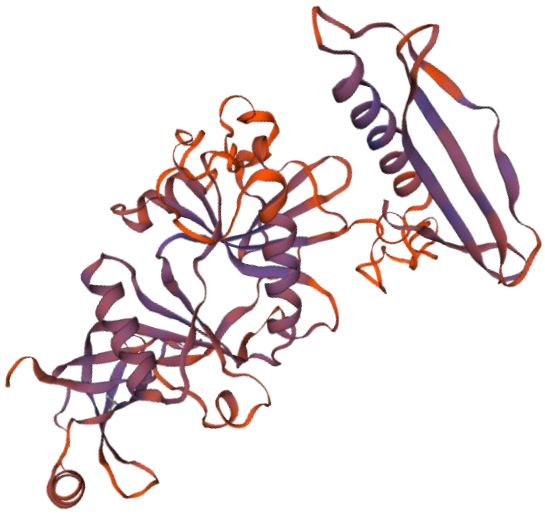
Characterisation of the activity of a novel DNA ligase enzyme encoded by a crAss-like bacteriophage from the human gut

Author: Polina Foteva
Supervisors: Stuart MacNeill and Carolin Kosiol



Contents

Abstract

In all forms of life, essential enzymes called DNA ligases play a key role in joining strands of DNA whenever new DNA is being made or when damaged DNA is being repaired. There are two classes of DNA ligases, which use as a cofactor either of two small molecules ATP or NAD. Both ATP-dependent and NAD-dependent DNA ligases are widely used in recombinant DNA technologies. That is why there are a variety of commercially available ligases with different properties and preferential uses in specific biotech processes.

Highly divergent organisms occupying diverse ecological niches are a rich source of novel enzyme activities for the biotech industry. In this regard, a recently discovered family of viruses that infect bacteria in human gut, called crAss-like phages are of an increasing interest for scientists. Some members of the family appear to encode their own ATP-dependent DNA ligases, but these have not been studied before. The aim of this research project was to purify and biochemically characterise the first example of a crAss-phage DNA ligase enzyme, as a prelude to exploring its utility in biotech applications. Initially, conditions were tested for optimal expression of four distinct crAss-like phage DNA ligases. One of them, named OJ ligase, was then purified at large scale to homogeneity, alongside an engineered catalytically inactive version of the same enzyme (OJ_m). The purified OJ enzyme (but not the catalytically inactive version OJ_m) was shown to possess ATP-dependent ligase activity in a series of biochemical assays. This research provides a solid foundation for continued exploration of the biotech potential of the crAss-like phage DNA ligases.

Introduction

Enzymes are biomolecules which increase the rate of biological processes. This property to ‘catalyse’ diverse set of reactions makes them essential for all living organisms. Moreover, enzymes with novel properties are of an interest in biotechnology, due to their potential use for improving current procedures or developing new methods. One key source of diverse, yet unstudied enzymes are bacteriophages (or phages for short): viruses that infect bacteria. Phages often encode enzymes that belong to the ubiquitously found classes of enzymes, but which differ in particular properties, e.g. they function at extremely high temperature or in high salt concentrations.

The recently discovered family of crass-like phages are found to contain genes for several important enzymes in their genome. These phages infect human gut bacteria from the phylum Bacteroidete and are found in roughly half of the human population, which makes them the most abundant viruses in human gut (Yutin et al. 2018). Some members of the crAss-like family can synthesise their own DNA ligase, an enzyme that catalyses the joining of newly made or newly repaired DNA. Indeed, there is a notable diversity of the ligase sequences across different phages, especially within zetacrassvirinae genera. As DNA ligases can join pieces of DNA together, they are essential for some cellular processes and a valuable tool in the recombinant DNA technologies.

Two factors have hindered scientists from studying crAssphages extensively. These are the recent discovery of the family in 2014 (Dutilh *et al*. 2014) and their challenging isolation. Unlike their host bacteria, which synthesise NAD-dependent DNA ligases, the ones encoded by crAssphages are ATP-dependent. Here we describe the recombinant expression of ATP-dependent DNA ligases encoded by four species of crAssphages. This is followed by purification of a representative of this group of enzymes (named OJ) and investigation of its properties. The OJ ligase was selected after small-scale purification and a mutagenesis, where it produced most promising results. The experimental work was supplemented by a phylogenetic reconstruction of the distribution of ATP-dependent DNA ligases across crAss-like phages.

Methods

Studying the properties of an individual enzyme ideally requires that the enzyme is first purified to homogeneity. This is most conveniently done by expressing the protein in bacteria (typically *E. coli*) and using standard protein purification technologies to separate the desired protein from the proteins that belong to the bacteria. To do this, the gene encoding the protein of interest is inserted into the bacteria in the form of a plasmid (a small circular DNA capable to being stably maintained inside the bacterial cells). This creates genetically modified bacteria, that can synthesise the desired protein in sufficient amount. The bacterial cells are then lysed, releasing a mixture of cellular proteins. The recombinant protein is purified from this solution. Depending on the function of the protein and the aim of the research, different experiments can be conducted to analyse its properties. The detailed steps of this strategy are described in the next few sections.

Molecular Cloning:

Initially, four distinct crAss-like DNA ligases were investigated. The sequences of these four presumptive ATP-dependent DNA ligases, named OH, OJ, OL and LM, were selected from a database and the genes were synthesised commercially as double-stranded DNA molecules (gBlocks). Before inserting the ligase genes into bacteria, they needed to be ligated into a plasmid to carry them. The four gBlocks and the plasmid vector type pEHISTEV were cut at specific positions (‘digested’) by adding restriction enzymes to the solutions. The resulting molecules had cohesive ends, meaning that their overhangs aligned and they could be easily joined together (fig.1).

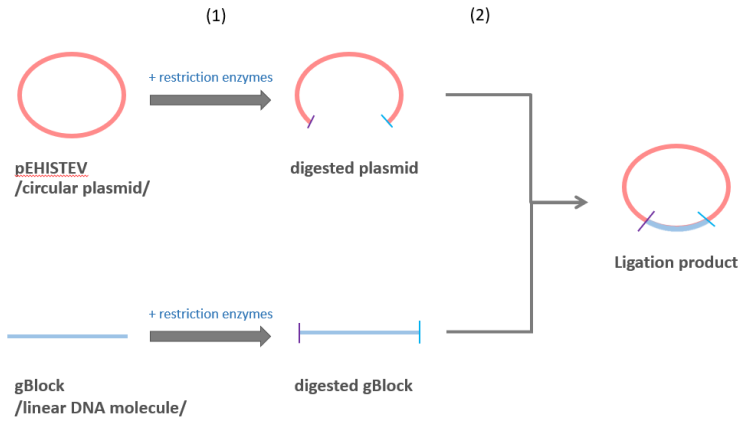


Figure 1. Schematic representation of DNA digestion by restriction enzymes (1) and ligation to a plasmid vector (2).

The obtained plasmids were transformed into special strain of *E. coli* (DH5α) and the cells were grown overnight to form separate colonies (fig.2). In addition to the transformed cells with each of the four ligation products (OH, OL, OJ and LM), there was a control sample, containing only the plasmid digest.

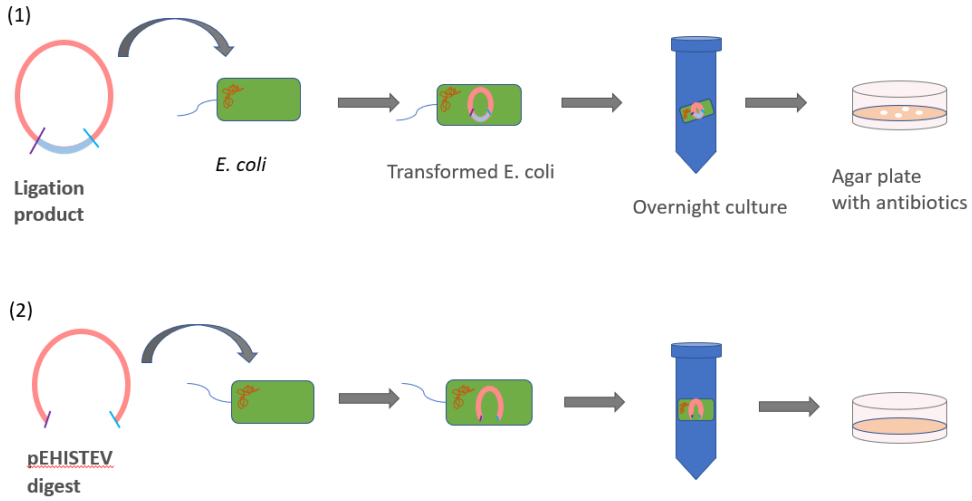


Figure 2. Transformation of E.coli DH5α with the product of ligation (line 1) and with the vector digest (line 2). Colonies are shown as white dots on the agar plate.

The plasmid vector contained a kanamycin resistance gene, which made the bacteria resistant to the antibiotic kanamycin. Thus, cells which had not been transformed successfully were not able to survive, when plated on kanamycin-containing agar medium. Similarly, as the negative control for the ligation were cells, transformed with the digested plasmid and no gBlock digest added, they could not express the kanamycin resistance genes.

Two colonies of each sample were selected, their plasmids were isolated and sent for DNA sequencing. This step was essential in order to check that no mutations had been inadvertently introduced during the procedure. It should be noted that a single colony was treated as an independent entity, because DNA sequences might differ between

colonies from the same plate. The plasmids that had the desired DNA sequence were transformed into another strain of *E. coli* (Rosetta 2 DE3) and a similar protocol as described above was followed.

Protein Expression:

The successfully transformed colony was taken from the plate, suspended in a liquid medium and incubated at 37°C overnight. The next day it was resuspended in larger amount of medium and incubated. The bacterial growth could be monitored by measuring the optical density at 600 nm (OD_{600}). As the number of cells increased, the solution became denser and less light could pass through it. Incubation continued until reaching OD_{600} between 0.6-0.8.

Genes contain the instructions for making proteins in a process called gene expression. This process is regulated by several mechanisms. In this case, the expression of the gene of interest was induced by adding a chemical (IPTG), to the samples. Following a second period of incubation at 37°C for 4 hours, the cells were expected to contain the recombinant crAssphage DNA ligase in their cytoplasm. Samples were taken before and after induction and analysed by SDS-PAGE gel. The cells were separated from the medium by centrifugation and frozen. Ultrasound was used to lyse the cells. The obtained solution was centrifuged and the supernatant was collected. Providing that the protein of interest was soluble, it would be in the supernatant.

Protein Purification:

The solution we had collected contained a mixture of cellular proteins, therefore we had to purify the one we wanted. This was possible because the protein of interest was synthesised with six additional histidine amino acids at the N-terminus. Addition of this ‘tag’ enabled binding to nickel cations, so the protein could be isolated from the solution by immobilised metal affinity column chromatography (IMAC). The protein mixture was run through a column, containing small nickel resin, so that the recombinant enzyme could bind it. At the end, a chemical (imidazole), which also binds the resin, was used in high concentration to elute the desired protein. The purity was additionally increased by size exclusion chromatography, where the protein solution was diluted in a buffer and run through a column. This resulted in a number of fractions, containing molecules of similar size (fig.3). Only the fractions, containing the desired protein were collected, mixed with glycerol (1:1 ratio) and the solution was stored in the freezer.

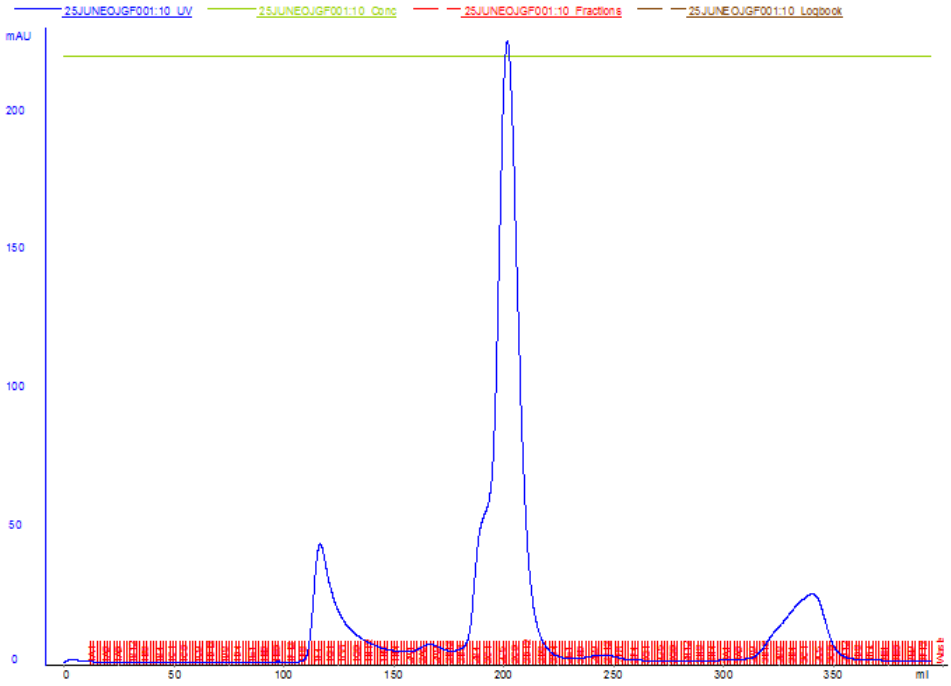


Figure 3. Size exclusion gel chromatography. The x-axis shows the elution volume (mL) and y-axis shows the absorbance (mAU). Peaks correspond to the fractions containing proteins of similar size and the height of the peak is associated with the concentration of the protein. The highest peak is due to the OJ DNA ligase and the fractions, collected around 200 mL contained the pure enzyme.

Mutagenesis:

A way to check that any activity subsequently detected was due to the recombinant DNA ligase, not because of native bacterial proteins, contaminating the purified samples, was to synthesise and purify a mutant form of the enzyme. The mutant could be used as a second negative control during the activity assays, in addition to the sample with no enzyme added.

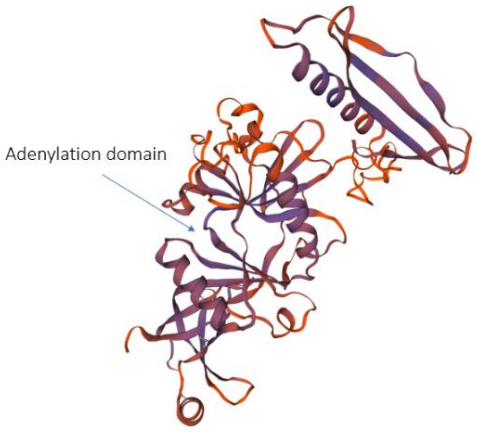


Figure 4. A hypothetical model of OJ DNA ligase. (Image generated in https://swissmodel.expasy.org/)

Previous studies have shown that a single lysine amino acid that is located at a key position in a highly conserved catalytic domain (fig.4 – the adenylation domain), is crucial for the activity of ATP-dependent DNA ligases (Doherty and Suh 2000; Martin and MacNeill 2002). Therefore, these enzymes can be easily inactivated by changing the genetic sequence of the plasmid, so that cells synthesise a mutant enzyme having alanine instead of lysine. For convenience, in the case of OJ this is shown as OJ K140A, meaning substitution of lysine at position 140 with alanine (fig.5).

Mutagenesis was carried out using a commercial kit, which required custom oligonucleotide primers (very short single-stranded DNA molecules) to introduce the desired mutation. The plasmid carrying the OJ gene was used as a template in a PCR

reaction with the custom primers. The template was then removed and the newly synthesised DNA was introduced into bacterial cells (*E. coli* strain NEB 5-α), following the manufacturer instructions. The same steps, as described above, were repeated with the successfully mutated plasmids, resulting in a purified mutated OJ_m DNA ligase.

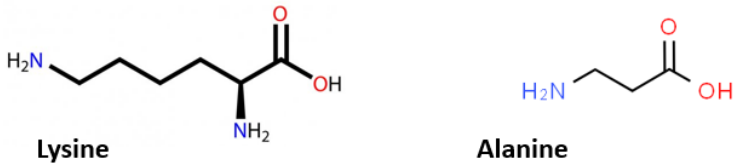


Figure 5. Chemical structure of the amino acid lysine and amino acid alanine.

Activity Assays:

The activity of the purified OJ DNA ligase was studied in a variety of activity assays. Assays 1 and 2 aimed to show that the purified protein functions as a DNA ligase enzyme by comparing its activity with a commercially available standard ATP-dependent DNA ligase encoded by a T4 phage (modified version of the protocols described by Muerhoff *et al.* 2004, and the activity assay for the T4 DNA ligase in the New England Biolabs website). The two assays differed in the type of DNA fragments used as substrates for the catalytic reaction. Assay 3 tested the dependence of OJ DNA ligase on ATP and the same substrate as in assay 1 was used.

Phylogenetic Reconstruction:

The protein sequences of ATP-dependent DNA ligases of crAssphages were collected from GenBank (Sayers *et al*., 2020), searching 'ATP-dependent DNA ligase' in the protein database and filtering for 'crAssphages', as well as from the supplementary data of publications (Yutin et al. 2018; Tisza and Buck 2021). Multiple sequence alignment was performed using 'MUSCLE' option in SeaView (Gouy *et al.* 2010). The best fit model was selected in IQ tree and the sequences that failed the composition chi2 test for divergence were removed. The best-fit model of the crAssphages DNA ligase tree was used to generate the final phylogenetic tree in high performance computing software. The tree was compared to the polymerase A (Pol A) and terminase L (Ter L) phylogenetic trees found in literature (Yutin et al. 2018).

Results

Protein Expression, Purification and Mutagenesis:

Following construction of plasmids to express tagged recombinant DNA ligases, the solubility of the proteins at various temperatures was tested.

Samples OH and OJ had higher intensity of the bands of interest (fig.6-A: bands indicated by a red arrow), suggesting a higher level of expression and/or solubility of the recombinant OH and OJ proteins. All of them were soluble. OL and LM appeared less soluble, but when protein induction was performed at lower temperature, their solubility increased. For subsequent studies, expression of the OJ ligase was scaled-up and a catalytically inactive mutant of the OJ ligase (called OJ_m) was successfully expressed and purified in parallel. Figure 4-B shows the purified OJ and OJ_m proteins, the first crAss-phage DNA ligases to be purified.

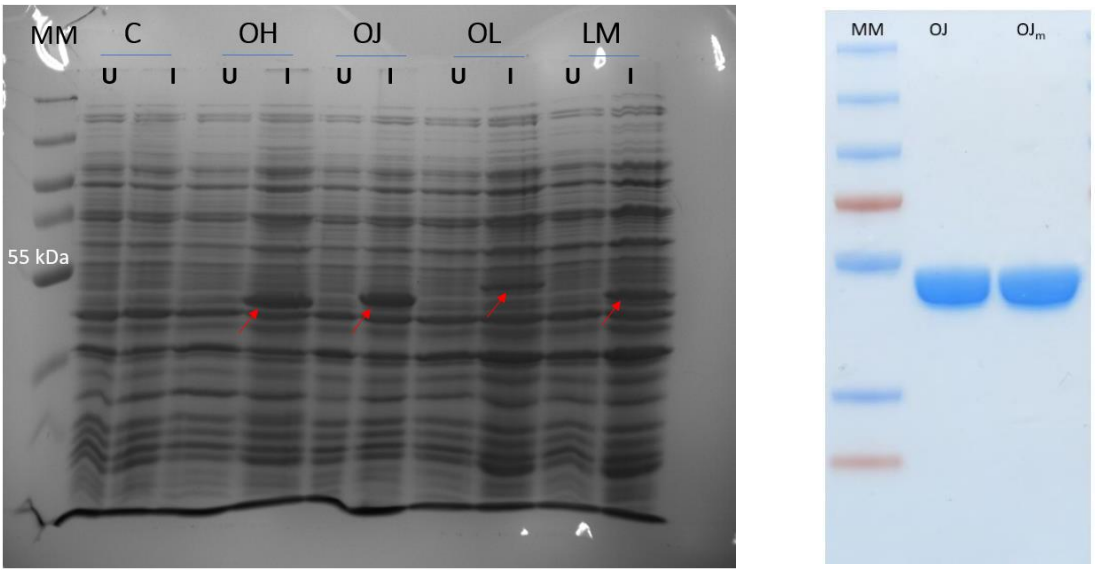
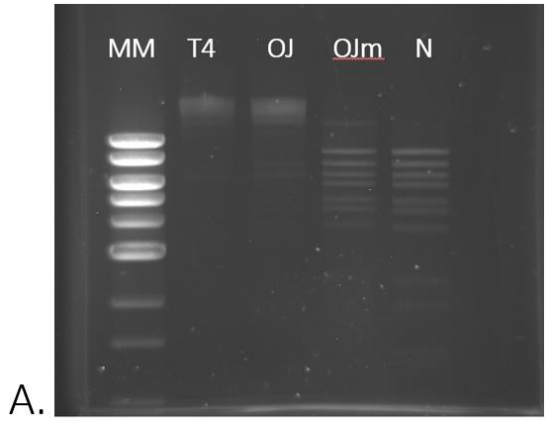


Figure 6. A – SDS-PAGE gel of the four enzyme solutions prior and post protein induction with IPTG, as well as a control sample. Arrows pointing at the additional bands, corresponding to the recombinant proteins. B - SDS-Page gel showing the purified OJ and the mutated OJ (K140A) DNA ligases. Line MM is a molecular weight marker, showing that the weight of the proteins is a little below 55 kDa.

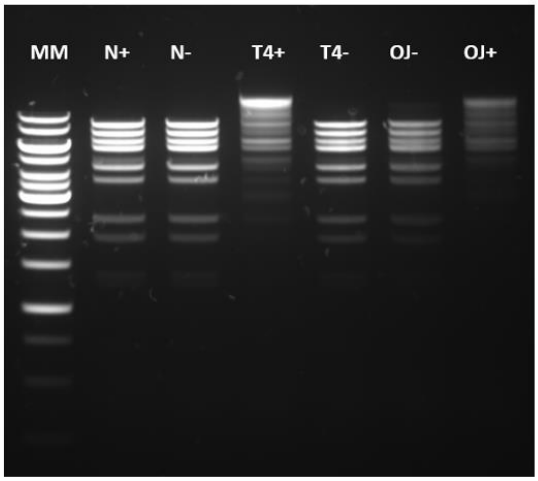
Activity Assays:

The observed activity of the OJ DNA ligase was comparable to the well studied and widely commercially available T4 DNA ligase. The OJ DNA ligase was highly active and ligated all the DNA fragments produced by prior digestion of phage λ DNA with the restriction enzyme BstEII. This was shown by the presence of one single DNA fragment of a higher molecular weight in lanes T4 and OJ (fig.7-A). The OJ DNA ligase was also shown to be active in the presence of ATP and to lose its activity, when no ATP was added to the buffer (fig.7-B.). The T4 and OJ DNA ligases produced a mixture of DNA fragments of different length, when phage λ DNA, digested with the restriction enzyme HindIII and HaeIII, was used as a substrate (fig.7-C.).

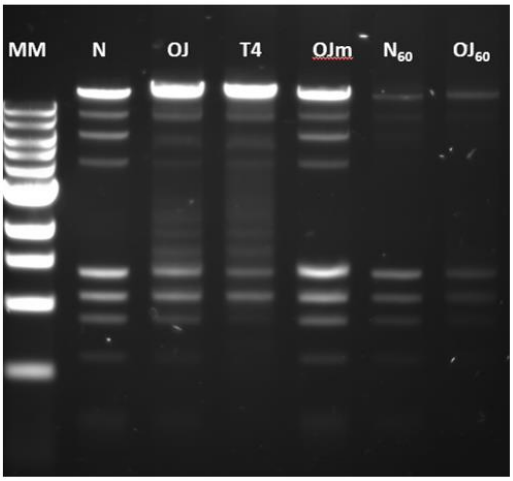
Overall, the results of these assays show that the OJ ligase is a highly active ATP-dependent DNA ligase that is capable of ligating both cohesive and blunt DNA ends. As expected, no activity was observed in the negative controls, where either the mutant OJ_m DNA ligase was used, or no DNA ligase was added.



A.



B.



C.

Figure 7. Catalytic activity assays testing the ligation activity of OJ DNA ligases. ***A.*** *Assay 1 - using λ DNA-BstEII Digest.* ***B.*** *Assay 3 - testing the dependence of the activity on ATP.* ***C.*** *Activity Assay 2, using λ DNA HindIII and HaeIII digests.* ***Legend:*** *MM – molecular weight marker (1kb DNA ladder); T4 – T4 DNA ligase; OJ – OJ DNA ligase; OJm – mutant (K140A) OJ DNA ligase; N – no enzyme added; '+' – ATP was added to the buffer; '-' – no ATP added;* N_{60} *and* OJ_{60} *– lower DNA concentration and shorter time for the reaction (60 min. instead of 120 min). DNA fragments were visualised by UV light.*

Phylogenetic Reconstruction

Overall, 103 sequences were aligned, but 4 of them were removed after running model selection in IQ-Tree (Minh et al., 2020). The best-fit model for the 98 sequences left was VT+F+R5. The phylogenetic tree of crAssphage DNA ligases, inferred on the High Performance Computing cluster Kennedy of the University of St Andrews, showed higher level of similarity between crAss-like bacteriophages from the same genera (fig.8).

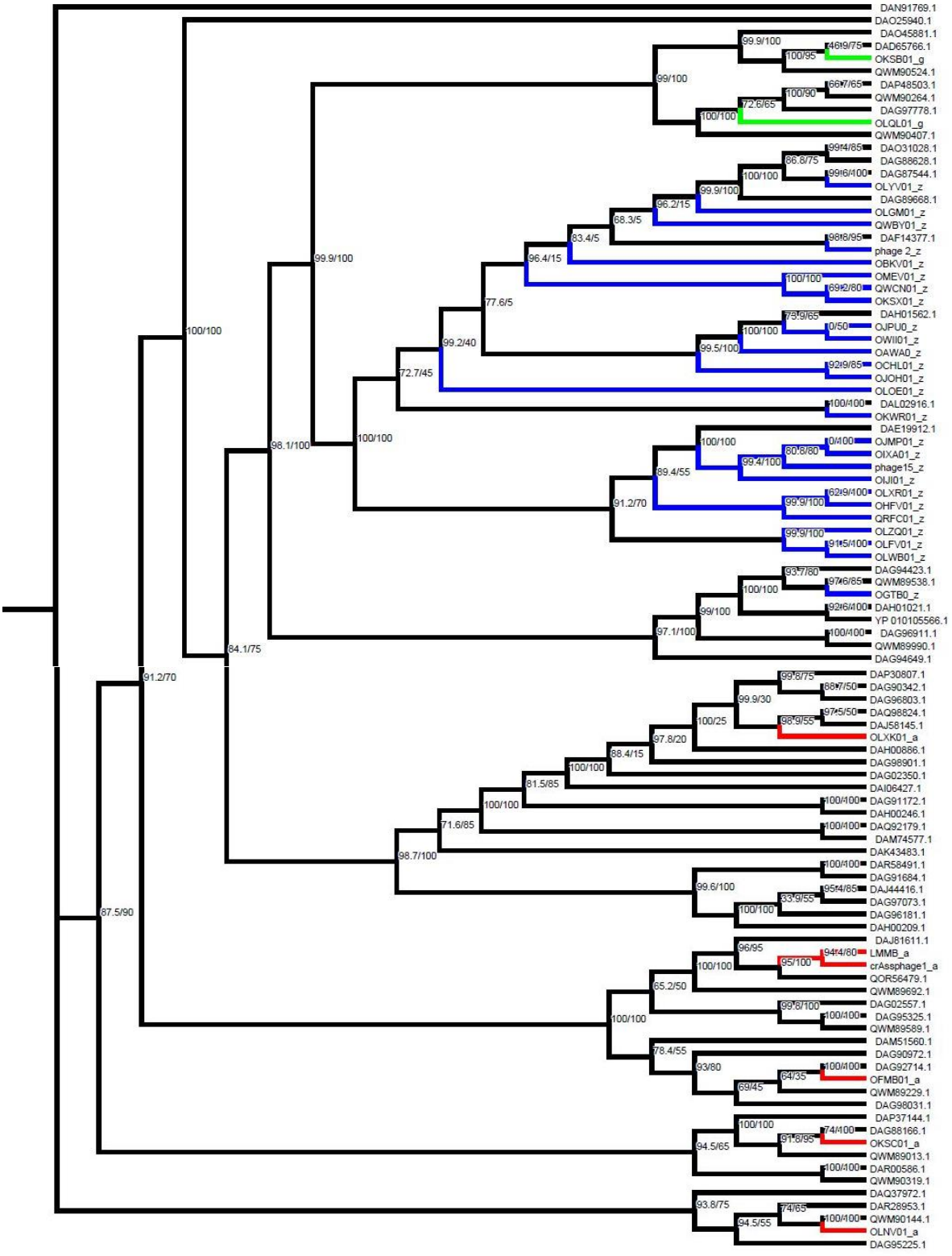


*Figure 8. Phylogenetic tree of DNA ligase protein sequences distribution across different representatives of the crAss-like phages family. Genera are indicated with a colour and the first letter of the genera name, following the name of the phage: green – gamma and *_g; red – alpha and *_a; blue – zeta and *_z. Numbers on the branches are boostrap values.*

Conclusion

This study shows that the most abundant viruses in human gut – crAss-like bacteriophages can encode soluble and functional ATP-dependent DNA ligase enzymes. Four divergent crAssphage DNA ligases (named OH, OJ, OL and LM) were successfully expressed in recombinant form in *E. coli* strain Rosetta 2 DE3 and purified. OJ DNA ligase was studied further and shown to catalyse the ligation of cohesive ends in the presence of ATP. No activity was detected when using ATP-depleted buffer. Similarly, the mutant form of OJ – OJ_m DNA ligase was catalytically dead and inactive.

The project offers a solid base for future development. The activity of OJ DNA ligase can be tested in other assays under various conditions with the potential of discovering novel properties of the enzyme. Provided that there is sufficient time and materials, there is no limit of the types of assays that can be performed. This might be followed by biochemical investigation of the OJ DNA ligase structure, which would give clues for its evolutionary origin and mechanism of catalysis. Considering the fact that OJ DNA ligase is the first crAss-like phage DNA ligase to be purified, and one of the first crAss-like phage protein to be expressed in a lab, the results of the research have the chance to be published in a scientific journal (Drobysheva *et al.* 2020).

The biotech industry is a dynamically developing field, where different classes of enzymes are used. The results of the current study imply that that crAss-like phages are a source of novel functional enzymes. As the four crAssphage DNA ligases, have divergent DNA (and hence protein) sequences, they are likely to possess different properties. Therefore, it would be not only interesting, but also reasonable to study all of them. Following optimisation of the activity assays, the catalytic activity of OH, OL and LM DNA ligases can be readily tested. A more extensive investigation of the enzymes, encoded by this family of bacteriophages, might reflect in their direct application in commonly used biotechnologies, improving the cost and efficiency of science. This advantageous prospect coincides with my leadership ambitions to make science equally accessible to every researcher. In a leadership aspect the project can be developed in two directions – contribution with other scientist who work on crAssphages, as well as, implementation of new strategies to tackle barriers that scientist from less developed countries face.

This research would not be possible without the generous support of Lord Laidlaw and the Laidlaw Foundation. A thousand thanks for providing everything I needed and encouraging me to follow my ambitions. I would also like to express my greatest gratitude to my supervisors Dr Stuart MacNeill and Dr Carolin Kosiol. I truly appreciate all your patience, support and positivity.

References:

Doherty, A., Suh, S., 2000. Structural and mechanistic conservation in DNA ligases. N*ucleic Acids Res*. 28(21):4051–4058.

Drobysheva, A., Panafidina, S., Kolesnik, M., Klimuk, E., Minakhin, L., Yakunina, M., Borukhov, S., Nilsson, E., Holmfeldt, K., Yutin, N. 2020. Structure and function of virion RNA polymerase of a crAss-like phage. Nat 2020 5897841. 589(7841):306–309.

Dutilh, B., Cassman, N., McNair, K., Sanchez, S., Silva, G., Boling, L., Barr, J., Speth, D., Seguritan, V., Aziz, R., 2014. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun*. 5(1):1–11.

Gouy M., Guindon S., Gascuel O., 2010. SeaView version 4 : a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27(2):221-224.

Martin, I. and MacNeill S., 2002. ATP-dependent DNA ligases. *Genome Biol.* 3(4):reviews3005.1.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D. , Woodhams M.D., Von Haeseler A., Lanfear R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37 (5): 1530-1534.

Muerhoff, A., Dawson, G. and Desai, S., 2004. A non-isotopic method for the determination of activity of the thermostable NAD-dependent DNA ligase from *Thermus thermophilus* HB8. *J Virol Methods.* 119(2):171–176.

Sayers E.W., Cavanaugh M., Clark K., Ostell J., Pruitt K.D., Karsch-Mizrachi I.. GenBank. Nucleic Acids Res. 2020; 48:D84–D86

Tisza, M. and Buck, C., 2021. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proc Natl Acad Sci.* 118(23): e2023202118

Yutin, N., Makarova, K., Gussow, A., Krupovic, M., Segall, A., Edwards, R., Koonin, E., 2018. Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. Nat Microbiol. 3:38–46.