To Reduce or not to Reduce, that is the question: Ancestral Sequence Reconstruction of Class II Ribonucleotide Reductases

Matthew Urban Watts

An Honours Thesis submitted in partial fulfilment of the requirements for the Bachelor of Science (Hons) in Biology, the University of Auckland, 2023. This Honours Thesis is for examination purposes only and is confidential to the examination process.

Contents

I. Abstract:

Ribonucleotide Reductases (RNRs) are an ancient family of proteins that reduce ribonucleotides to deoxyribonucleotides, and likely date back to the RNA to DNA transition in the early evolution of genomes. There are three classes of RNR, which differ in a number of respects, including their oxygen dependence. Class I enzymes are oxygen dependent, class III enzymes are strictly anaerobic, and class II enzymes are oxygen independent. It has been speculated that class II or III may be most similar to the ancestral RNR (dubbed the 'ur-reductase') in terms of function, but their low sequence similarity and divergent cofactor dependencies make ancestral reconstruction a challenge. As a first step towards reconstruction of the ur-reductase, we have started work on ancestral reconstruction of the class II RNRs. To this end, I created an alignment of Class II RNR sequences and built a phylogeny, using class I RNRs as a possible outgroup. We next performed Ancestral Sequence Reconstruction (ASR) to reconstruct probable ancestral sequences at the internal nodes of our tree. Our lab recently generated a line of *E. coli* that completely lacks ribonucleotide reduction meaning we can use this line to test ancestral sequences for function via a Phenotypic Rescue Experiment (PRE). As our knockout line cannot survive without deoxyribonucleoside supplementation, we can assay for ancestral RNR function as only functional RNR sequences should rescue deoxyribonucleoside dependency in the knockout.

II. Acknowledgments:

A short series of thank yous: Thank you Ant for taking me on for this project and for your kind mentorship during its duration. Thank you to Soon for dealing with my lab related questions. Thank you to Danni for advice for my transformations. Thank you to the rest of the Poole lab for much needed feedback on various presentations and projects throughout the semester. And thank you to my flatmates for dealing with spread of my laptop, notes and other possessions across the dining room table.

1. Introduction:

1.1 Ribonucleotide Reduction and the Transition from the RNA to DNA World

Ribonucleotide Reduction is an important reaction for the synthesis of deoxyribonucleotides (dNTPs), the building blocks of DNA, from the ribonucleotides which make up RNA (Poole, Logan, & Sjöberg, 2002). Ribonucleotide Reductases (RNRs) are the only known enzymes which catalyse Ribonucleotide reduction for the *de novo* synthesis of dNTPs (Reichard, 1997). Because of this central role in DNA synthesis RNRs are effectively ubiquitous across all domains of life, with only five intracellular parasites known to lack RNRs (Lundin, Torrents, Poole, & Sjöberg, 2009). The RNA world hypothesis is the leading view for the origin of life (Saito, 2022). In this scenario RNA initially serves as both the biocatalytic and coding components of early cells. This is supported by the wide range of catalytic processes extant ribozymes are capable of, as well the presence of ribozymes in fundamental structures, such as the ribosome (Bartel & Unrau, 1999; Doudna & Cech, 2002). The initial RNA world is thought to have been replaced by a Ribonucleoprotein (RNP) world, where RNAs served as the coding molecule while still maintaining a catalytic role in combination with protein scaffolds (Poole, Jeffares, & Penny, 1998). This is the point at which a protoRNR is hypothesised to have originated, as shown in Figure 1, providing the deoxyribonucleotides for a DNA genome. We can place the origins of RNRs here as proteins are required for the arrangement of the redox-active metals and cofactors required for RR, which ribozymes are unlikely to have performed as an RNA scaffold is too unstable for the utilisation of reactive cofactors such as strong redox metals (Poole, Penny, & Sjöberg, 2000). The transition from an RNA to a DNA genome is evolutionarily important as DNA genomes are capable of accurately storing much more genetic information than an RNA genome (Lundin, Berggren, Logan, & Sjöberg, 2015). This transition likely occurred via several intermediate steps, as the maximum hypothesised RNA genome was likely not large enough to encode a complex protein such as an RNR required for DNA synthesis (Poole, Penny, & Sjöberg, 2000). It is important to note that RR was not necessarily the only chemical pathway which allowed the production of RNRs present in the RNP world, with the Reverse Deoxyriboaldolase reaction hypothesised to be a potential alternative for early life (Poole A. M., On alternative biological scenarios for the evolutionary transitions to DNA and biological protein synthesis, 2011). However, the dominance of extant RNRs

indicates that this alternative pathway was outcompeted. Therefore, due to the ubiquity of RNRs, investigating their evolutionary history is important for an investigation of DNA utilisation in early life.

Figure 1: Shows the transition from the RNA world to the RNP world and then to a modern world of cell function. The period of time during which an ancestral RNR is hypothesised to have originated is also displayed, occurring between the RNP and modern world. Adapted from Lundin et al. (2015), Created with BioRender.com.

1.2 Ribonucleotide Reductases

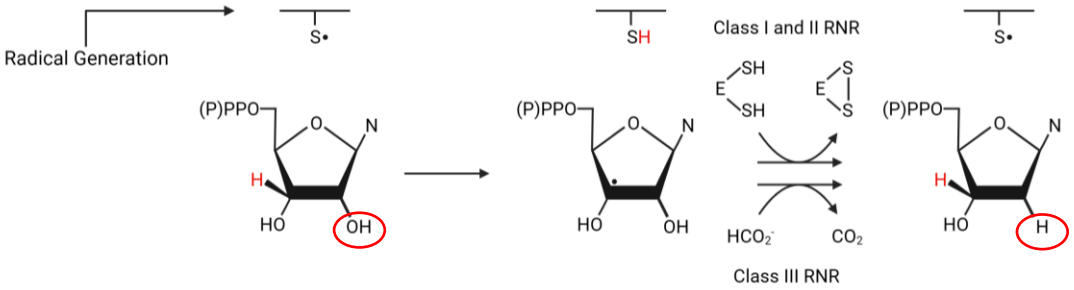

Figure 2: The general reaction catalysed by RNRs, with the key reduction of a hydroxide group to a singular Hydrogen indicated in red circle. Hydrogen atom (red) removal and subsequent rearrangements result in the loss of the 2′ hydroxyl group in the form of water. In class I and II RNRs, reduction is carried out via the oxidation of two thiol residues to a disulfide. In contrast, class III RNR obtains reduction is carried out via the oxidation of formate. Adapted from Sintchack et al. (2002). Created with BioRender.com.

Ribonucleotide Reductases are the enzymes which carry out ribonucleotide reduction. There are three classes of RNR: I, II and III. They are distinguished by their structure and the diverse cofactors they use to generate the initial radical to carry out reduction, (Lundin, Berggren, Logan, & Sjöberg, 2015). Class I RNRs use a di-iron-tyrosyl radical, Class II use coenzyme B12, also known adenosylcobalamin (AdoCbl), while Class IIIs make use of an Iron-sulphur cluster and S-adenosylmethionine to produce a radical (Torrents, Aloy, Gibert, & Rodríguez-Trelles, 2002). While the cofactors used to generate the initial radical may differ, the overall reaction they catalyse is shared. Additionally, they all share a conserved cysteine on a finger loop in the centre of a β/α Barrel (Eklund, Uhlin, Färnegårdh, Logan, & Nordlund, 2001). However, the reducing equivalents differ between the classes, as shown in figure 2, (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002; Wei, et al., 2014).In terms of structure Class I and III appear as two homodimeric proteins, one large α subunit and one smaller β subunit. In class I these are arranged as a heterotetramer (Torrents, Aloy, Gibert, & Rodríguez-Trelles, 2002). In Class III the larger catalytic α subunit binds to the smaller β subunit. Class II has both a monomeric and a homodimeric form, as shown in figures 3 and 4 respectively. The monomeric form appears to be the result of a 130-amino acid insertion of a dimeric interface into a monomer, as highlighted in figure 3 (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002). The RNRs have different

distributions across the domains of life: Class I RNRs are found in eukaryotes, bacteria, bacteriophage and viruses. Class II RNRs are present in bacteria, bacteriophage, eukaryotes and archaea while Class III RNRs are found in bacteria and bacteriophage (Lundin, Torrents, Poole, & Sjöberg, 2009). Additionally, different classes of RNRs can be found in the same organism (Crona, et al., 2013). Another distinguishing feature of the different classes of RNR is their sensitivity to oxygen. Class III is anaerobic, Class II operates independently of oxygenation and Class I is aerobic (Poole, Logan, & Sjöberg, 2002).

Figure 3: Structure of the monomeric class II RNR from L. leichmannii. Helices A and B are yellow and the dimer-mimicking monomeric insertion is shown in pink. Retrieved from *The Origin and Evolution of Ribonucleotide Reduction*, by Lundin et al., 2015, retrieved 2023, Jun. 4, from https://www.mdpi.com/2075-1729/5/1/604

1.3 The Evolutionary Relationship of the RNRs

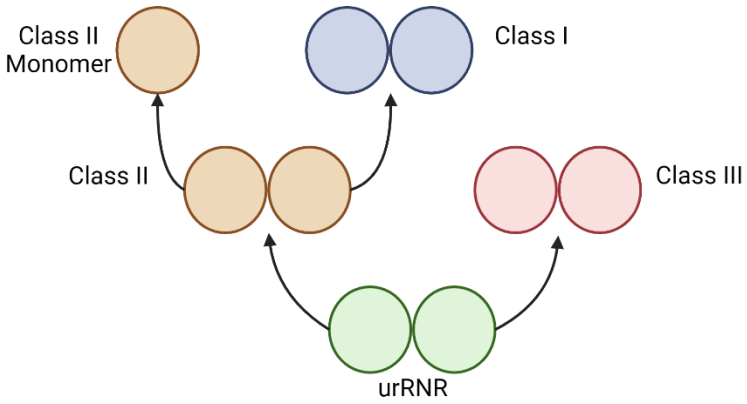

Figure 4: Model for the origin of the different classes of RNR from one another and placing a common ancestor, the urRNR. In this model the urRNR could have been similar to a dimeric class II RNR or a Class III RNR. Additionally, class I and the monomeric class II are hypothesised to have originated from dimeric class II RNRs. Adapted from Lundin et al (2015). Created with BioRender.com.

Based on the conservation of the RNRs reaction mechanism based on homologous structure it is widely agreed that RNRs trace to a single ancestor, often referred to as the urRNR (Lundin, Berggren, Logan, & Sjöberg, 2015; Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002). This is different to the protoRNR referenced in Figure 1, which is the first protein which carried out RR. The protoRNR is likely to have been a generalist reduction-based enzyme, which means it could have catalysed reactions other than just RR and therefore could be ancestral to enzymes other than the RNRs, such as proteins which also make use of a β/α Barrel (Stubbe, 2000; Lundin, Berggren, Logan, & Sjöberg, 2015; Eklund, Uhlin, Färnegårdh, Logan, & Nordlund, 2001). Therefore, when discussing the common ancestor of RNRs it should be referred to as the urRNR. The diversity of extant primary sequences of RNRs make claims about the urRNR difficult, because we are making inferences about an ancestor based only on what exists today (Burnim, Spence, Xu, Jackson, & N, 2022; Lundin, Berggren, Logan, & Sjöberg, 2015). This is an issue as some of the characters of modern RNRs could be derived, for example monomeric class IIs are confined only to specific habitats, which could mean there is some unknown selective pressure giving this distribution (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002; Poole, Logan, & Sjöberg, 2002). Additionally, the distribution of the different classes of RNR suggests

there has been HGT of sequences, which reduces the clarity of evolutionary relationships (Lundin, Gribaldo, Torrents, Sjöberg, & Poole, 2010). The arguments about what the urRNR would have looked like is therefore still unresolved. The different sensitivity of RNRs to oxygen is informative for determining the evolutionary history of the RNRs as we know the conditions of the early earth were likely anoxic, indicating that the urRNR was likely not a Class I RNR as Class I are aerobic (Lundin, Berggren, Logan, & Sjöberg, 2015). Class II is argued to be ancestral based on its ubiquitous distribution across all three domains, independence from oxygen and prebiotic or RNA world origin of its AdoCbl cofactor (Poole et al., 2002). Class III is proposed as being ancestral as it is anaerobic and use of an ancient motif in the form of an iron-sulphur cluster (Reichard, 1997). In their study Burnim et al (2022) carried out phylogenetic structural analysis of RNRs, which supported a common ancestry for Class III and Classes I and II, as well as introducing a novel class 0 sharing a common ancestor with Classes I and II, as shown in figure 5. This differs from the hypothesis in Lundin et al (2015) as Class III is presented as the most basal group and there is no explicit origin of class Is from class IIs as in Lundin et al (2015). The fact that new hypotheses are still being drawn from phylogenetic reconstructions of these proteins indicates that there is more to learn about their evolutionary relationships.

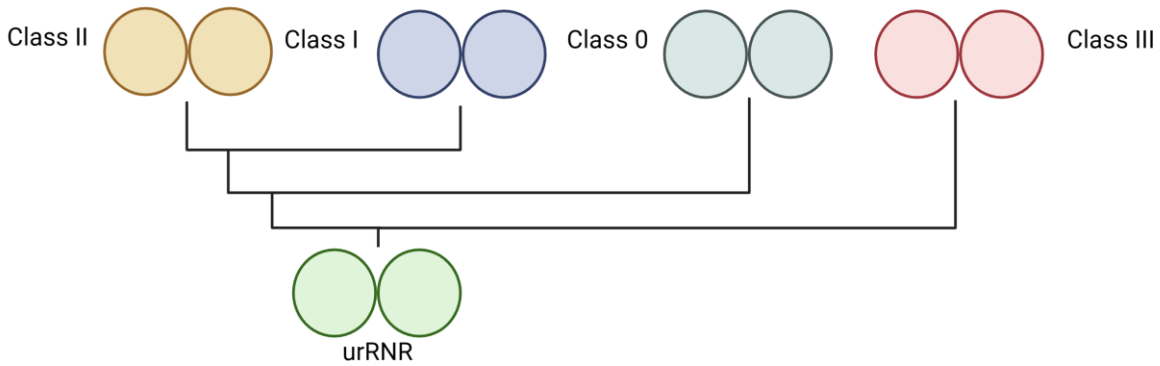

Figure 5: Evolutionary relationships of RNRs based on phylogenetic reconstruction by Burnim et al (2022). This phylogeny has Class III as the most basal group, as well as including a class 0, which is a claim specific to this paper. The relationship between class I and II is presented as descent from a common ancestor. .Adapted from Burnim et al (2022) Created with BioRender.com.

1.4 Ancestral Sequence Reconstruction

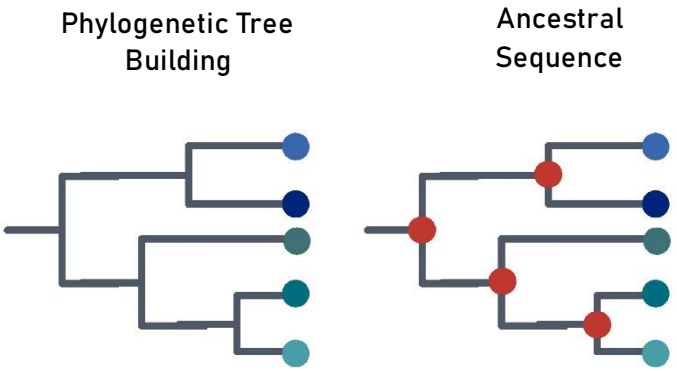

Figure 6: A visual comparison of Phylogenetic tree building and Ancestral Sequence Reconstruction. For a phylogeny we work out how the leaf nodes (in various shades of blue) are related to one another, but do not produce the sequences of the ancestors. For Ancestral Sequence Reconstruction we not only create a phylogeny, but we also recreate the sequences of the internal nodes (the red dots at the branching points), which represent the ancestors of the modern sequences. Created with BioRender.com

Ancestral Sequence Reconstruction (ASR) is a methodology based on compiling the modern sequences of proteins, aligning those proteins to one another, using the alignment to produce a phylogeny and then inferring the internal nodes of the phylogeny, which represent ancestral proteins, based on the modern proteins. A simplified version is shown in figure 6, highlighting the difference between phylogenetic tree building, where we are only interested in the relationships between the leaf nodes, and ASR, where we are also inferring internal nodes. Methodologies such as maximum parsimony, maximum likelihood or Bayesian methods can be used to infer ancestral sequences at nodes in the phylogenetic tree (Randall, 2016). Maximum parsimony selects the phylogeny which requires the fewest state changes to arrive at the modern day, effectively using Ockham's razor to select the outcome which requires the fewest changes (Poole A. M., 2021). Maximum likelihood and Bayesian methods are probabilistic, which means that multiple ancestral sequences are compared at each node and selected based on the probability of a particular sequence being the true ancestor (Spence, Kaczmarski, Saunders, & Jackson, 2021). For a Maximum likelihood methods, the sequence which has the highest likelihood of being the true ancestor based on an evolutionary model is selected. In Bayesian

methods the state which matches a posterior distribution of states at nodes in the model based on the a priori expectations the researchers had. The ability to produce ancestral proteins via the researcher's method of choice allows us to explore hypotheses concerning the functionality of those ancestral proteins, at what point important mutations occurred or other characteristics of ancestral proteins such as monomeric versus dimeric form. However, the limitations of this methodology is that it is highly dependent on the quality of the alignment and model used. Additionally deeper nodes in the phylogeny are less likely to be accurately reproduced compared to shallower nodes due to the accumulation of errors in the reconstruction, epistasis or contingencies during evolution or a change in the function a protein carries out compared to the ancestor. Epistasis occurs when one gene influences the evolution of another, which can affect its evolutionary trajectory (De Visser, Cooper, & Elena, 2011). Evolutionary contingencies occur where the outcomes of an evolutionary scenario could have been very different based on chance occurrences, such as the origin of citrate utilisation in the long term evolution experiment (Blount, Lenski, & Losos, 2018; Blount, Borland, & Lenski, 2008). If the function of an ancestral protein differs from that of its descendants, this can lead to modern sequences bearing little resemblance to the ancestor due to the different selective pressures being exerted (Clifton, et al., 2018). All of these factors can obscure the evolutionary signal, making it more difficult to work out ancestral sequences (Salverda, et al., 2011). Therefore, ASR works best for proteins where there is a large dataset of proteins, which are all under selective pressure and where function has been maintained across the phylogeny (Clifton, et al., 2018; Randall, 2016).

The function and reliability of different ASR methods was explicitly tested by Randall et al (2016) in a benchmarking test using a dataset of fluorescent proteins they had evolved. This lineage of fluorescent proteins was created via mutagenesis PCR. This allowed the researchers to record sequences and characteristics of the real ancestral proteins in their phylogeny, which allowed an explicit comparison between the ancestral nodes produced via ASR methods against a set of real ancestral sequences (Hillis et al, 1992, as cited in Randall et al, 2016; Randall et al, 2016). This experiment made use of Bayesian and Maximum Parsimony ASR models. As expected, deeper nodes in their ASR models accumulated more errors. However, It is

interesting to note that more complex Bayesian models didn't perform significantly better than Maximum Parsimony in terms of the number of incorrectly inferred amino acids. This suggests that ASR doesn't necessarily have to have a huge number of free parameters and be computationally heavy to produce meaningful results, so even a model with a few parameters can allow inferences (Hanson-Smith, Kolaczkowski, & Thornton, 2010; Randall, 2016). This is tempered by the fact that maximum parsimony performed significantly worse than the Bayesian methods at predicting phenotype. It is important to note that ASR is dependent on every step in the workflow, so alignment, tree building and selection of an appropriate model for the evolutionary scenario are very important for a successful analysis (Kapli, Yang, & Telford, 2020; Bromham, 2019).

1.5 Class II RNRs and ASR

Many of the traits of the Class II RNRs help ameliorate the weaknesses of ASR. For example, as previously mentioned, ASR can run into problems when proteins have epistasis or contingencies in their evolutionary history. However, in the case of RNRs their function is so fundamental that any loss of function would be highly deleterious. Additionally, the reaction mechanism for ribonucleotide reduction is highly specific and only occurs in the RNR protein family, which have an ancient origin and serve the same purpose across all the domains of life (Lundin, Berggren, Logan, & Sjöberg, 2015). Therefore, it is unlikely that evolutionary contingency or epistasis has played a significant role in the evolution of RNRs, while the monophyly of RNRs makes it unlikely that there has been a change of function between the urRNR and modern RNRs. This suggest that there should be strong evolutionary signal, supported by the presence of heavily conserved sites (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002). Therefore, for ASR of the RNRs we would expect all ancestral proteins to maintain the same reaction mechanism and function. This means that the success of ASR can be tested by establishing whether a functional RNR has been produced.

Class II RNRs are an interesting choice for an ASR experiment as they have a great deal of sequence and structural diversity, but still have a number of heavily conserved sites. As previously mentioned, they occur across all of the domains of life whereas classes I and III are more limited (Lundin, Berggren, Logan, & Sjöberg, 2015). Class IIs also have both a homodimeric and monomeric form, whereas the other classes

have only one form each (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002). This raises interesting questions concerning the timing and form of the split between the two different Class II forms. There is a limitation that many class II sequences contain inteins (Gogarten, Senejani, Zhaxybayeva, Olendzenski, & Hilario, 2002), which are a form of mobile genetic element which are spliced post translationally, as indicated in figure 7. While inteins can be phenotypically neutral (Friedrich, et al., 2007), they should be removed from proteins sequences for ASR as their presence reduces the quality of alignments, and thereby the quality of ASR. A positive for the ASR of Class IIs compared to the other classes of RNR is that there is no requirement to run ASR on multiple subunits which need to interact with one another to function as is the case if investigating the dimeric Classes I or III. This is important as ASR of interacting subunits could pose problems in terms of having to use concatenation or running parallel ASR models for proteins which interact but do have separate evolutionary histories (Kubatko & Degnan, 2007). Therefore, investigation of ancestral Class II RNRs via ASR provides a relatively convenient method to explore the evolution of RNRs and provide evidence as to which hypothesis is most likely for the relationship between Class I and II RNRs.

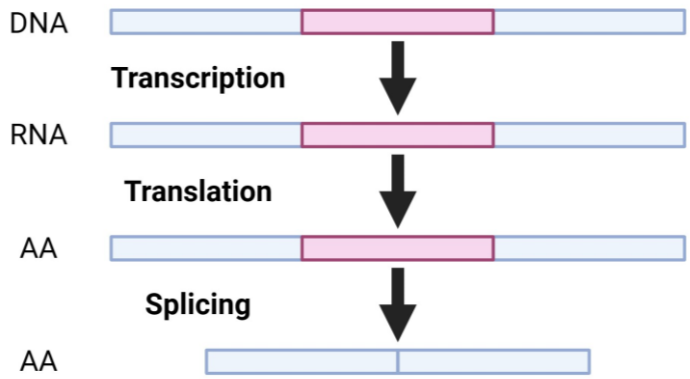

Figure 7: Removal of inteins, indicated by a red section, from functional proteins, indicated by the blue sections, after transcription, translation and post translational splicing. Noteworthy that Inteins are self-splicing, so no other proteins are required.

1.6 Phenotypic Rescue Experiments

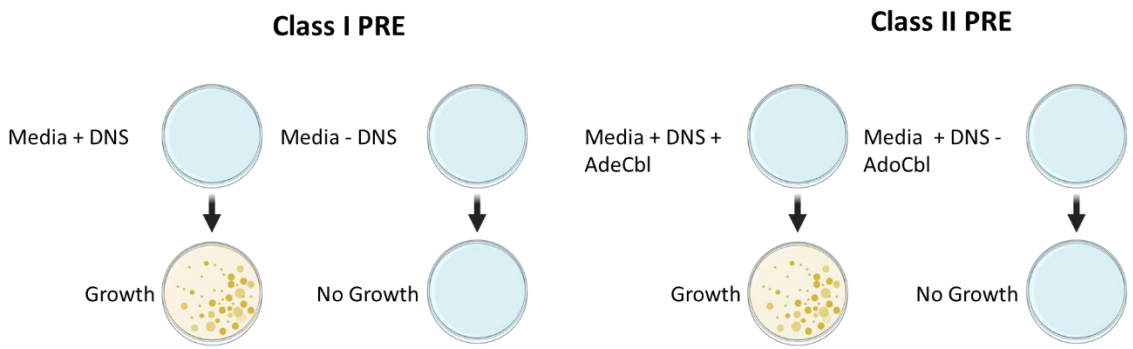

Figure 8: The basic premise of a phenotypic rescue experiment for class I and II RNRs. For class I proteins if a functional RNR is successfully transformed into an RNR knockout cell line it should allow the cell culture to grow without deoxyribonucleoside supplementation. For Class II RNRs this also requires the addition of the cofactor AdoCbl, without which it should not be able to grow. If the transformation is unsuccessful, the knockout cell line will not grow without DNTP supplementation. Created with BioRender.com

A Phenotype Rescue Experiment (PRE) offers a methodology to rapidly screen proteins for function. Phenotypic Rescue Experiment work by having a knockout cell line, which cannot survive without supplementation. If transforming a gene into that cell line allows the cell line to survive without supplementation, the gene which was introduced codes for a protein which carries out that function. Therefore, to test if a sequence is functional transform the gene of interest into the knockout cell line and place it on a growth medium without supplementation. If growth is observed this indicates the transformed cells have gained this function. This provides an explicit test of function, allowing us to determine whether ASR has been successful by testing whether or not a functional protein has been produced; a binary question, to reduce or not to reduce, as shown in figure 8. If the transformed cells are capable of a particular function, whereas other knockout cells are not capable of that function the gene which was transformed has conferred that function to the cell, as it is the only variable which has changed. This experimental methodology is possible because the Poole lab created a novel RNR knockout E. coli line (Arras, et al., 2023). A PRE is appropriate in this context as we are assuming that if our ASR has been successful

a functional protein will have been produced, supported by the fact that there are no alternative pathways which could perform RR (Arras, et al., 2023). Therefore, we want to initially screen for function, rather than analysing the protein structure using X-ray crystallography or alpha-fold (Jumper, et al., 2021). As we are expecting our ASR to produce a large number of nodes which will need to be screened for function, PRE is a suitable method as it offers a rapid and effective way to screen a large number of proteins.

1.7- Research Hypothesis and Aims

The hypothesis is that due to the conservation of important motifs and associated function across the evolutionary history of RNRs, ASR of Class II RNRs should produce ancestral sequences which function as RNRs at all internal nodes. To that end the aims of this study are to create a dataset of Class II RNRs, perform a multiple sequence alignment, carry out ASR based on that alignment, to transform the proteins produced at internal nodes by ASR into our RNR knockout and to then carry out a PRE using transformed cells to determine whether ASR has been successful in producing functional proteins across the phylogeny.

2. Analysis Methods:

2.1 Dataset Creation

The initial basis for creating a dataset of class II RNRs were known class II sequences from the literature which represented the diversity of class II sequences. These sequences included monomeric and dimeric class IIs as well as examples from across the domains of life where class II RNRs are present. The monomeric example was the amino acid sequence for the RNR *L. leichmannii* included in Sintchack et al. (2002). The specific sequence from this paper was easily accessible as the authors had shared their data to the protein databank under accession number 1L1L (Berman, et al., 2000). This sequence was entered into the Hmmer search tool, initially searching in the Swissprot dataset, with otherwise standard search parameters (Potter, et al., 2018). Swissprot was chosen as it is a curated dataset, although further curation will be required due to known errors with RNR annotation in large databases (Lundin, Torrents, Poole, & Sjöberg, 2009). The same procedure was used to look at an interesting example of a class II sequence

reportedly present in a eukaryote, *D. discoideum*, by Crona et al (2013). They had also saved their proteins in a public database, dictybase (Fey, Dodson, Basu, & Chisholm, 2013), which is a database of *Dictyostelium* sequences. The amino acid sequence was retrieved and used to carry out a search in HMMER in the Swissprot dataset under standard search parameters. For a search of dimers *Pseudomonas aeruginosa* was used as this RNR was referenced in Crona, Hofer, Astorga-Wells, Sjöberg, & Tholander (2015). In this case the sequence used wasn't included in any of the paper, so the top entry from uniprot (The UniProt Consortium, 2023) was used. None of the sequences used thus far had been archaeal, so to ensure that there was sufficient coverage of archaeal sequences in the database an archaeal sequence from a previous search result, *P. furiosus*, was used to carry out a HMMER search, in swissprot under standard search parameters. As the total number of sequences in my dataset was still low, to supplement the number of sequences included in the dataset HMMER searches for each of the previously mentioned sequences were re-entered into the search tool using the Hmmer reference proteomes rather than swissprot as the search dataset, as the reference proteomes is a larger dataset.

The resulting sequences were then renamed to simplify the identification and subsequent visualization of the distribution of sequences. For example, L. leichmannii was recorded as M_B_Lac_lei_sp|Q59490|, indicating that it is a monomeric class II and bacterial in origin, as well as including the sequence identifier, Q59490, for retrieval from uniprot. If multiple non identical sequences occurred in the same species each sequence was given a number, for example M_B_Lac_del_1_sp|Q1G7W2|.

2.2 Dataset Curation

Dataset curation is the process of removing sequences which had been mistakenly included (Lundin, Torrents, Poole, & Sjöberg, 2009) and removing sections of sequences which weren't found in the final protein, such as the afore mentioned inteins. This was an iterative process, carried out after dataset creation, with alignments and ASR allowing the identification of errors and removal of non-class II sequences, thereby improving the quality of my dataset. Initially all proteins

returned by HMMER from swissprot were included, as the total number was low. Extreme outliers in terms of E-value, a measure of similarity between sequences, were excluded. This list was later refined based on the presence of motifs conserved in class II RNRs and the absence of class I motifs, such as the YY double tyrosine motif, which are located at Tyr730 and Tyr731 in E. coli NrdA (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002). Alignments were used to identify class I proteins mislabelled as class IIs, as well as to distinguish class II monomers and dimer from each based on the presence or absence of the monomeric insert which defines class II monomers (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002). As a large number of my sequences contained inteins, curation of inteins using Uniprot was carried out to identify which sequences carried these motifs and where in the sequence they were locatedA text editing tool was then used to remove sections of the protein identified as inteins. Additionally, as duplicates of the same sequence were included due to the overlap of different searches these duplicates had to be removed from the dataset. Partial sequences based on shotgun sequencing, for which a warning appeared in Uniprot, were also not included in the dataset. The final dataset and list of sequences from which inteins were removed are included in the appendices in Appendix 1.

2.3 Sequence Alignment

Sequence Alignments were initially carried out across three different software: Muscle (Madeira, et al., 2022), Mafft (Madeira, et al., 2022) and Cobalt (Papadopoulos & Agarwala, 2007). These alignments were then compared to each other using the TCS tool on the Tcoffee server, which rates multiple sequence alignments based on the quality of the alignment, measured by how different each amino acid at each position in the alignment is relative to the other protein sequences (Chang, Di Tommaso, Lefort, Gascuel, & Notredame, 2015). Further curation of sequences and the alignments produced from these sequences were also scored using TCS on the Tcoffee server as an indicator of quality.

2.4 ASR

Ancestral sequence reconstruction was carried out in iqtree2 (Minh, et al., 2020) using the highest scoring Muscle sequence alignment. Initially a model finder

(Kalyaanamoorthy, Minh, Wong, von Haeseler, & Jermiin, 2017) was run to identify which model had the highest likelihood for the alignment entered. Two models, WAG+F+I+R6 and WAG+F+R6, had significantly higher scores than other models for this alignment. The ML trees produced by these models were compared to one another using a tanglegram made using eMPRess software (Santichaivekin, et al., 2021) to identify if there were major differences in topology between the two. As the two models showed only one difference in topology ASR was then carried out using the highest scoring model, WAG+F+I+R6, in iqtree. The command line was as follows:

iqtree2 -s (alignment file) -m (model) -B 1000 -asr.

Where -s indicates the alignment file, -m is the model for treebuilding, -B is the number of ultrafast bootstrap (Hoang, Chernomor, von Haeseler, Minh, & Vinh, 2018) replicates, which in this case was set to 1000, and -asr, which indicates that an output file containing the sequences of internal nodes is produced. The sequences of internal nodes were extracted from the output file using a text editor.

2.5 Structural Analysis of ASR Products

The internal nodes produced during ASR were aligned using Muscle to check if key residues had been conserved. Internal nodes were then entered into the PHYRE server (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015) under standard conditions. PHYRE is a tool which identifies what proteins are most similar to the query based on structure. This was done to check if internal nodes were still being identified as class II RNRs.

3. Analysis Results:

3.1 Dataset Creation and Curation

Table 1: Number of monomers and dimer in final dataset

Character	Number of Sequences
Monomers	45
Dimers	27
Total	72

Table 2: Number of sequences from each domain in final dataset

Character	Number of Sequences
Bacteria	44
Archaea	6
Eukaryotes	19
Viruses	3
Total	72

Table 1 indicates that there is representation of both Monomers and Dimers in the class II dataset, with more monomers than dimers. Table 2 shows that while the majority of sequences in the dataset are bacterial, there are examples of class II RNRs representing each of the domains of life, as well as viral sequences. In the dataset there are examples of both monomers and dimers in archaea and bacteria, while eukaryotes and viruses RNRs were only monomeric. The full list of sequences is in Appendix 1.

The initial dataset for Class II RNRs contained both monomers and dimers, as shown in figure 9. Additionally, there were examples of Class II from across the domains of life included in the dataset, as shown in figure 10. Outgrouping was initially done using examples of class I sequences, however these sequences were later removed as class I are a poor choice of outgroup as in the Lundin (2015) evolutionary scenario they would be an ingroup. Additionally potential interference with ASR results at deep nodes indicated, as indicated in Phyre results, provided another reason to remove class I from the dataset.

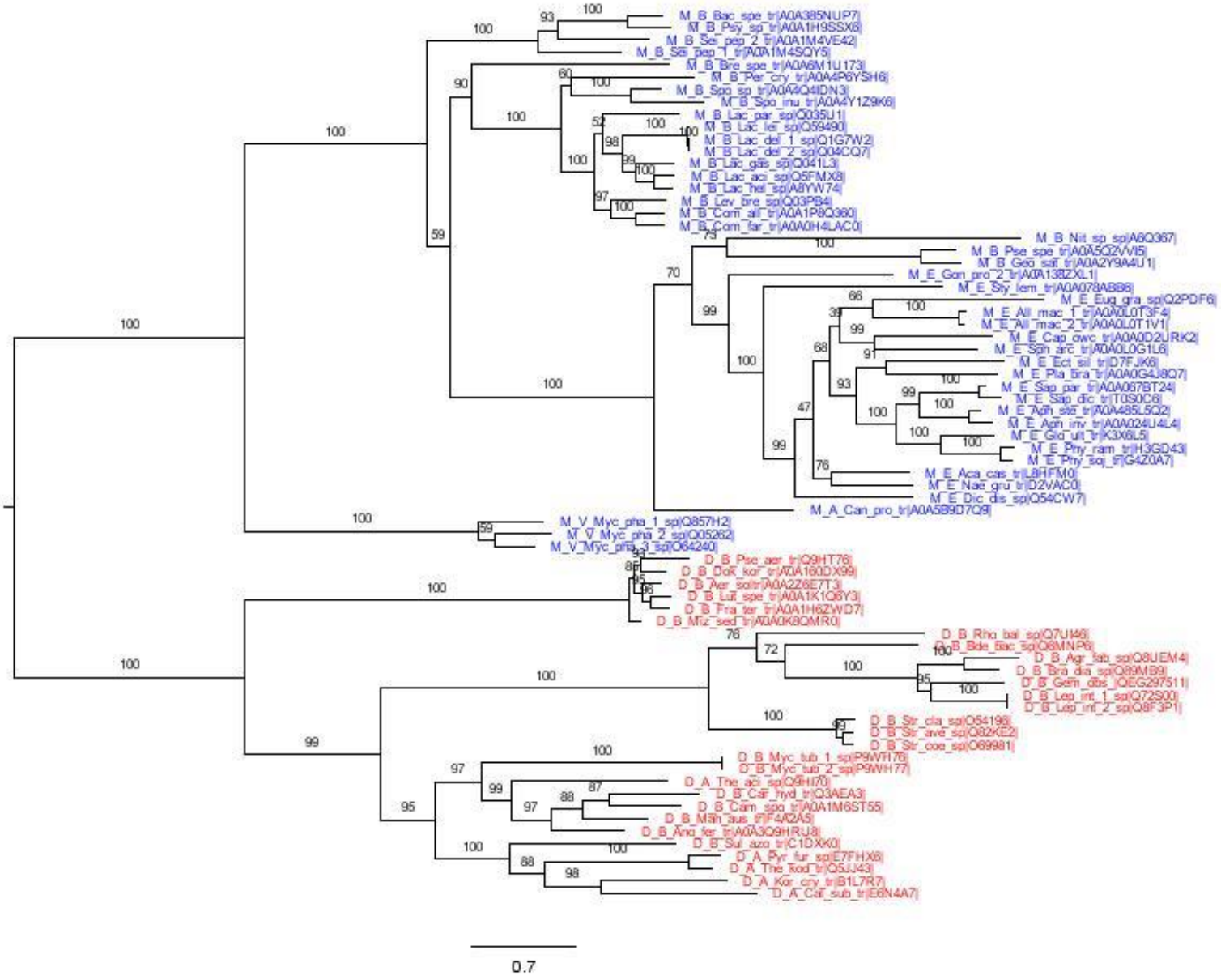

Figure 9: Tree of final dataset produced with WAG+F+I+R6 from Muscle alignment. Percent bootstrap support is indicated on the branch label. Monomers and dimers are highlighted, monomers are highlighted in blue, dimers are in red. Tree is rooted to the split between monomers and dimers.

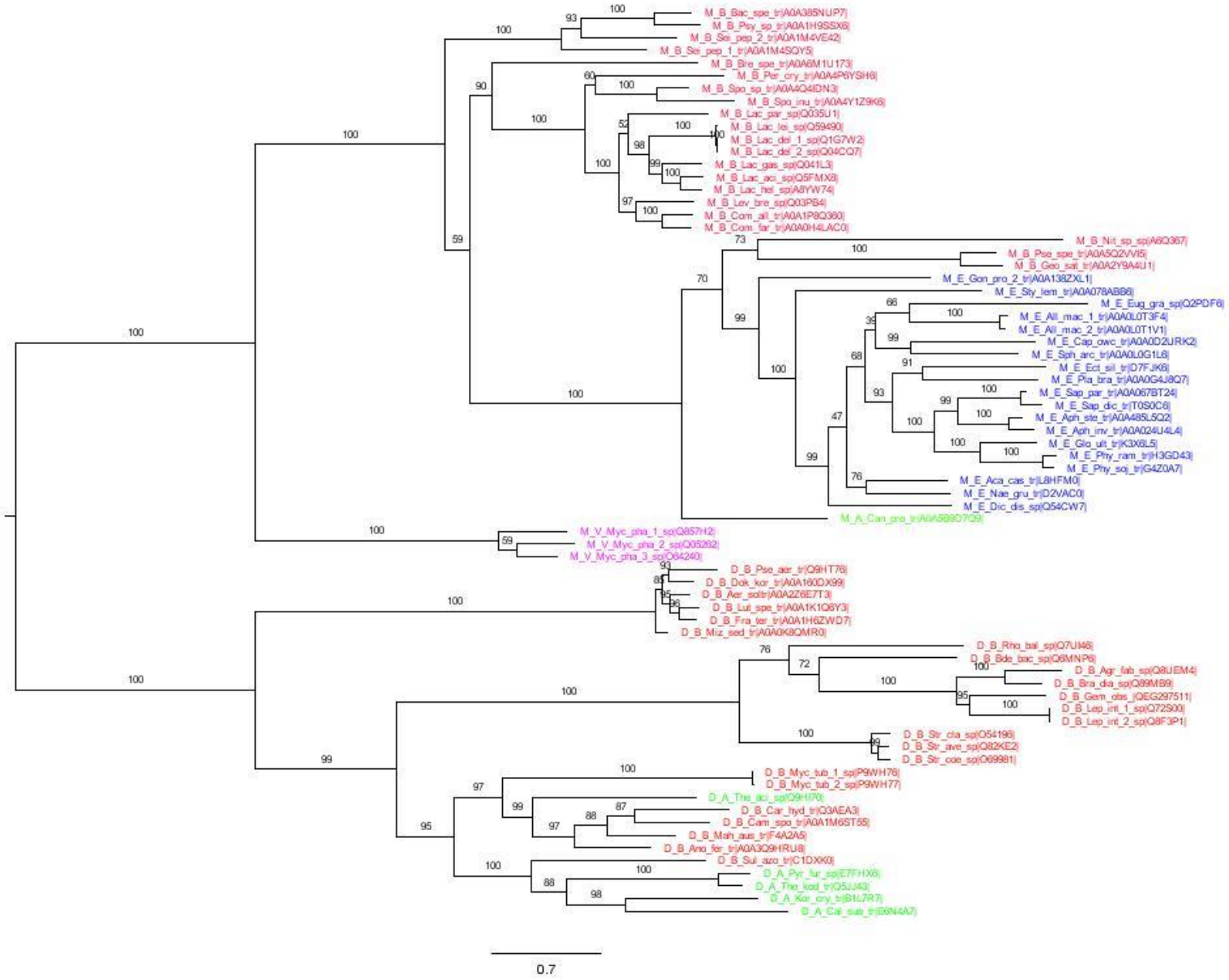

Figure 10: Tree of final dataset produced using WAG+F+I+R6 model from Muscle alignment. Percent bootstrap support is indicated on the branch label. Bacteria, archaea, eukaryotes and viruses are highlighted, with bacteria are in red, eukaryotes are in blue, archaea are in green and viruses are in pink. Tree is rooted to the split between monomers and dimers.

3.2 Sequence Alignment and ASR

Table 3: TCS scores from Tcoffee for different multiple sequence alignments.

Alignment Tool	Tcoffee Score
Cobalt Muscle	687
Mafft	692
Muscle	695
Muscle (– Class I)	703
Muscle (– Class I, no Inteins)	718

Alignment scoring using tcoffee resulted in the Muscle alignment tool being selected, with an initial alignment score of 695 in Tcoffee. as shown in table 3. Alignment using muscle maintained important motifs required for RNR function, such as the disulfide cysteins required for reduction (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002). Monomers could be distinguished from dimers based on the presence or absence of the monomeric insert (Sintchak, Arjara, Kellogg, Stubbe, & Drennan, 2002). Removing the Class I sequences which had previously been used as an outgroup increased the alignment score to 703. Additionally, muscle alignment showed that there had clearly been a mistake during curation as an intein was still included in the alignment, in the sequence D_B_Aer_soltr|A0A2Z6E7T3|. Upon removal of this intein the Tcoffee alignment score improved to 718. However, there were still a large number of areas only included in a few sequences, suggesting that alignment could be improved further. Visualistion of alignments was done using Aliview software (Larsson, 2014).

Alignment of internal nodes produced via ASR from the initial Class II dataset indicated that key conserved motifs required for RNR function were being conserved in all nodes irrespective of depth, as well as clear monomeric character for monomeric nodes and dimeric character for dimeric ancestral nodes. The most ancestral node had a dimeric character. It was noteworthy that the length of nodes was far larger than the input proteins, with the length of ASR products mirroring the length of the alignment used. For the initial ASR containing an Intein ASR products were circa 2013 positions in length. For the final ASR this was educed to circa 1780 positions upon removal of the intein from D_B_Aer_soltr|A0A2Z6E7T3|. As many of my files are not viewable in this format, alignment files, log files and tree files of sequences and of ASR nodes are included in the appendices.

3.3 Structural Analysis of ASR products

Table 4: Distribution of class I and class II identity in initial and final ASR

	Class II identity	Class I identity
Initial ASR nodes	69	6
Final ASR nodes	70	0

For the Initial ASR the majority of nodes were identified as being most similar to class II RNRs by PHYRE. However, as well as the nodes for class I sequences the deepest three nodes: Nodes 44 - 49 were classed as structurally similar to class I RNRs. Upon removal of the five Class I RNRs from the dataset for the final ASR, all resulting nodes were identified as being most similar to class II RNRs in Phyre, as shown in table 4. This suggests class I inclusion in the dataset had been affecting the ASR of deeper nodes.

4. Lab Methods:

4.1 Strains and Media

Strains and Plasmids

Table 5: Strains used and their purpose.

Strain	Purpose
REL606	PRE
Mach 1	pUc57 nrdAB extraction
ΔRNR	PRE, pUc57 nrdAB extraction
NEB 10-beta	Electroporation test

Table 6: Plasmids used and their purpose.

Plasmid	Purpose
NEB pMiniT 2.0 Vector	PRE
pUc57 nrdAB	PRE

REL606 was the main *E. coli* strain used throughout this experiment and was sourced from Richard Lenski's long term evolution experiment (University of Houston, Texas). Mach 1, ΔRNR and NEB 10-beta strains were used for extraction of a pUc57 plasmid containing the nrdAB operon and electroporation respectively. These strains were provided by Samantha Arras (University of Auckland) and Danielle Maddock (University of Auckland). ΔRNR RNR knockout lines from Arras et

al (2023), created by a scarless genome engineering protocol (Fehér et al., 2008, as cited in Arras et al., 2023) was provided to me by the Poole Lab (University of Auckland). Gemmata obscuriglobus was used as it was the cell line available in the lab which contained a Class II RNR. E. coli cell lines were grown at 37 °C, whereas G. obscuriglobus was grown at 28 °C due to it's slower growth rate. Plasmids used in this experiment include the NEB pMiniT 2.0 Vector from the NEB PCR cloning kit (NEB #E1203S), as well as a pUc57 plasmid (Thermo Fisher #SD0171) containing the nrdAB operon extracted from Mach 1, created by Samantha Arras (University of Auckland). Both plasmids carry an Ampicillin resistence gene.

Media

Liquid media used in this experiment were standard Luria Burtani (LB), prepared by weighing out 25g of pre-mixed tryptone, yeast extract and sodium chloride (NaCl) powder per litre of water. This mixture was autoclaved to ensure sterilisation. Super Optimal broth with Catabolite repression (SOC) was used for recovery of cells after transformation.

Instructions for SOC

1L	Final []	Component
5g	0.5%	Yeast extract
20g	2.0%	tryptone
0.5g	10mM	NaCl
0.186g	2 mM	KCL
2.4g	20mM	$MgSO_4$

Added 800ml of dH_2O and dissolved components. Adjusted to pH 7.5 prior to use with 1 M NaOH, then added dH_2O to the final volume of 960ml and autoclaved. When solution had cooled to 50°C added 40ul of sterile 10% glucose (Barrick Lab, 2023).

Solid media were all produced by adding two grams of agar to one hundred mL of liquid media to produce a 2% gel plate, which was autoclaved and poured in all cases except for MOPS minimal media. MOPS minimal media was produced following the protocol from the E. coli Genome project (University of Wisconsin – Madison). MOPS has to be filter sterilised as the mix is heat sensitive, so to produce solid media agar was added before gently heating and mixed to melt the agar and

allow pouring of plates. MOPS media was supplemented with equal concentrations of the four Deoxyribonucleotides to a total concentration of 1g/L of media where specified.

MOPS Minimal Medium:

990ml	Component
100ml	10X MOPS mixture
10ml	0.132 M K_2HPO_4
880ml	milliQ H_2O

Mixed ingredients above a2nd adjusted pH to 7.2 with 10 M NaOH, then filter sterilised. Added 10ml of 1X glucose for a final percentage of 1%, then stored at -20°C.

10X MOPS Mixture:
In a 1 L beaker with a sterile stir bar, added the following to ~300ml of millQ H_2O:

Grams	Formula Weight (FW)	Component
83.72	209.3	MOPS
7.17	179.2	Tricine

Added 10 M KOH for a final pH of 7.4, then added milliQ H_2O to 440ml, then made fresh $FeSO_4$ solution and added to the MOPS/Tricine solution.

Component	FW	grams	H_2O vol (ml)	Stock conc (M)
$FeSO_4 \cdot 7H_2O$	278	0.028	10	0.01

Then added the following solutions to the MOPS/Tricine/$FeSO_4$ solution in the order shown:

Volume	Component
50ml	1.9 M NH_4Cl
10ml	0.276 M K_2SO_4
0.25ml	0.02 M $CaCl_2 \cdot 2H_2O$
2.1ml	2.5 M $MgCl_2$
100ml	5 M NaCl
0.2ml	Micronutrient stock
387ml	Autoclaved milliQ H_2O

The total volume should come to 1000ml. Filter sterilised with a 1 L capacity 0.2-micron filter. Aliquoted into falcon tubes and frozen at -20°C.

Stocks used in 10X MOPS mixture:
Each made up separately and stored at room temperature.

Component	FW	Stock conc (M)	grams	Vol (ml)
NH_4Cl	53.49	1.9	50.82	500
K_2SO_4	174.3	0.276	4.8	100
$CaCl_2 \cdot 2H_2O$	147	0.02	0.294	100
$MgCl_2$	203.3	2.5	50.75	100
NaCl	58.44	5	292.2	1000

Micronutrient Stock (100ml)
Mixed all components in 40ml autoclaved milliQ H_2O, then brought total volume to 50ml and stored at room temperature.

Component	Formula	FW	Grams for 50ml
Ammonium molybdate	$(NH_4)_6Mo_7O_{24} \cdot 4H_2O$	1235.9	0.009
Boric acid	H_3BO_3	61.83	0.062
Cobalt chloride	$CoCl_2$	237.9	0.018
Cupric sulfate	$CuSO_4$	249.7	0.006
Manganese chloride	$MnCl_2$	197.9	0.040
Zinc sulfate	$ZnSO_4$	287.5	0.007

Potassium phosphate K_2HPO_4 Solution

Component	FW	Stock conc (M)	grams	Vol (ml)
K_2HPO_4	173.2	0.132	23.0	1000

Made sure to use dibasic K_2HPO_4 and not monobasic K_2HPO_4. Autoclaved then stored at room temperature.
Protocol from the E. coli Genome Project (University of Wisconsin – Madison)

M1 media (for Gemmata)

M1 medium
M1 media was made following the protocol from the Poole Lab (University of Aucklnd).

1 L	Component
2.5g	$CaCO_3$
0.05g	$Na_2HPO_4.12H_2O$
10mL	Hunter's Basal Salts
7.5g	Agar
to 500mL	Distilled water

For M1 media dissolved all components listed above (CaCO3 is insoluble) and autoclaved. Cooled to 55°C and add the following components:

Vitamin solution #6. 5mL required per 500mL of M1
N-acetylglucosamine. 1.0g (dissolved in 10mL of water and filter sterilized)

Mixed well to disperse $CaCO_3$ before pouring plates. After pouring, allowed to dry in the laminar flow for at least 30 minutes, then sealed plates with parafilm.

Hunter's Basal Salts (require 10mL per 500mL of M1 media)

1 L	Component
5g	Nitrilotriacetic acid
14.85g	$MgSO_4.7H_2O$
1.67g	$CaCl_2.2H_2O$
6.335g	$Na_2MoO_4.2H_2O$
49.5mg	$FeSO_4.7H_2O$
25mL	Metals 44 solution
to 500mL	Distilled water

Dissolved nitrilotriacetic acid first by neutralization with KOH before addition of other salts. Adjusted the pH to 7.2 with H_2SO_4, then autoclaved.

Vitamin solution #6 (only 5mL per 500mL of M1)

1 L	Component
2.0mg	Biotin

5.0mg	Calcium pantothenate
2.0mg	Folic acid
5.0mg	Nicotinamide
5.0mg	Pyridoxine HCl
10.0mg	Riboflavin
5.0mg	Thiamine HCl
1000mL	Distilled water

Mixed the above components, filter sterilized and stored at 4°C in the dark, aka the bottle was covered in foil and placed in the fridge

Metals 44 (25mL requires per 500mL of HBS)

1L	Component
250mg	Na-EDTA
1095mg	$ZnSO_4.7H_2O$
500mg	$FeSO_4.7H_2O$
154mg	$MnSO_4.H_2O$
39.2mg	$CuSO_4.5H_2O$
24.8mg	$Co(NO_3)_2.6H_2O$
17.7mg	$Na_2B_2O_7.10H_2O$
1000mL	Distilled water

Dissolved the EDTA and add a few drops of H_2SO_4 to slow precipitation of metal ions. Does not require autoclaving as it will be added to HBS before that is autoclaved.
Streptomycin and ampicillin were added to media when appropriate to concentrations of 50µg/mL and 100µg/mL of media.

4.2 Primer Design
Primers for the region containing the nrdJ in Gemmata obscuriglobus were created using Geneious Prime 2023.1.2 (https://www.geneious.com), making use of the inbuilt Primer creator tool on the genome of *Gemmata obscuriglobus*, available from the NCBI website (NCBI, 2023). The RNR gene runs from 2,691,273-2,695,014bp. Two forward and two reverse primers were designed, with one located as close as possible to the Gene, while the other was placed further away from the gene to include the native promotor. The two forward primers were placed at 2,690,000bp and 2,685,000bp, while the two reverse primers were located at 2,695,300bp and 2,700,000bp. Primers for REL606 RNR operon were taken from (Arras, et al., 2023), and are referred to as the +800 primer and -800 primer based on their position

relative to the RNR nrdAB operon. All Primers were ordered from Integrated DNA Technologies (2023 Integrated DNA Technologies, Inc).

Table 6: Primers used in experiments, including sequence and target strain

Primer	Sequence	Strain
nrdAB +800	5' –CGATACTCCAGTCCTGCGTAATGC– 3'	REL606
nrdAB -800	5' –CGACCAACGATTGTCCGTGAGG– 3'	REL606
GoCII2.685 F	5'-CGGGCAGTAATAGGTTTTAACTGC-3'	*G. obscuriglobus*
GoCII2.691 F	5'-GTATGCGCGGTTCATGTC -3'	*G. obscuriglobus*
GoCII2.695 R	5'-AAAAACAAACCGCGGTCACG-3'	*G. obscuriglobus*
GoCII2.700 R	5'-TCTGGATTCACTGCCGCCG-3'	*G. obscuriglobus*

4.3 RNR Gene Extraction

Initially a direct PCR of *G. obscuriglobus* carried out across a temperature gradient was used to identify which temperature was most suitable for the primers created, as these can differ from the values provided by simulations of the primers. The PCR was set up to run a test and a control for each temperature from 51-62°C, totalling 24 PCR tubes. This was done using a master mix of 25x the values of a standard 10µL reaction, totalling 125µL of Kappa 2 hotstart polymerase, 6.25µL GoCII2.691 F, 6.25µL, GoCII2.695 R and 82.5 µL of water, and 5 µL of Triton X 100, a detergent included to aid the lysing of cells (Shatzkes, Teferedegne, & Murata, 2014) as *G. obscuriglobus* has an intracytoplasmic membrane (Devos, 2013). For test repetitions 4 µL of extracted DNA was added to 6 µL of Master Mix, whereas for controls 4 µL of water was added. The conditions of the PCR were a hot start with a 10-minute initial period at 98°C, then 35 repetitions with a denaturation time of 15 seconds at 98°C, annealing for 15 seconds between 51-62°C and extension at 72°C for 1 minute and 10 seconds. This was followed by a hold period of 5 minutes at 72°C. To visualise whether the PCR had run successfully 5 µL of each PCR reaction was run on a 1% agarose gel with redsafe at a concentration of 4 µL/100mL across 24 lanes. Repeats were carried out using the same conditions.

For a whole genome extraction *E. coli* was grown in 10ml of LB in a McCartney bottle shaken at 37°C in an incubator. *G. obscuriglobus* cells were taken directly from the freezer stock, with a scraping was placed in sterilised water to carry out

the extraction protocol. Extraction was carried out using the wizard kit (Promega #A1225). 1ml of overnight culture was placed in a 1.5ml microcentrifuge tube and centrifuged at 14,000 × g for 2 minutes to pellet the cells. Remove the supernatant. Add 600µl of Nuclei Lysis Solution and gently pipet until the cells are resuspended. Incubate at 80°C for 5 minutes then cool to room temperature. Add 3µl of RNase Solution and invert the tube 2–5 times to mix. Incubated at 37°C for 60 minutes, then cool the sample to room temperature. Add 200µl of Protein Precipitation Solution then vortex at high speed for 20 seconds to mix. Incubated the sample on ice for 5 minutes then centrifuged at 14,000 × g for 3 minutes. Transferred the supernatant containing the DNA to a clean 1.5ml microcentrifuge tube containing 600µl of room temperature isopropanol. Gently mixed by inversion until the thread-like strands of DNA form a visible mass. Centrifuged at 14,000 × g for 2 minutes. Carefully pour off the supernatant and drain the tube on clean absorbent paper. Add 600µl of room temperature 70% ethanol and gently invert the tube several times to wash the DNA pellet. Centrifuge at 14,000 × g for 2 minutes, then carefully aspirate the ethanol. Drain the tube on clean absorbent paper and allow the pellet to air-dry for 10–15 minutes. Added 100µl of DNA Rehydration Solution to the tube and rehydrate the DNA by incubating the solution overnight at 4°C and stored at 4°C. This was done with three repetitions, T1, T2 and T3 DNA concentration from these repetitions was recorded using a nanodrop spectrophotometer. PCR of the nrdAB operon in REL606 was then carried out using 20µL of Kappa hotstart, 2 µL of +800 primer and 2 µL of the -800 primer. The conditions of the PCR were a hot start with a 10-minute initial period at 98°C, then 35 repetitions with a denaturation time of 15 seconds at 98°C, annealing for 15 seconds at 60°C and extension at 72°C for 6 minutes. This was followed by a hold period of 5 minutes at 72°C. There was also a subsequent repeat for the REL606 extraction where the annealing temperature was dropped to 57°C. To visualise whether the Genome Extraction and the PCR of the RNR gene had been successful 5 µL of Genome extract and PCR reaction was run on a 1% agarose gel with redsafe at a concentration of 4 µL/100mL to test whether the PCR had been successful. Later repetitions of the PCR reaction were followed by a PCR cleanup.

PCR cleanup was carried out using DNA from genome extractions T1 and T2 were used with 4 repetitions from each extraction. These were named with the first digit

representing the genome extraction they were from while the second digit indicated what the repetition was. The kits used for the cleanup were a NEB (# T1030S) and a Macherey Nagel (#740609.50) kit.

4.4 Plasmid Miniprep

Strains containing nrdAB plasmids, Mach 1, 2 strains of REL606 and RNR31 (?), removed from freezer stocks and grown overnight in 10 ml of LB. Thermo Fisher's GeneJET miniprep kit was used as per the protocol provided with the kit in appendix 7. Resulting material was run on a 1% agarose gel with redsafe at a concentration of 4 µl/100ml.

4.5 Making Electrocompetent Cells

Initially the protocol used was from the Barrick lab (Barrick Lab, 2023). Overnight cultures of REL606 or ΔRNR were grown in LB medium + Strep at 50µg/mL. 10 ml of fresh LB in a 50ml flask was then inoculated with 200 µl of overnight culture to an OD_{600} of ~0.05. Cells were grown incubated at 37°C shaking for 2-3 hours until they reached an OD_{600} of ~0.6. Cells were then transferred to 15 ml Falcon conical tubes. Pellet the cells by centrifugation for 5 minutes at 4000 RPM. Remove and pour off supernatant. Washed by adding 10 ml of chilled 20% glycerol to each tube, then vortexing to resuspend the pellet. Centrifuge for 3.5 minutes, remove and pour off supernatant. Repeat for at least four wash cycles. Resuspend in approximately 100 µl of 20% glycerol to produce a 100x concentration of the initial culture. These were then divided into 50 µl aliquots and frozen.

Later repetitions were carried out using a protocol provided to me by Danielle Maddock (University of Auckland). Use a 3ml aliquot of overnight culture was to inoculate 350 ml of LB in a sterile conical flask, incubate at 37°C shaking until the OD_{600} of the culture reached 0.35-0.4. decant cells into six (or more) 50 mL Falcon tubes and cooled on ice for 20 minutes. Cells were then pelleted by centrifugation at 4°C, 1700 × g for 15 minutes. Pour off supernatant. Resuspend each pellets in ~5 mL ice-cold, sterile H2O, and then top up to 45 ml. Re-pellet at 4°C, 1700 × g for 15 minutes. Resuspend three pellets in ~5 mL ice-cold sterile H2O, then use each suspension to resuspend a second pellet, so there are three tubes containing bacteria. Top these up to 45 ml with sterile ice-cold H2O. Re-pellet at 4°C, 1700 × g for 15 minutes. Re-suspended cells in ~5 mL ice-cold sterile 10% (w/v) glycerol and

combine into a single 50 mL Falcon tube. Top up to 45 mL, then pellet at 4°C, 1700 × g for 15 minutes. Discard supernatant, and re-suspend pellet in residual liquid. A 1:100 dilution of cells should give an OD600 ~0.1 – either: pellet and remove further supernatant, or add additional 10% (w/v) glycerol until this value is reached. Split cells into 50 µL aliquots and store at -80°C

4.6 Ligation and Transformation

Ligation was carried out using the NEB cloning kit without electrocompetent cells (NEB #E1203S). Ligation reactions were assembled with 1 µL of Linearised pMiniT™ 2.0 Vector, 4 µL of DNA insert or 2 µL of the provided Amplicon Cloning control, with H_2O to 5 µL, then 4 µL of Cloning Mix 1 and 1 µL of Cloning Mix 2. This was incubated at room temperature for the maximum 15 minutes recommended in the kit due to the 5.3kb insert size of the RNR gene extraction. Later iterations also included incubation for 20 minutes. This was then incubated on ice for 2 minutes and then stored at -20°C. PCR was carried out using the primers provided with the NEB kit, located either side of the insertion site. The conditions for the PCR were a 10-minute initial period at 98°C, then 30 repetitions with a denaturation time of 15 seconds at 98°C, annealing for 15 seconds at 53°C and extension at 72°C for 2 minutes for the control insert and 6 minutes for test inserts. This was followed by a hold period of 5 minutes at 72°C. To visualise whether the PCR had run successfully 5 µL of each PCR reaction was run on a 1% agarose gel with redsafe at a concentration of 4 µL/100mL across 24 lanes.

Transformations were initially carried out using a heat shock protocol detailed in the NEB cloning kit (NEB #E1203S). Thaw a 50 µl tube of competent cells on ice for 10 minutes. Add 2 µl of ligation reaction; gently mix by flicking the tube 4–5 times. Incubate on ice for 20 minutes. Heat shock at 42°C for 30 seconds. Chill on ice for 5 minutes. Add 950 µl of SOC and place at 37°C for 60 minutes shaking at 250 rpm prior to plating. These repetitions were carried out using electrocompetent cells made using REL606, using both protocols to produce electrocompetent cells as described in the making electrocompetent cells section. Transformation was attempted with the ligation reactions as well as the pUc57 plasmid from the plasmid miniprep. Electrocompetent cells and vectors were first thawed on Ice for 20-30 minutes, then 2 µL of vector was added to electrocompetent cells and mixed by

flicking the tube. These were then incubated on ice for 30 minutes, followed by heat shocking the bottom half of the tube in a 42°C water bath for 60 seconds. Samples were placed back on ice for 2 minutes, before recovery in 300 µL of SOC in the 37°C shaking incubator for 45 minutes. Cells were then plated on LB + Amp plates to select for successful transformants and incubated overnight at 37°C.

Later transformations were carried out via electroporation. For these transformations three repetitions were carried out. An initial repetition was done using NEB-10 Beta electrocompetent cells. Two subsequent repetitions were carried out, one using REL606 electrocompetent cells, the other using ΔRNR. Transformations were done using the pUc57 plasmid extracted from Mach 1 in the miniprep. 1 µL of vector was added to 50 µL of electrocompetent cells, then shocked at 2500 volts for 5 milliseconds. Cells were then recovered in 1 ml of SOC, incubated at 37°C shaking for 30 minutes. 100 µL of the recovery media was then plated on LB + Streptomycin (50µg/mL) plates to test for cell viability, as well as LB + Amp plates to select for successful transformants and incubated overnight at 37°C. Additionally plates of MOPS with ampicillin and with and without DNS added were used to test whether the media could select for cells with a functioning RNR. Test of whether MOPS with vs without dNS could be used to test for the absence of a functional RNR gene was tested by growing ΔRNR and wt REL606 on MOPS and MOPS + DNS in a 4 x 6 well plate.

Table 7: Layout of growth experiment of ΔRNR and wt REL606 on MOPS and MOPS + dNS

X	1	2	3	4	5	6
A	ΔRNR MOPS	ΔRNR MOPS	ΔRNR MOPS	X	X	X
B	ΔRNR MOPS + dNS	ΔRNR MOPS + dNS	ΔRNR MOPS + dNS	X	X	X
C	Wt REL606 MOPS	Wt REL606 MOPS	Wt REL606 MOPS	X	X	X
D	ΔRNR MOPS + dNS	ΔRNR MOPS + dNS	ΔRNR MOPS + dNS	X	X	X

5. Lab Results:

5.1 RNR Gene Extraction

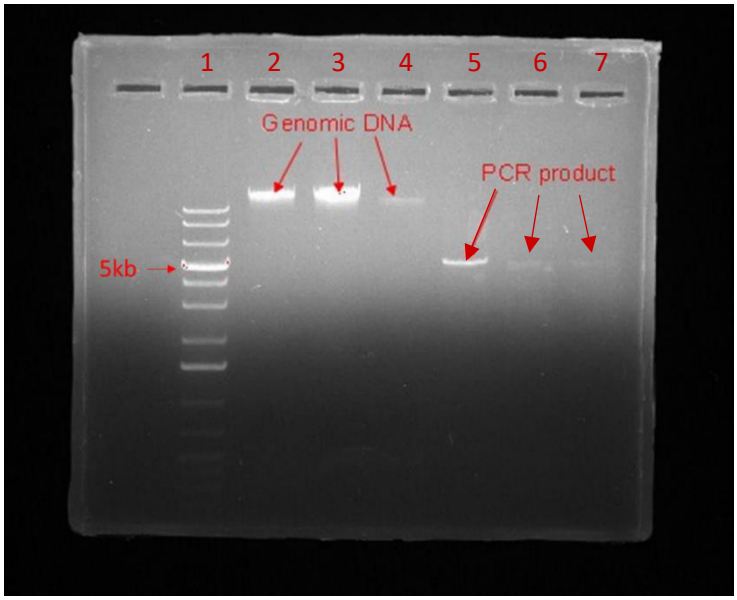

Figure 11: Gel of Genome extraction and PCR of three repetitions from REL606 with a 1kb ladder. Expected bands were Genomic DNA above the ladder and PCR products of circa 5.3kb in size. Genome extractions 1 and 2 appear to be successful, while only PCR 1 has been successful. Lanes are numbered 1-7 for reference.

Attempts at genome extraction from *G. obscuriglobus* proved unsuccessful. While the cells did grow in M1 media, gels of the genome extraction products were blank other than the 1kb ladder. Genome extraction and PCR from REL606 proved more successful, as seen in figure 23. The whole genome extraction appears to have been successful for all three repetitions due to the presence of bands above the highest band of the 1kb ladder in lanes 2,3 and 4. Repetition 1 of the PCR appears to have been successful as there is a clear band present in lane at circa 5.3kb, which is the size of the region between the +800 and -800 primer containing the RNR gene, while repetition 2 and 3 have fainter bands.

Table 8: DNA concentration of PCR cleanup of PCR of nrdAB operon from wt REL606.

NEB Buffer		MN Buffer	
Repetition	DNA Concentration (ng/µl)	Repetition	DNA Concentration (ng/µl)
11	2.7	13	39.95
12	11.1	14	13.75
21	10.75	23	11.5
22	19.1	24	13.5

PCR cleanup of the PCR product from the PCR of the nrdAB operon from wt REL606 had mixed results, as can be seen in the table above. In this table the first digit indicates the repetition of the PCR product used in the cleanup, while the second digit indicates what repetition of the PCR cleanup it is. The highest DNA concentration was from repetition 13, the third PCR cleanup repetition from repetition one of PCR, at 39.95 ng/µl, while the lowest concentration was from repetition 11 at 2.7 ng/µl, with other results ranging between 10-20 ng/µl. There did not appear to be a major difference between results based on the genome extraction or the protocol which was used.

5.2 Plasmid Miniprep

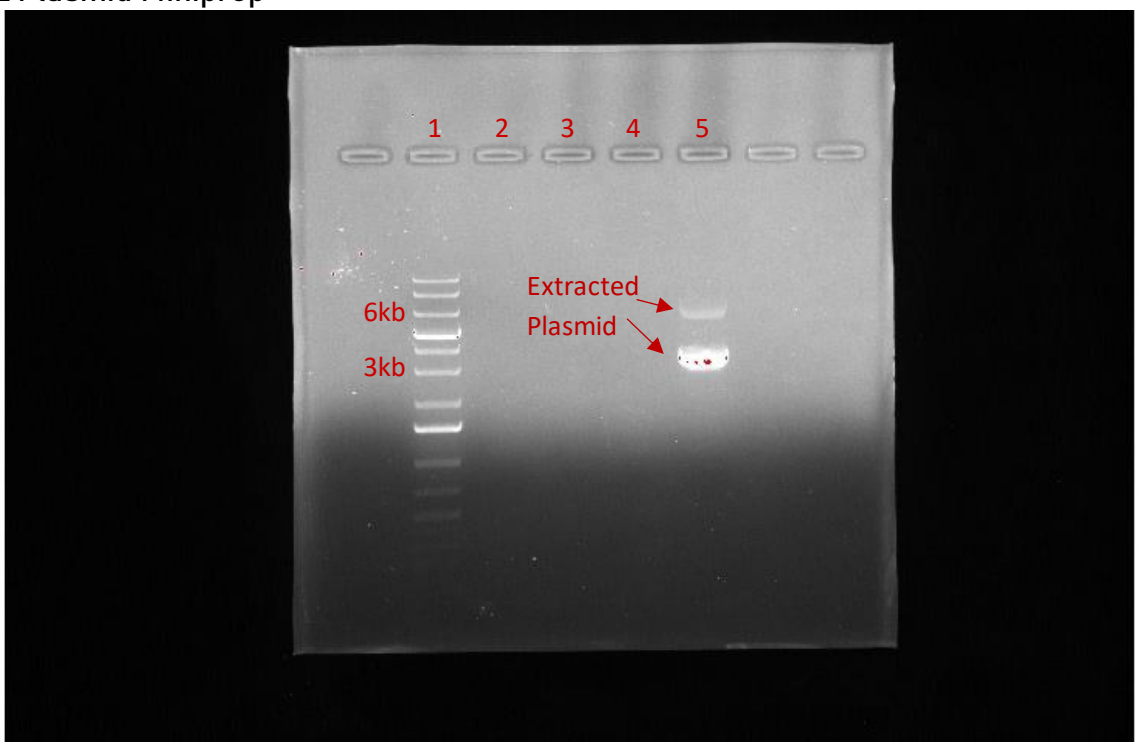

Figure 12: Gel of Plasmid Miniprep from 4 strains of E. coli with lanes and band size indicated. Only lane 5 containing the plasmid miniprep of a puc57 plasmid from Mach 1 E. coli appears to have run, with a strong band at circa 3kb and a weaker band at circa 6kb, as expected with different forms plasmids can take in electrophoresis (Schmidt, Friehs, Schleef, Voss, & Flaschel, 1999).

Miniprep from REL606 and RNR31 strains appear to have failed as no product ran in the first 3 lanes. Extraction from Mach 1 appears to have been successful as there is a strong band at circa 3kb on the 1kb ladder, indicating extraction of the puc57 plasmid containing the nrdAB operon.

5.3 Ligation and Transformation

Figure 13: Transformation of puc57 nrdAB plasmid into NEB-10 Beta E. coli, restreaked onto LB with Ampicillin. Growth indicates transformation has been successful.

Table 9: Table of initial transformations via electroporation and their results.

Repetition	Result
NEB-10 Beta + Ligation Product	Failed transformation
NEB-10 Beta + pUc57 nrdAB	Successful transformation
wt REL606 (Barrick Lab prep) + pUc57 nrdAB	Failed Transformation

While multiple repetitions of the ligation reaction using the NEB kit were attempted, no successful ligations for either the control or the insert were produced. This was indicated by a lack of growth for transformed cells on selective media, as well as a lack of PCR product using the test primers provided in the NEB kit. Therefore, subsequent transformations were attempted using the puc57 plasmid extracted from Mach 1. The initial test transformation into NEB-10 Beta prepared using Danielle Maddocks cell preparation method and electroporation rather than heat shock. Repetitions using cells prepped using the barrick cell prep method as well as the product from ligation attempts both failed, indicating that there were likely issues with the previous protocols.

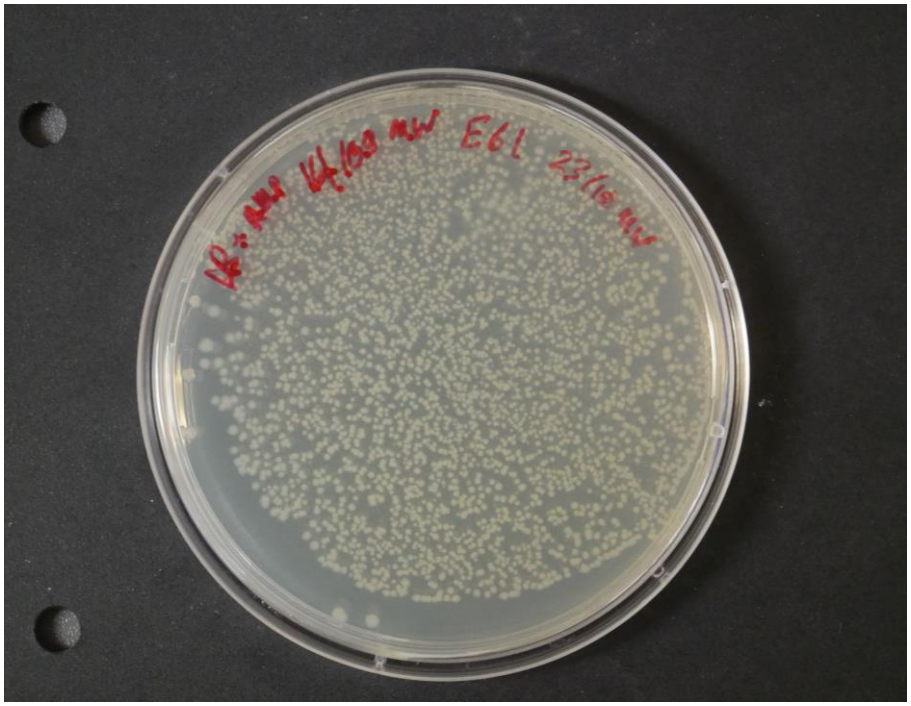

Figure 14: Transformation of puc57 nrdAB plasmid into wt REL606. Plated onto LB with Ampicillin. Growth indicates transformation has been successful.

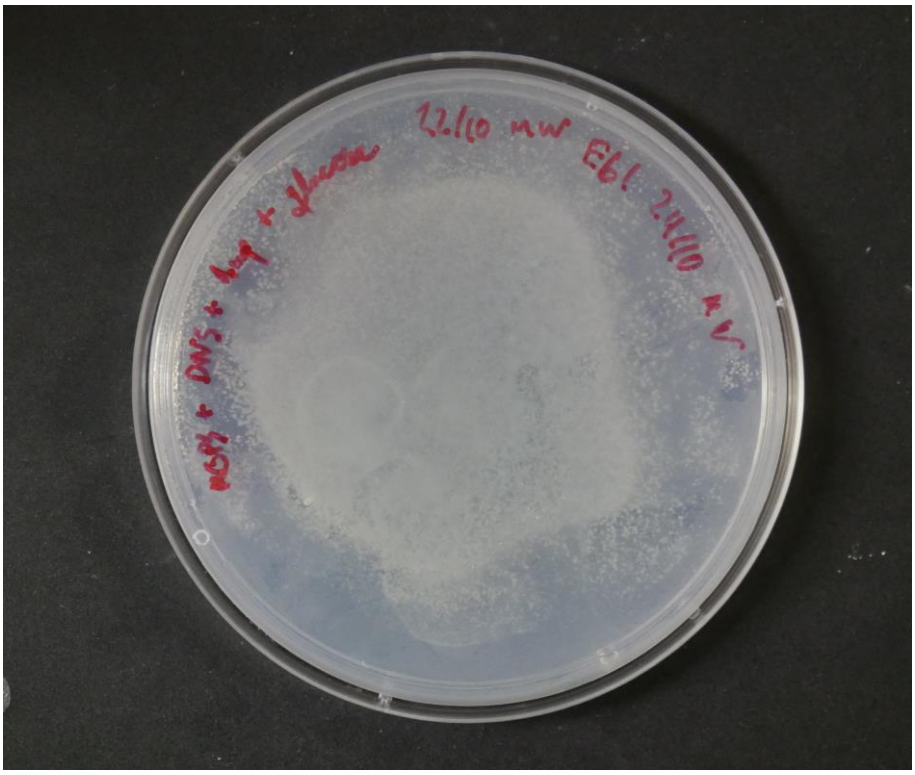

Figure 15: Transformation of puc57 nrdAB plasmid into wt REL606. Plated onto MOPS with dNS, Ampicillin and 1% glucose. Growth indicates transformation has been successful.

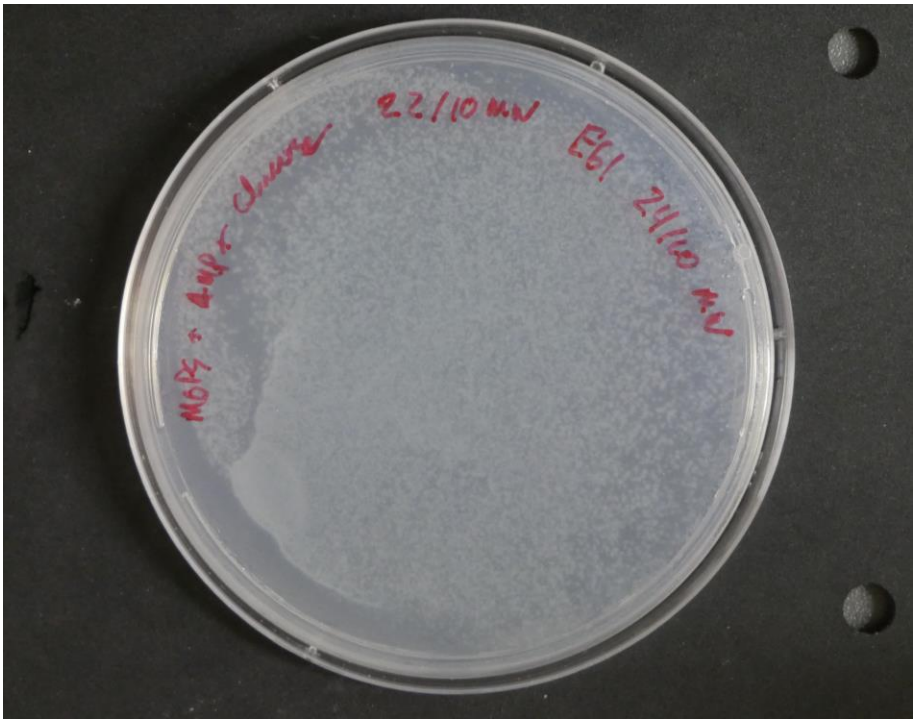

Figure 16: Transformation of puc57 nrdAB plasmid into wt REL606. Plated onto MOPS with Ampicillin and 1% glucose. Growth indicates transformation has been successful.

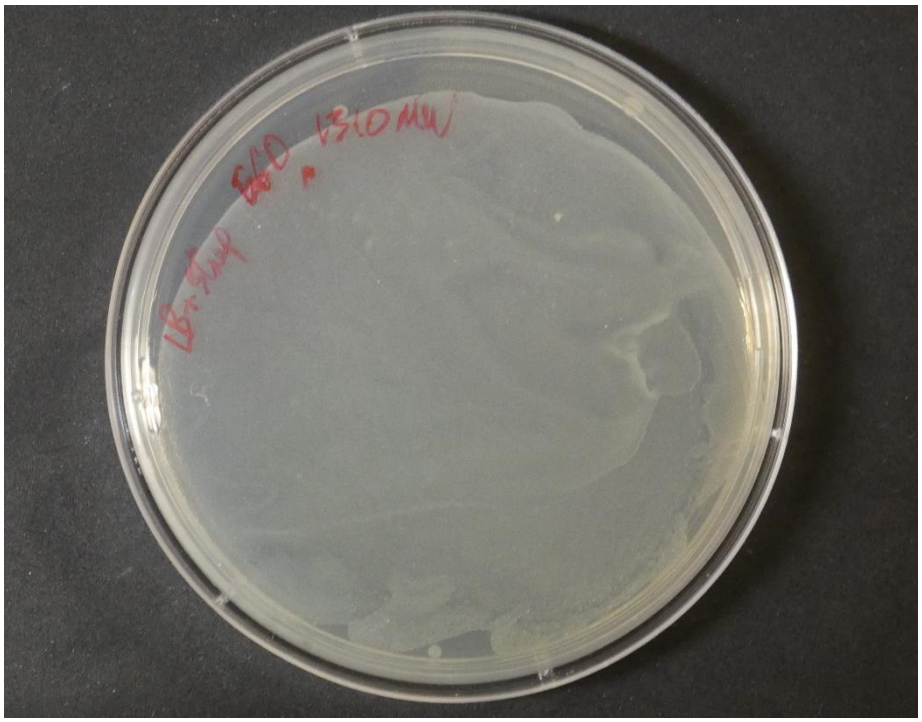

Figure 17: Transformation of puc57 nrdAB plasmid into wt REL606. Plated onto LB + Streptomycin. Growth indicates the number of competent cells.

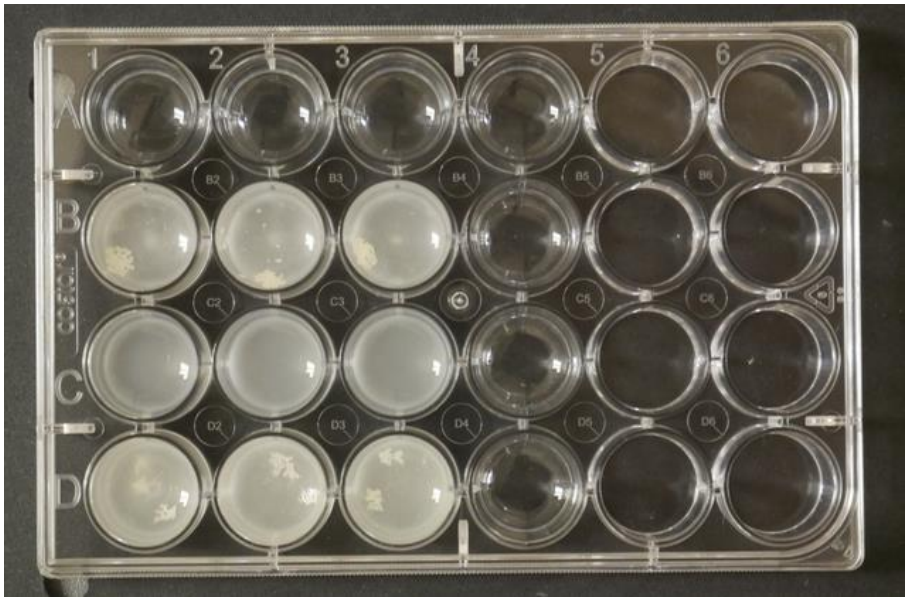

Figure 18: Growth test of ΔRNR and wt REL606 on MOPS and MOPS + DNS. The first row is ΔRNR with just MOPS, with no growth. The second row has ΔRNR on MOPS + DNS and has grown. The third and fourth rows have wt REL606 on MOPS and MOPS + DNS respectively, with both rows indicating growth. The lack of growth in the first row indicates that the knockout cannot survive on MOPS without supplementation with DNS.

Figure 19: Transformation of puc57 nrdAB plasmid into ΔRNR. Plated onto LB + streptomycin. Growth indicates that cells survived electroporation.

Using the pUc57 plasmid extracted from mach 1 successful transformation into both Neb 10 beta, as shown in figure 13 and wt REL606 were achieved, as shown by growth on selective media including with repetitions on both LB, shown in figure 14 and MOPS in figures 15 and 16. Background growth was tested on Lb with streptomycin added as per the media chapter. Transformations into ΔRNR were unsuccessful, as growth was only observed on plates with only streptomycin and no ampicillin present, shown in figure 19 indicating that the plasmid with amp resistance hasn't been taken up. However, the growth pattern of ΔRNR on MOPS with or without DNS supplementation was as expected, as shown in figure 18, indicating that a successful transformation should have allowed for growth on MOPS + Ampicillin + 1% glucose.

6. Discussion:

For the analysis section the results were largely positive. The final dataset produced during this research is representative of the diversity present in class II RNRs both in terms of their structure and the domains in which they are found. The alignments of these proteins improved as the experiments progressed and issues with the dataset were addressed. ASR of potential Class II ancestors ran successfully and did produce internal nodes which were identified as being most closely related to Class II RNRs. The iterations of datasets and alignments used provide a snapshot of issues the which can arise in phylogenetics and how they can be identified and ameliorated. For example, the failure of initial curation to remove all inteins from the dataset was only identified during the alignment phase and led to a rerun of those alignments, resulting in a higher alignment score, as did the removal of class I sequences from the dataset, while simultaneously improving ASR results. The iterative refinement which occurred during dataset curation was gratifying and showed that there were always questions to be asked about what sequences were included and why.

However, it is important to note that are unavoidable biases in the databases this research will be using in terms of what RNR sequences have been gathered. This will be due to biases in what microbes are researched (Lemos, Fulthorpe, Triplett, & Roesch, 2011). By searching databases using multiple query sequences followed by curation I hope to have captured as much of the class II diversity as possible. A

limitation of ASR is detailed in the Randall et al (2016) paper where they found lower accuracy for phenotype prediction than sequence prediction, which in their experiment leading to bioluminescent ASR sequences with differing colours and brightness compared to the actual ancestral protein. This difference in phenotype could occur during ASR for this experiment especially for deeper nodes, resulting in proteins with reduced function. Additionally, this reduced function may not be detected if purely using a PRE as a binary question of 'does growth occur or not?'. There isn't a great deal that can be done directly to increase phenotype prediction as ASR methods are probabilistic, and therefore have inherent uncertainty. If the sequences included in the dataset and the phylogeny for the ASR are well curated, this will reduce the likelihood of ASR producing non-functional nodes. Additionally, if growth times for transformed cells are recorded during the PRE this could give an indication of how functional the proteins produced via ASR are without adding to the time required for this research or significantly adding to the workload of the researcher.

A particular issue with alignments during this experiment was the large number of gaps caused by unique poorly aligned areas in the alignment only being present in a small subset of the dataset proteins. This resulted in the internal nodes produced by ASR being far longer than the RNRs in the dataset. This could be addressed by identifying whether the areas increasing the length of the alignments are located in regions essential to function and could potentially be removed, or whether these gaps are due to a small number of sequences which could be excluded from further study. Another potential option is constraining what areas of the protein are included for ASR using a chassis model, whereby the majority of the protein is kept constant with only certain regions of interest produced vias ASR (Islam, Lin, Lai, Matzke, & Baker, 2020). However this approach could interfere with one of the key purposes of this research, which was to use the diversity of class II RNRs to try and identify the character of ancestral RNRs An additional feature of the internal nodes was that there weren't a high number of gaps introduced, despite large differences between monomers and dimers which should have led to gaps in, for example, the monomeric insertion for dimeric ancestors. A potential way to address this is to use parsimony to determine what insertions and deletions are most likely at different

points in the tree, as suggested by Mascotti (2022), ordering multiple copies of nodes where parsimony doesn't provide a clear answer.

In terms of the lab work associated with this project the ability to pivot and find alternative methodologies proved hugely important in eventually producing a successful transformation. Initial whole genome extraction from wt REL606, RNR operon extraction via PCR with +800 and -800 primer and PCR cleanup were successfully carried out. However, as the ligation reaction of this extracted DNA into a plasmid did not work an alternative had to be found. Carrying out a plasmid miniprep of an existing nrdAB containing plasmid allowed subsequent transformations to be attempted, allowing successful transformation into the wt REL606, as shown by its survival on media containing ampicillin. It is unfortunate that transformation into ΔRNR was unsuccessful, as this would have been a true example of a phenotypic rescue. Additionally on the control plates for this experiment using Lb with streptomycin ΔRNR grew only a few colonies, while REL606 grew rapidly. ΔRNR does have more specific growth requirements than the wt REL606, so modification to the protocols used in this experiment could involve greater supplementation of media with dNTPs during the growth phase for producing electrocompetent cell, as using LB without dNTPs for both wt REL606 and ΔRNR may be resulting in a lower number of competent cells for ΔRNR. An improvement to the transformation protocol would also be to include a serial dilution to allow calculation of transformation efficiency, as this could be helpful for identifying issues with protocols chosen. Overall, the experimental results still have value as they indicate that the media, electrocompetent cell prep and transformation protocols can be effective, which indicates that PRE is possible.

The results generated both during analysis and working in the lab, while mixed in their success, are still meaningful for this area of research as the dataset gathered, successful transformation of an RNR bearing plasmid and iterative refinement of methods via trial and error provides potentially valuable information for researchers interested in researching RNRs, or for those interested in using ASR or a PRE in their research.

7. Appendices

Appendix 1: Sequences in Final Dataset

Sequences included in final dataset	Inteins Curated (Y/N)
M_B_Lac_lei_sp|Q59490|	N
M_B_Lac_del_1_sp|Q1G7W2|	N
M_B_Lac_del_2_sp|Q04CQ7|	N
M_B_Lac_aci_sp|Q5FMX8|	N
M_B_Lac_hel_sp|A8YW74|	N
M_B_Lac_gas_sp|Q041L3|	N
M_B_Lev_bre_sp|Q03PB4|	N
M_B_Lac_par_sp|Q035U1|	N
M_E_Dic_dis_sp|Q54CW7|	N
M_V_Myc_pha_1_sp|Q857H2|	N
M_V_Myc_pha_2_sp|Q05262|	N
M_V_Myc_pha_3_sp|O64240|	N
M_B_Nit_sp_sp|A6Q367|	N
D_B_Myc_tub_1_sp|P9WH76|	Y
D_B_Myc_tub_2_sp|P9WH77|	Y
D_A_The_aci_sp|Q9HI70|	Y
M_E_Eug_gra_sp|Q2PDF6|	N
D_A_Pyr_fur_sp|E7FHX6|	Y
D_B_Lep_int_1_sp|Q72S00|	N
D_B_Lep_int_2_sp|Q8F3P1|	N
D_B_Str_ave_sp|Q82KE2|	N
D_B_Bde_bac_sp|Q6MNP6|	N
D_B_Str_coe_sp|O69981|	N
D_B_Str_cla_sp|O54196|	N
D_B_Agr_fab_sp|Q8UEM4|	N
D_B_Bra_dia_sp|Q89MB9|	N
D_B_Rho_bal_sp|Q7UI46|	N
M_B_Com_all_tr|A0A1P8Q360|	N
M_B_Com_far_tr|A0A0H4LAC0|	N
M_B_Spo_sp_tr|A0A4Q4IDN3|	N
M_B_Per_cry_tr|A0A4P6YSH6|	N
M_B_Spo_inu_tr|A0A4Y1Z9K6|	N
M_B_Bre_spe_tr|A0A6M1U173|	N
M_B_Sei_pep_1_tr|A0A1M4SQY5|	N
M_B_Bac_spe_tr|A0A385NUP7|	N
M_B_Sei_pep_2_tr|A0A1M4VE42|	N
M_B_Psy_sp_tr|A0A1H9SSX6|	N
M_E_Aca_cas_tr|L8HFM0|	N
M_E_Glo_ult_tr|K3X6L5|	N
M_E_Sph_arc_tr|A0A0L0G1L6|	N
M_E_Phy_ram_tr|H3GD43|	N
D_B_Pse_aer_tr|Q9HT76|	N
D_B_Dok_kor_tr|A0A160DX99|	N
D_B_Miz_sed_tr|A0A0K8QMR0|	N
D_B_Lut_spe_tr|A0A1K1Q6Y3|	N
D_B_Fra_ter_tr|A0A1H6ZWD7|	N
D_B_Aer_soltr|A0A2Z6E7T3|	Y

D_A_The_kod_tr|Q5JJ43|	Y
D_A_Kor_cry_tr|B1L7R7|	N
D_B_Sul_azo_tr|C1DXK0|	N
D_B_Ano_fer_tr|A0A3Q9HRU8|	Y
D_B_Cam_spo_tr|A0A1M6ST55|	Y
D_A_Cal_sub_tr|E6N4A7|	Y
D_B_Car_hyd_tr|Q3AEA3|	Y
D_B_Mah_aus_tr|F4A2A5|	Y
M_E_All_mac_1_tr|A0A0L0T3F4|	N
M_E_All_mac_2_tr|A0A0L0T1V1|	N
M_A_Can_pro_tr|A0A5B9D7Q9|	N
M_E_Nae_gru_tr|D2VAC0|	N
M_B_Pse_spe_tr|A0A5Q2VVI5|	N
M_E_Phy_soj_tr|G4Z0A7|	N
M_B_Geo_sat_tr|A0A2Y9A4U1|	N
M_E_Pla_bra_tr|A0A0G4J8Q7|	N
M_E_Aph_ste_tr|A0A485L5Q2|	N
M_E_Sap_par_tr|A0A067BT24|	N
M_E_Ect_sil_tr|D7FJK6|	N
M_E_Cap_owc_tr|A0A0D2URK2|	N
M_E_Sap_dic_tr|T0S0C6|	N
M_E_Aph_inv_tr|A0A024U4L4|	N
M_E_Sty_lem_tr|A0A078ABB6|	N
M_E_Gon_pro_2_tr|A0A138ZXL1|	N
D_B_Gem_obs_|QEG297511|	N

Appendix 2: Alignment files

Cobalt_Alignment.aln

Mafft_Alignment.txt

Muscle_alignment.txt

Muscle_class_I_Inteins
_removed_alignment.t

Initial_ASR_output.txt

Final_ASR_output.txt

Final_ASR_alignment.txt

Appendix 3: Log files

Muscle_Model_Finder_log.txt.log

Initial_ASR_log.txt.log

Final_ASR_log.log

Appendix 4: Tree Files

MUSCLE_alignment.txt.iqtree

MUSCLE_alignment.txt.contree

MUSCLE_alignment.txt.treefile

Initial_ASR.txt.treefile

Initial_ASR.txt.iqtree

Initial_ASR.txt.contree

Final_ASR.txt.iqtree

FInal_ASR.txt.contree

Final_ASR.txt.treefile

8. References:

Arras, S., Sibaeva, N., Catchpole, R., Horinouchi, N., Si, D., Rickerby, A., . . . Poole, A. (2023). Characterisation of an Escherichia coli line that completely lacks ribonucleotide reduction yields insights into the evolution of parasitism and endosymbiosise. *Life*.

Bartel, D. P., & Unrau, P. J. (1999). Constructing an RNA world. *Trends in cell biology*, 9-13.

Blount, Z. D., Lenski, R. E., & Losos, J. B. (2018). Contingency and determinism in evolution: Replaying life's tape. *Science*.

Bromham, L. (2019). Six Impossible Things before Breakfast: Assumptions, Models, and Belief in Molecular Dating. *Trends in Ecology & Evolution*, 474-486.

Burnim, A., Spence, M., Xu, D., Jackson, C., & N, A. (2022). Comprehensive phylogenetic analysis of the ribonucleotide reductase family reveals an ancestral clade. *eLife*.

Clifton, B. E., Kaczmarski, J. A., Carr, P. D., Gerth, M. L., Tokuriki, N., & Jackson, C. J. (2018). Evolution of cyclohexadienyl dehydratase from an ancestral solute-binding protein. *Nat Chem Biol*, 542–547.

Crona, M., Avesson, L., Sahlin, M., Lundin, D., Hinas, A., Klose, R., . . . Sjöberg, B. M. (2013). A rare combination of ribonucleotide reductases in the social amoeba Dictyostelium discoideum. . *The Journal of biological chemistry*, 8198-8208.

Crona, M., Hofer, A., Astorga-Wells, J., Sjöberg, B. M., & Tholander, F. (2015). Biochemical characterization of the split class II ribonucleotide reductase from Pseudomonas aeruginosa. *PloS one*.

DeepMind and EMBL-EBI. (n.d.). *AlphaFold Protein Structure Database*. Retrieved from alphafold.com: http://alphafold.com/

Doudna, J. A., & Cech, T. R. (2002). The chemical repertoire of natural ribozymes. *Nature*, 222-228.

Hanson-Smith, V., Kolaczkowski, B., & Thornton, J. W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular biology and evolution*, 1988-1999.

Kapli, P., Yang, Z., & Telford, M. (2020). Phylogenetic tree building in the genomic age. *Nat Rev Genet* , 428–444.

Lundin, D., Berggren, G., Logan, D., & Sjöberg, B.-M. (2015). The Origin and Evolution of Ribonucleotide Reduction. *Life*, 604-636.

Lundin, D., Torrents, E., Poole, A. M., & Sjöberg, B. M. (2009). RNRdb, a curated database of the universal enzyme family ribonucleotide reductase, reveals a high level of misannotation in sequences deposited to Genbank. *BMC genomics*, 1-8.

NIH. (2023). *Epistasis*. Retrieved from genome.gov: https://www.genome.gov/genetics-glossary/Epistasis

Poole, A. M. (2011). On alternative biological scenarios for the evolutionary transitions to DNA and biological protein synthesis. *Origins of Life: the primal self-organization*, 209-223.

Poole, A. M., Jeffares, D. C., & Penny, D. (1998). The path from the RNA world. Journal of molecular evolution. *Journal of molecular evolution*, 1-17.

Poole, A. M., Logan, D. T., & Sjöberg, B. M. (2002). The evolution of the ribonucleotide reductases: much ado about oxygen. . *Journal of molecular evolution*, 180-196.

Poole, A., Penny, D., & Sjöberg, B. M. (2000). Methyl-RNA: an evolutionary bridge between RNA and DNA? *Chemistry & Biology*, 207-216.

Randall, R. R. (2016). An experimental phylogeny to benchmark ancestral sequence reconstruction. *Nat Commun*.

Reichard, P. (1997). The evolution of ribonucleotide reduction. *Trends Biochem Sci*, 81-85.

Saito, H. (2022). The RNA world 'hypothesis'. *Nat Rev Mol Cell Biol* , 582 .

Salverda, M. L., Dellus, E., Gorter, F. A., Debets, A. J., Van Der Oost, J., Hoekstra, R. F., & ... & de Visser, J. A. (2011). Initial mutations direct alternative pathways of protein evolution. *PLoS genetics*.

Sintchak, M. D., Arjara, G., Kellogg, B. A., Stubbe, J., & Drennan, C. L. (2002). The crystal structure of class II ribonucleotide reductase reveals how an allosterically regulated monomer mimics a dimer. *Nature structural biology*, 293-300.

Stubbe, J. (2000). Ribonucleotide reductases: the link between an RNA and a DNA world? *Current Opinion in Structural Biology*, 731-736.

Stubbe, J., & van Der Donk, W. A. (1998). Protein Radicals in Enzyme Catalysis. *Chem. Rev.*, 2661-2662.

Torrents, E., Aloy, P., Gibert, I., & Rodríguez-Trelles, F. (2002). Ribonucleotide reductases: divergent evolution of an ancient enzyme. *Journal of molecular evolution*, 138-152.