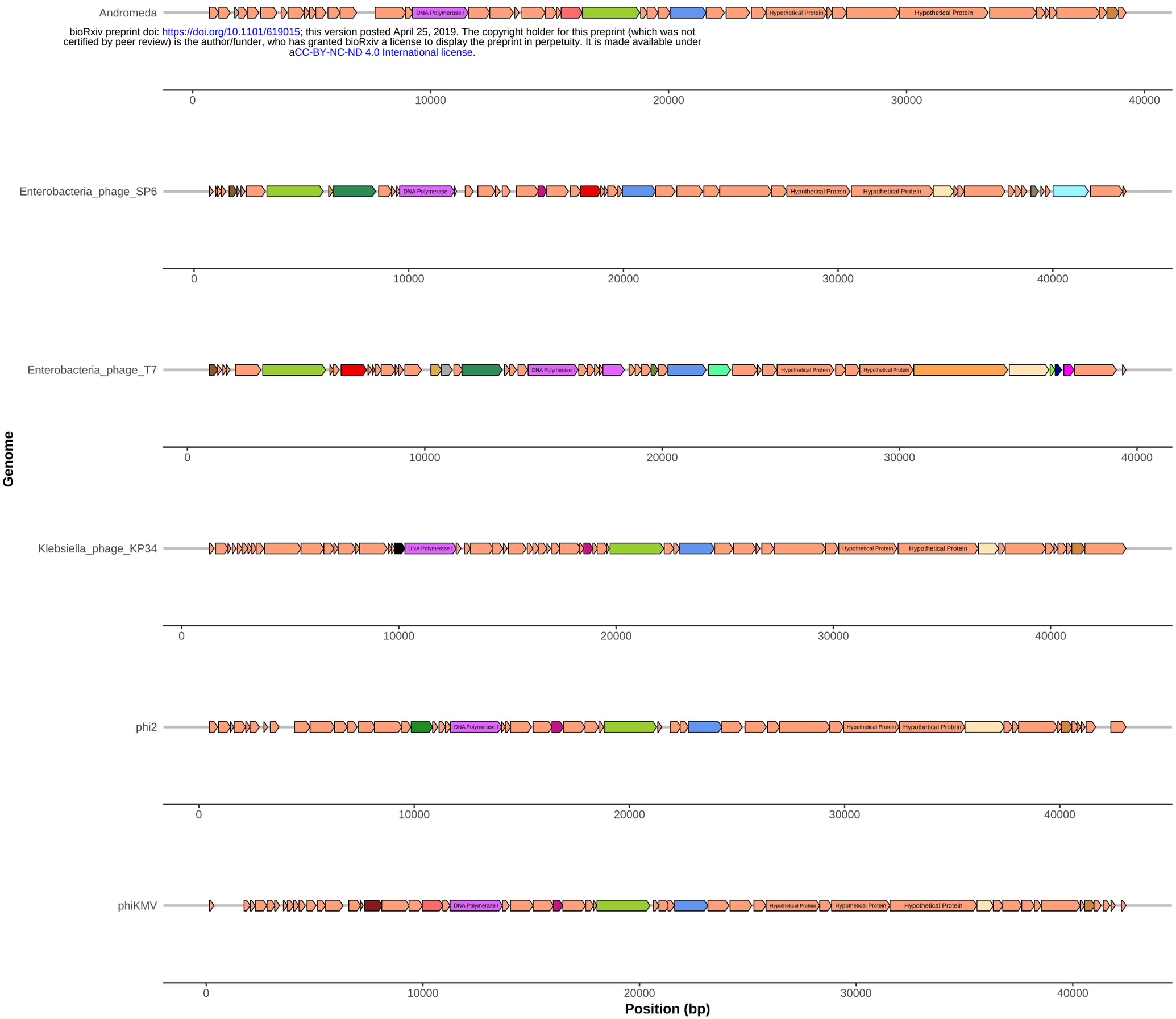

Distribution of *Autographivirinae* Genome Sizes
50000
45000
40000
35000
30000
Genome Size (bp)
Phage
SP6viruses
"Core" phiKMVviruses
T7virus
ExophiKMVviruses
KP34viruses

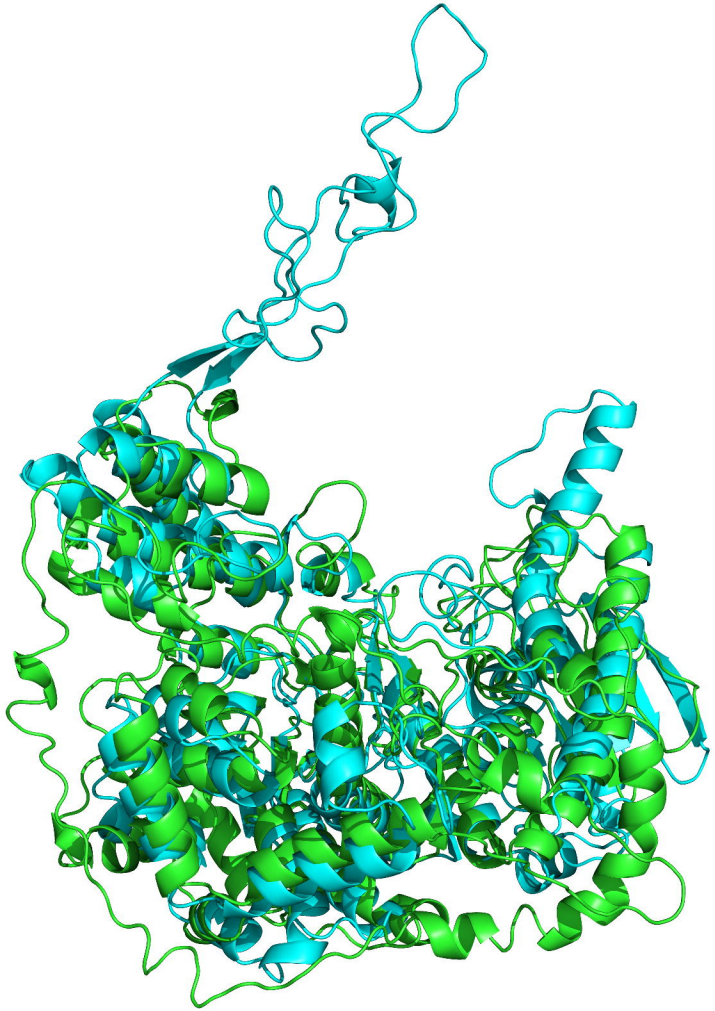
https://doi.org/10.1101/619015

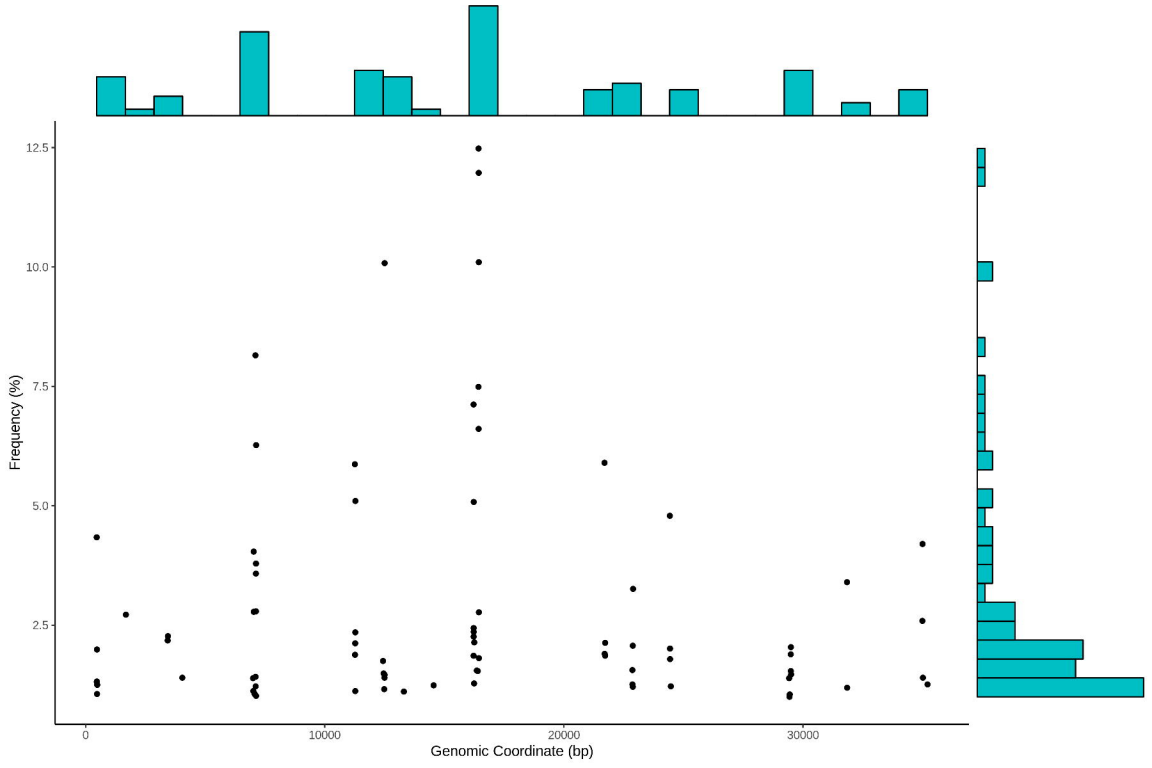
Tree scale: 1

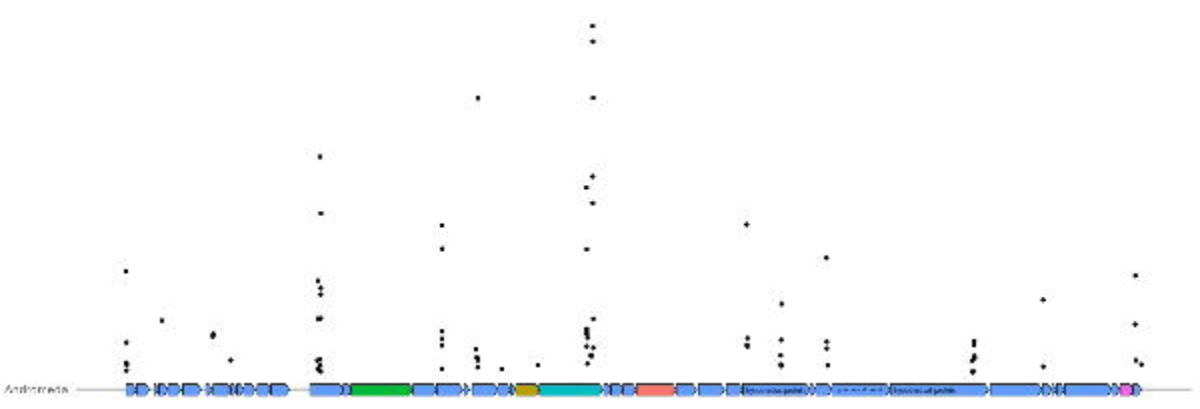

Tree scale: 1

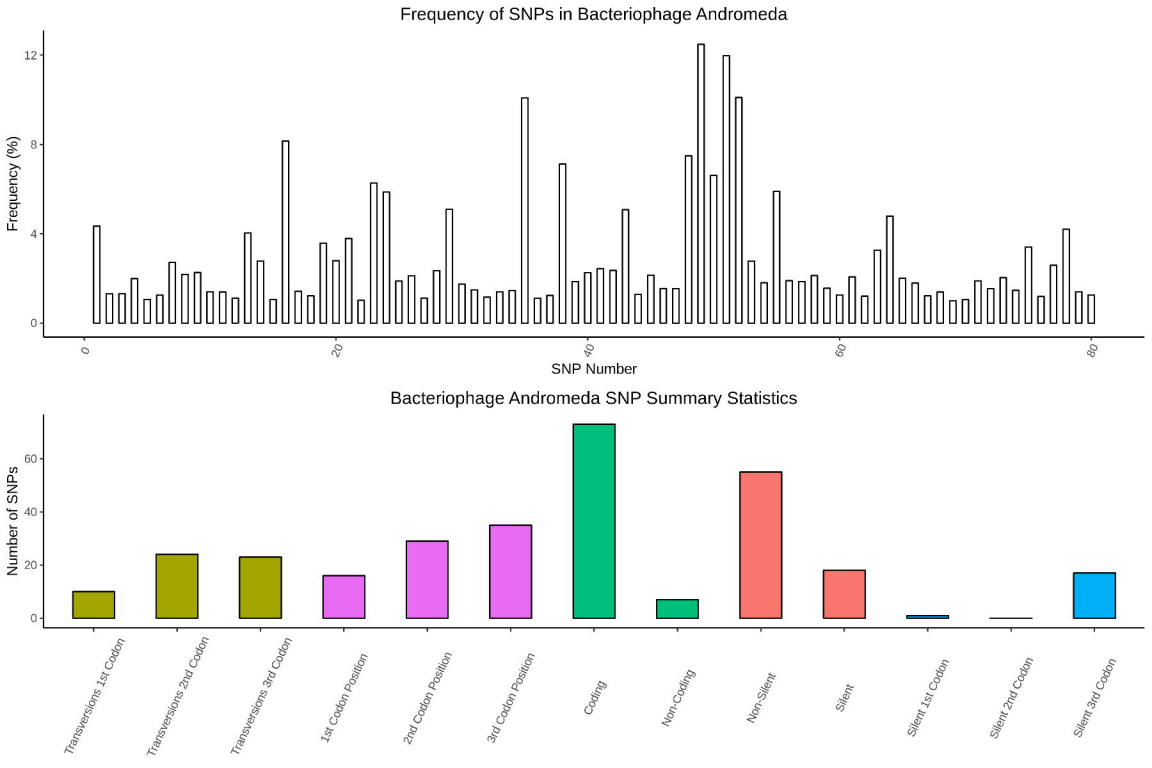

Acinetobacter_phage_IME-200
Andromeda
Burkholderia_phage_Bp-AMP1
Caulobacter_phage_cd1
Pseudomonas_phage_VSW3
Xanthomonas_phage_XAJ24
Xylella_phage_paz
Hypothetical Protein
0
10000
20000
30000
40000
Genome
Position (bp)
gene
ATP-dependent DNA ligase
Bacteriophage head to tail connecting protein
dephospho-CoA kinase
DNA ligase
DNA Polymerase I
DNA primase
DNA-dependent RNA polymerase
Hypothetical Protein
Lysozyme RrrD
Phage holin family 6
Phage T7 tail fibre protein
Recombination endonuclease VII
replicative DNA helicase
site-specific tyrosine recombinase XerC
tRNA-Trp(cca)






12.5
10.0
7.5
5.0
2.5
Frequency (%)
0
10000
20000
30000
Genomic Coordinate (bp)




Frequency of SNPs in Bacteriophage Andromeda
Frequency (%)
0
4
8
12
0
20
40
60
80
SNP Number
Bacteriophage Andromeda SNP Summary Statistics
Number of SNPs
0
20
40
60
Transversions 1st Codon
Transversions 2nd Codon
Transversions 3rd Codon
1st Codon Position
2nd Codon Position
3rd Codon Position
Coding
Non-Coding
Non-Silent
Silent
Silent 1st Codon
Silent 2nd Codon
Silent 3rd Codon