Discovery of cyanophage genomes which contain mitochondrial DNA polymerase

Yi-Wah Chan[1], Remus Mohr[2], Andrew D. Millard[3], Antony B. Holmes[4,5], Anthony W. Larkum[6], Anna L. Whitworth[7], Nicholas H. Mann[3], David J. Scanlan[3], Wolfgang R. Hess[2] and Martha R. J. Clokie[8]

[1]Reproductive Health, Clinical Sciences Research Institute, University Hospital, Coventry, UK

[2] Institute of Biology III, Genetics and Experimental Bioinformatics, Faculty of Biology, University of Freiburg, Freiburg, Germany

[3]School of Life Sciences, University of Warwick, Coventry, UK

[4]Department of Biomedical Informatics, Columbia University College of Physicians and Surgeons

[5]Center for Computational Biology and Bioinformatics, Columbia University, College of Physicians and Surgeons,

[6]School of Biological Sciences, University of Sydney, Sydney, Australia

[7]Department of Chemistry, University of Warwick, Coventry, UK

[8]Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, UK.

Corresponding author: Martha Clokie[9] (Tel: +44 (0)116 252 2959, Fax: +44 (0)116 252 5030, Email: mrjc1@le.ac.uk)

Keywords: *Acaryochloris*, cyanophage, evolution, mitochondrial DNA polymerase gamma,

Running Head: Cyanophage genomes contain mtDNA polymerase

DNA polymerase γ is a family A DNA polymerase responsible for the replication of mitochondrial DNA in eukaryotes. The origins of DNA polymerase γ have remained elusive because it is not present in any known bacterium, though it has been hypothesized that mitochondria may have inherited the enzyme by phage-mediated nonorthologous displacement. Here, we present an analysis of two full-length homologues of this gene which were found in the genomes of two bacteriophages which infect the chlorophyll-*d* containing cyanobacterium *Acaryochloris marina*. Phylogenetic analyses of these phage DNA polymerase γ proteins show they branch deeply within the DNA polymerase γ clade and therefore share a common origin with their eukaryotic homologues. We also found homologues of these phage polymerases in the environmental Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) database, which fell in the same clade. An analysis of the CAMERA assemblies containing the environmental homologues together with the filter fraction metadata indicated some of these assemblies may be of bacterial origin. We also show that the phage encoded DNA polymerase γ is highly transcribed as the phage genomes are replicated. These findings provide data which may assist in reconstructing the evolution of mitochondria.

Introduction

DNA polymerase is a high fidelity enzyme that catalyses the polymerisation of deoxyribonucleotides into a DNA strand (Wang 1996). There are currently seven families of DNA polymerase (A, B, C, D, X, Y and RT) (Filée et al. 2002; Harada et al. 2005). DNA polymerase is essential to all known forms of life on earth, and therefore, the genes, which encode DNA polymerase, have served as a useful tool to interpret evolutionary relationships between different organisms (Braithwaite, Ito 1993; Filée et al. 2002; Labonté et al. 2009).

However, Filée et al. observed that no DNA polymerase family was universally conserved between the Archaea, Bacteria and the Eukarya and that there was no simple explanation for this. These authors performed a phylogenetic study of family A DNA polymerases which included those from bacteria, phage, metazoa and fungi,which indicated that bacteriophages were likely to play an important role via lateral gene transfer (LGT) in the evolution of replication machinery. Specifically, they found that eukaryotic mitochondrial DNA (mtDNA) polymerase γ sequences formed a sister group to bacteriophage DNA polymerases. DNA polymerase γ is a family A polymerase found only in mitochondria (Bolden et al. 1977). The origin of mitochondrial DNA polymerase γ has remained elusive because a direct descendant of the bacterium, which led to the modern mitochondria of eukaryotes, has never been identified. Furthermore, the gene has not been previously identified in any non-mitochondrial genomes. Filée et al. thus proposed that it originated from a phage. This phage-derived gene was proposed to have replaced that of the protomitochondrion by nonorthologous displacement. Further phylogenetic studies indicated that T3/T7-like prophages were present in several proteobacterial genomes (Filée and Forterre 2005). As mitochondria are thought to be derived from α-proteobacteria, this provided tantalising evidence for the phage-replacement hypothesis. However, the hypothesis

has remained speculative because prior to this study no DNA polymerase γ genes have ever been found in a phage genome or indeed from any other non-eukaryotic genome.

Non-presence within a phage genome may be due to the fact that relatively few bacteriophages have had their entire genomes sequenced (to date 461 *Caudovirales* tailed-phages have been sequenced, November 2010). One group of bacteriophages that have never been previously isolated are those that infect *Acaryochloris* spp., a genus of cyanobacterium with a sedentary lifestyle that lives either symbiotically with, or epiphytically on, metazoans or associated with rocks (Miyashita et al. 1996; Murakami et al. 2004).

We isolated and sequenced the genomes of two novel bacteriophages that infect *Acaryochloris* spp and we show that they encode a gene that appears to be related to mitochondrial polymerase γ. This is the first time such a gene has been found either in a phage or a bacterium, and its presence has implications for the origin of mitochondria. We also show that the gene is highly expressed at the initial stages of infection and thus is likely to be functional during bacteriophage DNA replication.

Materials and Methods

Phage isolation

Phages were isolated from material collected on the reef flat at Heron Island, Australia, in March 2006 (Kühl et al., 2005; Larkum and Kühl 2005). Phage nomenclature: A: *Acaryochloris*, HI: Heron Island, S: *Siphoviridae*. Both phages came from the same crude unfiltered seawater sample obtained by incubating an *Acaryochloris*-associated ascidian (*Lissoclinum patella*) with SM buffer (50 mM Tris-HCl (pH 7.5), 100 mM NaCl and 8 mM $MgSO_4$).

Sequencing of *Acaryochloris* phage genomes

Phage DNA was extracted (Wilson et al. 1993) and the genomes of A-HIS1 and A-HIS2 were commercially sequenced by AGOWA (LGC) using a shotgun approach. Phage DNA was sonicated to 2–3 kb fragments and the ends were polished with T4 DNA polymerase/T4 polynucleotide kinase and fragments were ligated into pMCL200. Clones were end-sequenced from both sides to get a 6× coverage using standard Dye Terminator Sequencing (BigDye version 3.1) on a 96 capillary 3730XL DNA Analyzer (Applied Biosystems). Remaining gaps were closed by walking reads on shotgun clones or on polymerase chain reaction (PCR) products spanning gaps. Assembly was performed in GAP4 (Staden Package). No gap was left that was not covered either by a shot gun clone or by a PCR product for both phage genomes. Therefore, the genome DNA molecules are circular. ORFs were predicted by GeneMark.hmm 2.0 (Besemer and Borodovsky 1999).

Phylogenetic analysis

Amino acid alignments were created in Clustal X 1.83 (Thompson et al. 2002). Sequences for alignment were selected by Basic local alignment search tool (BLAST)/CAMERA-searches. The DNA polymerase A alignment was based on the alignment by Filée et al. (http://www-

archbac.u-psud.fr/Projects/dnapol/Ali_polA.htm). The Mus308 sequences were removed as they were not relevant to our study. Bacterial, phage and the *A. marina* phage sequences were subsequently added to this alignment.

A search for 'DNA polymerase gamma' was performed on the NCBI database and sequences were selected to include in the alignment. CAMERA sequences were obtained from the metagenomic ORFs (open reading frames) peptides dataset by BLAST analysis using the phage mtDNA polymerases. This search only recovered metazoan and fungal mtDNA polymerase sequences and not those of other eukaryotes such as plants or protists. The final alignment was formed by combining sequential alignments and appropriate truncations. First, the deletions made to form the original alignment were recreated by comparing a new bacterial or equivalent bacterial sequence to one from the original alignment. Similar sequences were then aligned (e.g. DNA polymerase γ sequences). The new sequence alignments were then truncated based on the original deletions. The alignment was then manually adjusted by eye in BioEdit. Alignments were converted to nexus format using Mesquite 2.01 for MrBayes with interleave=yes. The family A DNA polymerase phylogenetic tree was computed by MrBayes 3.1.2 (Huelsenbeck, Ronquist 2001) using a WAG model, rates=invgamma, 1,000,000 generations, sampling frequency of 100 and 25 % burnin. The sump function was used to summarise the sampled parameter values, to assess the log likelihood values and to check that the potential scale reduction factor approached 1. The phylogenetic tree was drawn in Treeview 3.2 in unrooted format (Page 1996).

RNA extraction and cDNA synthesis

A. marina was grown to the exponential phase in ASW supplemented with 0.5 gL^{-1} $NaHCO_3$ in 4 L glass jars at 28°C under 30 µmol photons $m^{-2}s^{-1}$ of continuous white fluorescent light. For each experimental replicate, 1 L of exponential phase culture was infected with phage A-

HIS1 and 1 L was infected with phage A-HIS2 in sterile 1 L conical flasks. The phages were inoculated to a multiplicity of infection of 1. Cultures were subsequently shaken under 30 μmol photons $m^{-2}s^{-1}$ and 100 mL of sample was collected per infected culture 10 minutes prior to 2, 4, 6 and 8 h after infection. The samples were centrifuged in a Hettich Rotina 46R centrifuge at 4,754 g for 15 min at 4°C and the supernatant was poured away. The cell pellet was resuspended in 0.75 mL of TRIzol® and frozen in liquid nitrogen at -80°C. RNA was extracted as described in (Clokie et al. 2006). RNA pellets were resuspended in 90 μL RNase-free water (Qiagen) with 10 μL DNase buffer and 4 μL Ambion TURBO™ DNase. Samples were then left for 20–30 min at 37°C. A Qiagen RNeasy Mini Kit was used to purify the RNA samples as described by the manufacturer. For the final step, RNA was eluted from the column with 50 μL of RNase-free water heated to 50°C in a water bath. PCR was used to check for host and phage DNA contamination using primers for the *Acaryochloris rpo*B gene and primers for the putative major capsid gene of both *Acaryochloris* phages, respectively (Supplementary Table S1). If a PCR product could be amplified using the phage and host specific primers, the RNA purification procedure was repeated until no such PCR product could be detected. For cDNA synthesis a fixed amount of RNA for each time point sample was used per 20 μL cDNA reaction. cDNA was synthesised using a SuperScript™ III Reverse Transcriptase kit (Invitrogen) and a Biometra TGradient or T3000 Thermocycler. cDNA samples were stored at -80°C.

Real-time quantitative PCR analysis

Primers for real-time quantitative PCR (qPCR) were designed using the Taq-Man® 'Probe & Primer Design' option (default parameters) in Primer Express™ v.2.0.0 (Applied Biosystems™) and acquired from Invitrogen™. The primers used in this study are listed in Supplementary Table S1. qPCR reactions were prepared in 96 well plates by combining 12.5

μL SYBR® Green PCR Master Mix (Applied Biosystems^{TM}), 9.5 μL nuclease free water (Qiagen) and 1 μL each of the forward and reverse qPCR primers (stock concentration of 3.75 pmol μL^{-1} for final concentration of 150 nM). Finally 1 μL template was added, i.e. cDNA (or genomic DNA for standard curves). No template controls were carried out for each primer set. Plates were subsequently sealed with adhesive covers (ABgene®) and analysed using an Applied Biosystems^{TM} 7500 Fast Real-Time PCR System. The program used was: 50 °C for 2 min, 95 °C for 10 min and 40 cycles of 95 °C for 15 s. This was followed by 95 °C for 15 s, 60 °C for 20 s, 95 °C for 15 s and 60 °C for 15 s at the end of each run to perform a melting curve analysis. Results were analysed using Applied Biosystems^{TM} Sequence Detection Software version 1.4 (7500 Fast System). Intra-plate variation was monitored by including triplicate reactions for each of A-HIS1 ORF 14, AHIS1 ORF 20 and A-HIS1 ORF 30 with a fixed amount of A-HIS1 genomic DNA (5 ng).

Standard curves were used to check similar amplification efficiency across a range of DNA concentrations, using a standard primer concentration of 150 nM for all primers, and were also used for absolute quantitation of transcript abundance. A serial dilution was performed with a known mass of genomic DNA (phage or host), m_{DNA}. The dilutions ranged between 150 ng (the most concentrated) and ~70 fg for host genomic DNA and ~0.5 fg for phage genomic DNA (the lowest concentration corresponds to ~8 gene copies). Standard curves (scatter plots) of C_T values versus log (concentration of genomic DNA) were generated with either genomic DNA from the host *A. marina* for host genes or phage genomic DNA from A-HIS1 or A-HIS2 for A-HIS1 or A-HIS2 genes, respectively. C_T is the cycle value at a certain threshold in the exponential phase of the fluorescence signal from the SYBR® Green in the qPCR reaction. Standard curve equations and corresponding R^2 values for each gene were calculated in Microsoft® Office Excel. Amplification efficiency, E (%), was calculated using $E = (10^{-1/m} -1)*100$ where *m* is the slope value obtained from the line of

best fit of the standard curve. R^2, the coefficient of determination, is a measure of how close the line of best fit corresponds to the actual data. The criteria for the parameters were: $-3.0 < m < -3.9$, $R^2 > 0.985$ and $80 < E < 110$ % as is generally accepted (Sigma-Aldrich 2008). The equation of the standard curve was $C_T = m*x + c_i$, where x is log (DNA concentration (ng)) and c_i is the y-axis intercept. x was converted to log (gene copy number) (ABI 2003) so that given a C_T value, a copy number could be computed. The parameters for the different forms of the standard curve for each gene tested are detailed in Supplementary Table S2.

Results and Discussion

Phage DNA polymerase γ

We identified mtDNA like polymerases in the genomes of both A-HIS1 and A-HIS2 using BLAST analysis. To establish their relationships with related sequences, we performed a phylogenetic analysis of the phage mtDNA polymerases, using the DNA polymerase A domain alignment by Filée et al. (2002) as a framework (http://www-archbac.u-psud.fr/Projects/dnapol/Ali_polA.htm). We included homologous DNA polymerase γ sequences from metazoa and fungi in the analysis. Finally, we included homologous sequences retrieved from the metagenomics data repository CAMERA database (Seshadri et al. 2007). Twelve assemblies which contained a partial or complete mtDNA polymerase A conserved domain like that of the phages were identified from the CAMERA database. Of these, three (A05ORF9, A06ORF10 and A08ORF1) could be completely aligned and one (A02ORF1) was truncated at the C-terminal end of the alignment (see Supplementary Data S1, S2 and Table S1). These four sequences were therefore included in our analysis.

We identified four main clades in the phylogenetic tree (Fig. 1, see Supplementary Table S4 for sequence details): a *Siphoviridae* clade, a bacterial clade (including both bacteria and cyanobacteria), a *Podoviridae* clade and a DNA polymerase γ clade. The *Myoviridae* sequences included did not form a specific clade. In particular, the myovirus *Bacillus* phage SPO1 fell deeply in the bacterial clade (suggesting it may be of bacterial origin). Intriguingly, the two *Acaryochloris* phage and CAMERA mtDNA polymerases rooted with the mtDNA polymerase γ clade of metazoa and fungi. Although the support value for the node where coliphage phiEcoM-GJ1 branches off is not supported 0.51, this does not affect interpretation of the relationships that are seen in the phylogenetic tree, which are the focus of this

manuscript. Similarly, were the tree collapsed at this node, it would not change the interpretation of the tree in terms of the clades that it forms.

The observation that *Acaryochloris* phage sequences cluster within the mtDNA polymerase γ clade provides the first evidence that this gene may have originated in a bacteriophage, supporting the hypothesis of Filée et al. (2002). We then looked at the identity of the other predicted ORFs in the CAMERA assemblies on which the phage mtDNA polymerase-like genes were found to see whether we could identify the origin of each assembly (i.e. whether they were likely to be from other bacteriophages or bacteria).

Of the four phage mtDNA polymerase γ-like sequences from CAMERA included in the phylogenetic tree of family A DNA polymerases, A05ORF9 and A06ORF10 may have originated from phage or bacteria based on the predicted gene content of the assemblies analysed by BLAST (Fig. 1, see Supplementary Table S3 for BLAST summary). The other two homologues (A02ORF1 and A08ORF1) are of unknown origin as they were the only ORF in the respective assemblies. One ORF in A05 produced a BLASTP hit to the cyanophage protein S-PM2p158 (a potential lipoprotein A). S-PM2 is the only known phage containing this ORF, and all other homologues are found in bacteria and eukaryotes. These analyses suggest that perhaps some of these scaffolds arose from extant marine bacteria which are descendants of protomitochondrion-like bacteria. Indeed, if these environmental sequences are from an extant lineage of bacteria, the phylogenetic tree in Fig. 1 shows that they do not fall within the bacterial clade and therefore represent a new bacterial phylum according to this gene.

Expression of phage DNA polymerase γ

Primers were designed to observe the expression of phage DNA polymerase γ at different stages during the bacteriophage infection cycle. Amplification efficiency was similar across

the entire dynamic range (see Materials and Methods, data not shown). To aid interpretation of the expression data, primers were also designed to amplify the putative major capsid gene of the bacteriophage as we know these genes are usually highly expressed and their expression reaches a peak at the end of an infection cycle (Clokie et al. 2006). The putative major capsid protein is termed as such since it does not have homology with any other capsid proteins but is the most abundant protein on an SDS-PAGE gel (data not shown). In all other phages that have been characterised on SDS-PAGE gels the most abundant protein corresponds to the major capsid protein.

Samples were collected at 2, 4, 6 and 8 h based on phage growth parameters that were determined (data not shown). The 8 h time point was used to affirm the level of expression after lysis. For both phages the eclipse period (i.e. the period from phage infection to formation of mature phage particles within host cells) was from 0–3.25 h and the latent period (i.e. the period from phage infection to release of phage particles from cells) took a further 1.75 h, so lasted ~5 h. Interestingly, the copy number for the phage mtDNA polymerase γ reached a peak of 57×10^3 and 113×10^3 copies at 2 h for A-HIS1 and A-HIS2, respectively, before decreasing over the remainder of the infection cycle (Fig.2A). This indicated that mtDNA polymerase γ is likely to be expressed early in the replication cycle, considering that for both phages the level of expression of this gene decreased between 2 and 4 h and their eclipse period is 3.25 h.

In comparison, the putative major capsid showed an increase in expression between 2 and 4 h after infection but then decreased between 4-8 h post-infection. In particular, the expression profiles of the putative A-HIS1 and A-HIS2 major capsid genes were consistent with the 5 h latent period of the phages. It was also noted that A-HIS2 produces ~18 times more copies of its putative major capsid than A-HIS1 (Fig.2B and C). However, the significance of this is not yet known.

In eukaryotes, mtDNA polymerase γ replicates mitochondrial DNA (Bolden et al. 1977). Although the expression work shows that the mtDNA polymerase γ is highly expressed in bacteriophage A-HIS1 and A-HIS2, whether these transcripts are translated and subsequently produce functional protein remains to be determined. In any case, this data suggests that the phage mtDNA polymerase γ may play a role in the replication of A-HIS1 and A-HIS2 DNA during their infection, which would be the first such case of a mtDNA polymerase γ-like enzyme being used for this purpose.

The most parsimonious explanation for the presence of mitochondrial gene-like sequences inside these phage genomes is that the *Acaryochloris* phage DNA polymerases are not a recent gene acquisition directly from a eukaryote but were acquired from a proteobacterium that was, or was related to, a mitochondrial progenitor. Further indirect support for phages changing their host from proteobacteria to cyanobacteria comes from the surprising observation of cyanobacterial aminoacyl-tRNA synthetases grouping with proteobacteria instead of other cyanobacteria (Zhaxybayeva et al. 2006; Luque et al. 2008). An alternative scenario for the presence of these genes in these bacteriophage is that they were acquired in the other direction, i.e. from mitochondria. In order for this to have occurred, very specific and unusual conditions would have been necessary because phages do not infect eukaryotes. For example, a phage could have infected a bacterium that was an intracellular parasite of a eukaryotic organism, and acquired DNA from the 'cellular soup' present following the bacterial infection. Potentially, if this happened, phage gene homologs could subsequently have undergone rapid evolution. However, this scenario is clearly less parsimonious than these phages having acquired the gene from another bacterium.

Certainly, this is the first time that a mtDNA polymerase has been found in a phage genome. It is not found in the corresponding *Acaryochloris* host genome (Swingley et al. 2008), nor indeed any other cyanobacterial or bacterial genome, with the possible exception

of the bacteria-like genomes from the CAMERA data set presented in this study. Currently, it is generally accepted that *Rickettsia* spp. and other α-proteobacteria are the most closely related extant group of bacteria to mitochondria, which suggests the protomitochondrion was related to these bacteria (Andersson *et al.* 1998). DNA polymerase γ has not been found in any of these Proteobacteria and therefore its presence in mitochondria may be explained by the orthologous replacement hypothesis of Filée et al. (2002). This hypothesis states that mtDNA polymerase originated from within a phage, and indeed, the identification of the gene in the *Acaryochloris* phages A-HIS1 and A-HIS2 supports this hypothesis.

The ability to isolate phages on a host which does not possess a mtDNA polymerase γ, suggests that these phages may also infect (or, have infected in the past) a protomitochondrion-like bacteria that does contain this gene. Indeed, it is unclear how these phages came to infect the cyanobacterium *A. marina*. The hypothetical host(s) may either be extinct, and thus the polymerase may have an ancient ancestry, or, it may be extant and susceptible to these phages, which would suggest that these phages at least, have a wide host range.

Undoubtedly, viruses encode significant uncharacterised genetic diversity and they also contribute to key roles in biogeochemical cycling, host physiology, population genetics and bacterial evolution through LGT. LGT events can be difficult to reconstruct, but here, we have presented a definitive example of phages encoding a gene with a clear evolutionary history, providing insight into the origin of mtDNA polymerase γ, which has long puzzled evolutionary biologists.

Supplementary Material

Supplementary data S1 and S2 and Tables S1-S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgements

Cyanophage genome sequencing was funded by NERC grant NE/E01089X/1. YWC was supported by the Molecular Organisation and Assembly in Cells Doctoral Training Program, via the EPSRC Life Science Initiative. WRH, RM and AWL thank the German Academic Exchange Service (DAAD) for supporting international cooperation. We acknowledge the IP of Panama, Ecuador, Fr. Polynesia, the USA, Italy, Tanzania, Bermuda and Canada as the countries of origin of the read, assembly and protein data obtained from the CAMERA website http://camera.calit2.net, which were used in this study. Sequences from International waters (1) between Madagascar and South Africa and (2) 500 miles west of the Seychelles in the Indian Ocean from CAMERA were also used (the assemblies used to construct Supplementary Table S3 are listed with their respective country of origin). We also acknowledge Shaun Heaphy for helpful discussions on the manuscript and the comments of various anonymous referees. Sequence data from this article have been deposited in EMBL under the accession numbers FN393744, FN393745.

Figure Legends

Figure 1. Phylogenetic relationships amongst family A DNA polymerases.

Unrooted Bayesian phylogenetic tree constructed in MrBayes. Nodes are labelled by a circle (support=1.00) or with support values. Where direct labelling was not possible, the corresponding support values are placed next to the two labels. Scale bar: amino acid substitutions per site. * denotes previously aligned sequences (Filée et al. 2002)). Nodes labelled AxORFy refer to sequences from CAMERA (see Supplementary Table S3). Numbers in parentheses refer to sequence details provided in Supplementary Table S4.

Figure 2. Expression of *Acaryochloris* phage genes.

Absolute gene copy number to estimate transcript abundance (converted from C_T values) of *Acaryochloris* phage genes (A) mtDNA polymerase γ: A-HIS1 ORF 14 and A-HIS2 ORF 20; (B) putative major capsid: A-HIS1 ORF 73 and A-HIS2 ORF 82; and (C) graph for A-HIS1 as in (B) re-plotted for clarity. Phages were infected at a MOI=1 at time zero. For A-HIS1, n=2 and A-HIS2, n=3. Error bars are one standard deviation.

References

Andersson, SG, A Zomorodipour, JO Andersson, T Sicheritz-Ponten, UC Alsmark, RM Podowski, AK Naslund, AS Eriksson, HH Winkler, CG Kurland. 1998. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* 396:133-140.

Applied Biosystems. 2003. Creating standard curves with genomic DNA or plasmid DNA templates for use in quantitative PCR (Appled Biosystems) California (CA): Carlsbad.

Besemer, J, M Borodovsky. 1999. Heuristic approach to deriving models for gene finding. *Nucleic Acids Res*. 27:3911-3920.

Bolden, A, G Pedrali-Noy, A Weissbach. 1977. DNA polymerase of mitochondria is a gamma-polymerase. *J Biol Chem*. 252:3351-3356.

Braithwaite, DK, J Ito. 1993. Compilation, alignment, and phylogenetic relationships of DNA polymerases. *Nucleic Acids Res*. 21:787-802.

Clokie, MR, J Shan, S Bailey, Y Jia, HM Krisch, S West, NH Mann. 2006. Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environ Microbiol*. 8:827-835.

Filée, J, P Forterre. 2005. Viral proteins functioning in organelles: a cryptic origin? *Trends Microbiol*. 13:510-513.

Filée, J, P Forterre, T Sen-Lin, J Laurent. 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol*. 54:763-773.

Harada, F, T Nakano, T Kohno, S Mohan, H Taniguchi, K Sano. 2005. RNA-dependent DNA polymerase (RT) activity of bacterial DNA polymerases. *Bull Osaka Med Coll*. 51:35-41.

Huelsenbeck, JP, F Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754-755.

Kühl M, Chen M, Ralph PJ, Schreiber U, Larkum AW. 2005. Ecology: a niche for cyanobacteria containing chlorophyll *d*. *Nature* 433:820.

Labonté, J, KE Reid, CA Suttle. 2009. Phylogenetic analysis indicates evolutionary diversity and environmental segregation of marine podovirus DNA polymerase gene sequences. *Appl Environ Microbiol*. 75:3634-3640.

Larkum AW, Kühl M. 2005. Chlorophyll d: the puzzle resolved. *Trends Plant Sci* 10:355-357.

Luque, I, ML Riera-Alberola, A Andujar, JA Ochoa de Alda. 2008. Intraphylum diversity and complex evolution of cyanobacterial aminoacyl-tRNA synthetases. *Mol Biol Evol*. 25:2369-2389.

Miyashita, H, H Ikemoto, N Kurano, K Adachi, M Chihara, S Miyachi. 1996. Chlorophyll *d* as a major pigment. *Nature* 383:402.

Murakami, A, H Miyashita, M Iseki, K Adachi, M Mimuro. 2004. Chlorophyll *d* in an epiphytic cyanobacterium of red algae. *Science* 303:1633.

Page, RD. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci*. 12:357-358.

Seshadri, R, SA Kravitz, L Smarr, P Gilna, M Frazier. 2007. CAMERA: a community resource for metagenomics. *PLoS Biol*. 5:e75.

Sigma-Aldrich. 2008. The Quantitative PCR Technical Guide St. Louis (MO): Sigma-Aldrich Co.

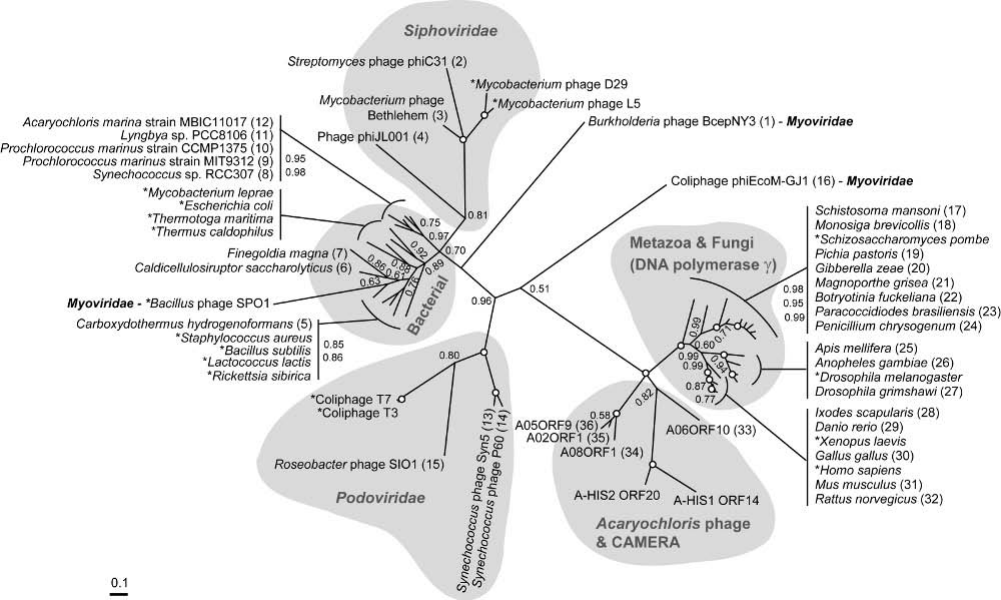
Swingley, WD*et al.* 2008. Niche adaptation and genome expansion in the chlorophyll *d*-producing cyanobacterium *Acaryochloris marina*. *Proc Natl Acad Sci USA*. 105:2005-2010.

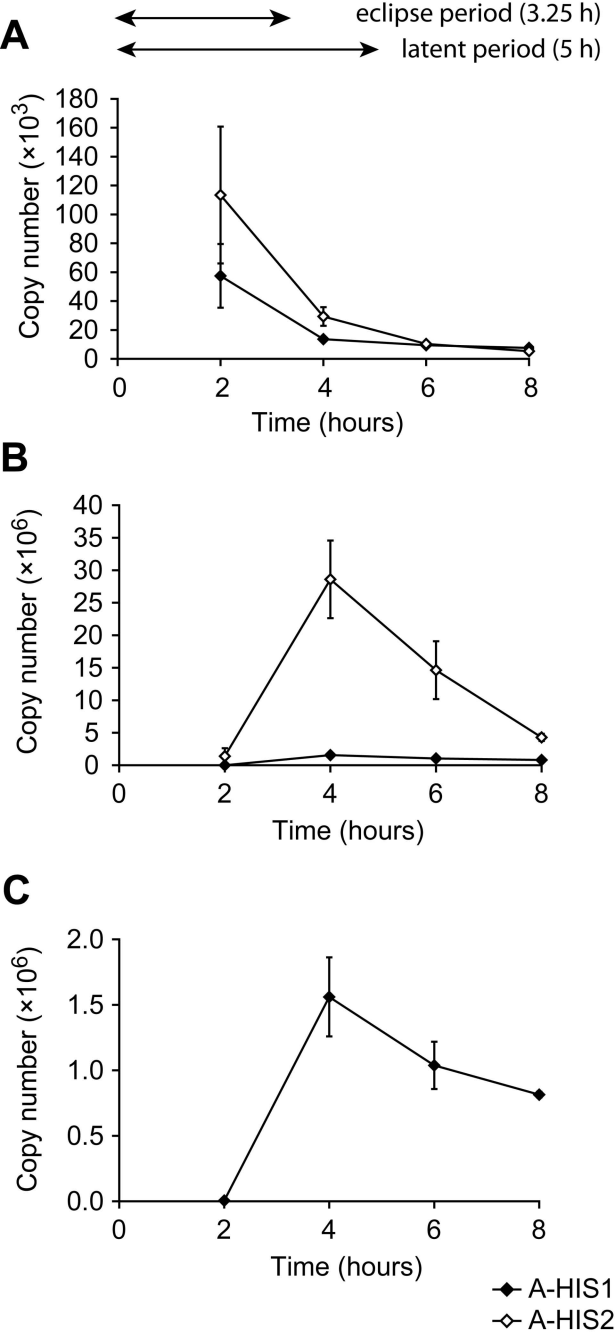
Thompson, JD, TJ Gibson, DG Higgins. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* Chapter 2:Unit 2.3.

Wang, TS-F. 1996. Cellular DNA Polymerases. In: ML DePamphilis, editor. DNA replication in eukaryotic cells. New York: Cold Spring Harbor Laboratory Press. p. 461-493.

Wilson, WH, IR Joint, NG Carr, NH Mann. 1993. Isolation and molecular characterization of five marine cyanophages propagated on *Synechococcus* sp. strain WH7803. *Appl Environ Microbiol*. 59:3736-3743.

Zhaxybayeva, O, JP Gogarten, RL Charlebois, WF Doolittle, RT Papke. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res*. 16:1099-1108.


Siphoviridae
Streptomyces phage phiC31 (2)
*Mycobacterium phage D29
*Mycobacterium phage L5
Mycobacterium phage Bethlehem (3)
Burkholderia phage BcepNY3 (1) - Myoviridae
Phage phiJL001 (4)
Acaryochloris marina strain MBIC11017 (12)
Lyngbya sp. PCC8106 (11)
Prochlorococcus marinus strain CCMP1375 (10)
Prochlorococcus marinus strain MIT9312 (9)
Synechococcus sp. RCC307 (8)
0.95
0.98
Coliphage phiEcoM-GJ1 (16) - Myoviridae
*Mycobacterium leprae
*Escherichia coli
*Thermotoga maritima
*Thermus caldophilus
0.75
0.97
0.81
0.70
Finegoldia magna (7)
Caldicellulosiruptor saccharolyticus (6)
Myoviridae - *Bacillus phage SPO1
Bacterial
Carboxydothermus hydrogenoformans (5)
*Staphylococcus aureus
*Bacillus subtilis
*Lactococcus lactis
*Rickettsia sibirica
0.85
0.86
0.96
0.51
Metazoa & Fungi (DNA polymerase γ)
Schistosoma mansoni (17)
Monosiga brevicollis (18)
*Schizosaccharomyces pombe
Pichia pastoris (19)
Gibberella zeae (20)
Magnoporthe grisea (21)
Botryotinia fuckeliana (22)
Paracoccidiodes brasiliensis (23)
Penicillium chrysogenum (24)
0.98
0.95
0.99
Apis mellifera (25)
Anopheles gambiae (26)
*Drosophila melanogaster
Drosophila grimshawi (27)
Ixodes scapularis (28)
Danio rerio (29)
*Xenopus laevis
Gallus gallus (30)
*Homo musculus (31)
Rattus norvegicus (32)
0.80
*Coliphage T7
*Coliphage T3
Roseobacter phage SIO1 (15)
Podoviridae
Synechococcus phage Syn5 (13)
Synechococcus phage P60 (14)
0.62
0.58
A05ORF9 (36)
A02ORF1 (35)
A08ORF1 (34)
A06ORF10 (33)
A-HIS2 ORF20
A-HIS1 ORF14
Acaryochloris phage & CAMERA
0.1


A
eclipse period (3.25 h)
latent period (5 h)
Copy number (×10^3)
180
160
140
120
100
80
60
40
20
0
0
2
4
6
8
Time (hours)
B
Copy number (×10^6)
40
35
30
25
20
15
10
5
0
0
2
4
6
8
Time (hours)
C
Copy number (×10^6)
2.0
1.5
1.0
0.5
0.0
0
2
4
6
8
Time (hours)
A-HIS1
A-HIS2

Supplementary Table S1. PCR and qPCR primer sequences.

Method	Gene	Primer name[a]	Primer sequence (5'-3')
PCR	*rpo*B	rpoBF	TGGGTAAAGTTACGCCCAAG
		rpoBR	CTTGTCACCCACCTGGAGTT
	putative major capsid	mcapF	GAASGCACTACCCGWCAATC
		mcapR	ATCRACGCCAGCRTAMGTAT
qPCR	mtDNA polymerase	F114	CGAAAAGCGGATCAACCTTCT
		R114	TGCGGGTAGCGAGTCCAA
		F220	CAAGGGTTCCTCGTACTGCTATG
		R220	GACGACTGCGGGCCTTAA
	RNase T	F120	GCGATGGAAATCAACGGTTT
		R120	CGCCAGATTTGCGAAAGC
	DNA pol I-like FEN	F130	CGGGTTACCCTCCGTTACTCA
		R130	TTGCGCGGCGTTAAAGA
	putative major capsid	F173	CGGTATGCTGTTACAGGACGAA
		R173	AGCGGCGGCGATAGAGT
		F282	CGTTCAGTCGGGACCAATG
		R282	CCGCAGGACCAGCGATT

[a]F/R = forward/reverse primer followed by phage number (1:A-HIS1, 2:A-HIS2) and then ORF number. F/R-114, 120 and 130 primers were used to assess intra-plate variation.

Supplementary Table S2: Standard curve parameters[a].

Phage	Gene	ORF	m	c_1	c_2	R^2	E(%)
A-HIS1	mtDNA pol γ	ORF 14	-3.4816	12.6157	37.7344	0.9987	93.74
	RNase T	ORF 20	-3.5115	11.6785	37.0132	0.9953	92.65
	DNA pol I-like FEN	ORF 30	-3.1472	10.6148	33.3207	0.9992	107.85
	putative major capsid	ORF 73	-3.8704	9.916	37.8402	0.9915	81.29
A-HIS2	mtDNA pol γ	ORF 20	-3.1561	13.2216	35.9495	0.9993	107.42
	putative major capsid	ORF 82	-3.1191	12.5364	34.9981	0.9998	109.22

[a]m is the slope of the line of best fit of the standard curve, R^2 is the coefficient of determination. The y-intercept, c_i is c_1 for x = log (DNA concentration) and c_2 for x = log (gene copy number). E is the amplification efficiency.

Supplementary Table S3. ORFs from CAMERA assemblies.

Assembly	Assembly ID	Location (region, country, habitat)	ORF[a]	Comment	Putative origin
A01	JCVI_SCAF_ 1096627020476	Lake Gatun (Panama Canal, Panama, freshwater)	A01ORF1	similar to hypothetical protein S-PM2p148/gp191 phage syn9	Phage or bacterial
			A01ORF2	-	
			A01ORF3	hypothetical protein; various hits including bacterial (specifically proteobacterial), archaeal, phage, eukaryota and *Phycodnaviridae*	
			A01ORF4	DNA polymerase gamma (no DNA pol A domain)	
A02	JCVI_SCAF_ 1096627024361		**A02ORF1**	DNA polymerase gamma (complete DNA pol A domain, possibly incomplete gene)	Unknown
A03	JCVI_SCAF_ 1101667175471		**A03ORF1**	DNA polymerase gamma (partial DNA pol A domain)	Unknown
A04	JCVI_SCAF_ 1096627138623	Punta Cormorant, Hypersaline Lagoon, Floreana Island (Galapagos Islands, Ecuador, hypersaline)	A04ORF1	putative high light inducible protein, cyanobacterial hits	Phage or bacterial
			A04ORF2	similar to hypothetical protein S-PM2p148/gp191 phage syn9	
			A04ORF3	DNA polymerase gamma (no DNA pol A domain)	
A05	JCVI_SCAF_ 1096627359208		A05ORF1	-	Phage or bacterial
			A05ORF2	-	
			A05ORF3	-	
			A05ORF4	-	
			A05ORF5	-	
			A05ORF6	-	
			A05ORF7	DPS (DNA Protecting protein under Starved conditions) domain, superfamily of ferritin-like diiron-carboxylate proteins	
			A05ORF8	rare lipoprotein A (DPBB_1 superfamily); cyanobacterial and other bacterial hits, also hit cyanophage S-PM2p158	
			A05ORF9	DNA polymerase gamma (complete DNA pol A domain, possibly incomplete gene)	

Supplementary Table S3. Continued.

Assembly	Assembly ID	Location (region, country, habitat)	ORF	Comment	Putative origin
A06	JCVI_SCAF_ 1096627384653		A06ORF1	top hit serine/threonine protein phosphatase (*Thermotoga neopolitana*)	Phage or bacterial
			A06ORF2	-	
			A06ORF3	Partial P-loop NTPase superfamily domain, top hits bacterial, less significant hits -phage	
			A06ORF4	Partial PP2Ac superfamily domain - bis(5'-nucleosyl)-tetraphosphatase PrpE, bacterial	
			A06ORF5	-	
			A06ORF6	eukaryota hits including methyl-CpG bindin domain proteins, no putative conserved domain detected	
			A06ORF7	-	
			A06ORF8	-	
			A06ORF9	-	
			A06ORF10	DNA polymerase gamma (complete DNA pol A domain, possibly complete gene)	
A07	JCVI_SCAF_ 1096627182099	Rangirora Atoll (Polynesia Archipelagos, Fr. Polynesia, Coral Reef Atoll)	A07ORF1	-	Unknown
			A07ORF2	-	
			A07ORF3	-	
			A07ORF4	-	
			A07ORF5	DNA polymerase gamma (partial DNA pol A domain)	
A08	JCVI_SCAF_ 1096627376200		**A08ORF1**	DNA polymerase gamma (complete DNA pol A domain, possibly incomplete gene)	Unknown
A09	JCVI_SCAF_ 1096628021027		**A09ORF1**	DNA polymerase gamma (partial DNA pol A domain)	Unknown
			A09ORF2	-	
			A09ORF3	-	
A10	JCVI_SCAF_ 1096628023999		A10ORF1	Fungal DNA polymerase gamma hits containing DnaQ like exonuclease superfamily domain	Unknown
			A10ORF2	DNA polymerase gamma (partial DNA pol A domain)	
A11	JCVI_SCAF_ 1101668141779	Chesapeake Bay, MD (North American East Coast, USA, estuary)	**A11ORF1**	DNA polymerase gamma (partial DNA pol A domain)	Unknown
A12	JCVI_SCAF_ 1101668711231	Cabo Marshall, Isabella Island (Galapagos Islands, Ecuador, coastal)	**A12ORF1**	DNA polymerase gamma (partial DNA pol A domain)	Unknown

[a]DNA polymerase gamma ORF numbers highlighted in bold. AxORFy denotes assembly x ORF y (the assembly numbering was arbitrary).

Supplementary Table S4. Sequence details of family A DNA polymerase alignment (Supplementary Data S1, S2).

Sequence	Name	NCBI accession number
1	*Burkholderia cenocepacia* HI2424 phage BcepNY3	ABR10600
2	*Streptomyces* (strain Norwich stock) phage phiC31	CAA07135
3	*Mycobacterium smegmatis* mc2155 phage Bethlehem	AAR89764
4	alpha-proteobacterium sp. JL001 phage phiJL001	AAT69504
5	*Carboxydothermus hydrogenoformans* Z-2901	YP_360471
6	*Caldicellulosiruptor sacchorolyticus* DSM 8903	AAR11871
7	*Finegoldia magna* ATCC 29328	YP_001691651
8	*Synechococcus* sp. RCC307	YP_001227925
9	*Prochlorococcus marinus* str. MIT 9312	YP_397731
10	*Prochlorococcus marinus* sp. CCMP1375	NP_875626
11	*Lyngba* sp. PCC8106	ZP_01624605
12	*Acaryochloris marina*	YP_001516479
13	phage *Synechococcus* sp. WH8109 Syn5	YP_001285436
14	*Synechococcus* sp. WH7803 phage P60	NP_570330
15	*Roseobacter* sp. SIO67 phage SIO1	AAG02598
16	*Escherichia coli* serotype O149:H10:F4 phage phiEcoM-GJ1	ABR68749
17	*Schistosoma mansoni*	XP_002578415
18	*Monosiga brevicollis* MX1	XP_001745166
19	*Pichia pastoris* DSMZ 70382	Q01941
20	*Gibberella zeae* PH-1	XP_385692
21	*Magnaporthe grisea* 70-15	XP_364068
22	*Botryotinia fuckeliana* B05.10	XP_001550165
23	*Paracoccidioides brasiliensis* Pb18	EEH49594
24	*Penicillium chrysogenum* Wisconsin 54-1255	CAP95318
25	*Apis mellifera*	XP_395230
26	*Anopheles gambiae* str. PEST	XP_311006
27	*Drosophila grimshawi*	XP_001988346
28	*Ixodes scapularis*	EEC18857
29	*Danio rerio*	XP_001921328
30	*Gallus gallus*	AAC60018
31	*Mus musculus*	AAA98977
32	*Rattus norvegicus*	CAB56206
33	A06ORF10	JCVI_SCAF_1096627384653
34	A08ORF1	JCVI_SCAF_1096627376200
35	A02ORF1	JCVI_SCAF_1096627024361
36	A05ORF9	JCVI_SCAF_1096627359208