ABSTRACT

It is important to understand ancient evolutionary relationships in order to understand cell and genome evolution. Shared derived genomic characters such as horizontal gene transfers (HGT), intron insertions, insertions and deletions within open reading frames, and gene fusion events can help as they can be useful for polarising ancient phylogenetic relationships. Here, I report a newly identified tri-fusion of DHNA-HPPK-DHPS protein domains which encode the consecutive folate biosynthesis enzymes: dihydroneopterin aldolase, 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase and dihydropteroate synthase. Using a combination of comparative genomics and phylogenetics to identify placement of HGTs, gene loss, and gene fission events it was possible to determine several key events in the evolution of the folate biosynthesis genes within the eukaryotes. The HPPK-DHPS bi-fusion was identified as a possible synapomorphy for the eukaryotes with several subsequent losses in the eukaryotic groups such as the Rhizaria and Excavata (depending on root placement). Furthermore, the DHNA-HPPK-DHPS tri-fusion was identified as a possible synapomorphy for the unikonts (i.e. Metazoa, Fungi and Amoebozoa). Several reversions of this gene fusion character were identified within the unikonts along with a putative DHNA duplication event which appears to be a synapomorphy for the Ascomycota and Basidiomycota (Dikarya). Two putative inter-domain HGTs were also identified: between the Metazoa and the crenarchaeaon *Sulfolobus acidocaldarius* and an additional transfer into the cercozoan *Paulinella chromatophora* from a bacterium. In conclusion, the DHNA-HPPK-DHPS tri-fusion is a putative synapomorphy for the unikonts, however, the loss and transfer of this character makes it difficult to evaluate the reliability of this synpomorphy.

TABLE OF CONTENTS

TABLE OF FIGURES

ABBREVIATIONS

ACT	Aspartate carbamoyltransferase
AIC	Akaike's information criterion
BCM	Baylor college of medicine
BIC	Bayesian information criterion
BLASTp	Protein basic local aligment search tool
CDD	Conserved domain database
cDNA	Complementary deoxyribonucleic acid
CEEM	Center for eukaryotic evolutionary microbiology
CPSII	Carbamoyl phosphate synthase II
DHFR	Dihydrofolate reductase
DHFS	dihydrofolate synthetase
DHNA	Dihydroneopterin aldolase
DHO	Dihydroorotase
DHPS	Dihydropteroate synthase
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
dUMP	deoxyuridine monophosphate
EC	Enzyme commission (number)
FOLB	Dihydroneopterin aldolase
FOLBKP	Dihydroneopterin aldolase - 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase - dihydropteroate synthase
FOLK	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase

FOLP	Dihydropteroate synthase
FPGS	Folylpolyglutamate synthase
HGT	Horizontal gene transfer
HMM	Hidden markov model
HPPK	6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase
JGI	Joint genome institute
LBA	Long branch attraction
LG	Le gascuel
MSA	Multiple sequence alignment
MUSCLE	Multiple sequence comparison by log-expectation
NCBI	National center for biotechnology information
ORF	Open reading frame
PCR	Polymerase chain reaction
PTSP	6-pyruvoyltetrahydropterin synthase
RGC	Rare genomic changes
RNA	Ribonucleic acid
RPS-BLAST	Reverse posistion-specific basic local aligment search tool
SDC	Shared derived character
SSU-rRNA	Small subunit ribosomal ribonucleic acid
TBE	Tris, borate, and ethylenediaminetetraacetic acid
TS	Thymidylate synthase

INTRODUCTION

TREE OF LIFE

The basis of modern phylogenetic research lies with Wili Hennig's cladistics which introduced the idea of grouping organisms into a series of monophyletic nested groups or "clades" containing all the descendents of a common ancestor (Hennig, 1966). This original research was based upon morphological characters and enabled the construction of rooted phylogenies. However, more recent research has moved to encompass molecular phylogenetic methods and thus enabled the first tree of life based on molecular data to be reconstructed (e.g. Woese's ribosomal RNA (SSU rRNA) phylogenies (Woese, 1996)). With the innovation of wide-scale genomic sequencing projects, increases of computational power, and development of new tools multi-gene molecular phylogenies have begun to emerge. However, even with these sophisticated methodologies vast numbers of genes (100s) are required to robustly resolve relatively recent evolutionary relationships such as the Archaeplastida (land plants, green algae, red algae and glaucophytes) (Rodriguez-Ezpeleta et al., 2005). This had led to some researchers to suggest that there are not enough genes free from HGT to resolve a meaningful phylogeny of the complete tree of life (Bapteste et al., 2004).

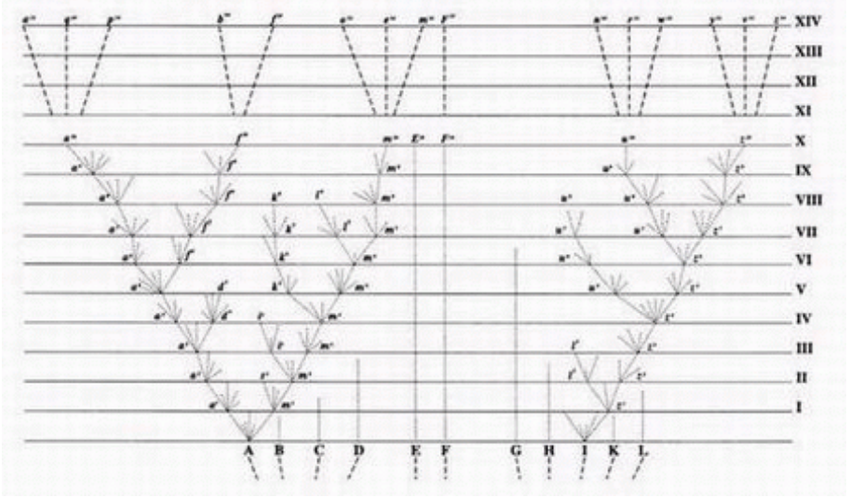

FIGURE 1 – DARWIN'S BRANCHING TREE OF EXISTENCE (DARWIN, 1859)

SHARED DERIVED CHARACTERS

Molecular phylogenies however, still had one disadvantage over morphological phylogenies in that they were based principally upon gross sequence divergence and thus did not allow trait polarisation. Trait polarisation is the idea that you can place a root between two clades based on the presence of a trait in one group and the absence in another and therefore requires discrete shared derived characters (SDCs) (Richards, 2005, Rokas and Holland, 2000). These were traditionally morphological characters such as Gaffney's testudinid skull architectures (Gaffney, 1977), however, derived genomic characters (or rare genomic changes – RGCs) such as gene fusions or HGTs can also be used as effective SDCs. These molecular SDCs have the advantages of both morphological phylogenies (through trait polarisation), and molecular phylogenies (through the objective analysis of character divergence). There has been considerable success elucidating eukaryotic phylogenies using genomic SDCs (Richards et al., 2003, Richards, 2005, Richards et al., 2006, Cavalier-Smith, 2002, Jenkins and Fuerst, 2001, Rogozin et al., 2009). This is seen in such examples as the use of an intron insertion in the homeobox *engrailed* gene of the Diptera and Lepidoptera to resolve the debate (where data was equivocal and the 18s SSU rDNA molecular studies were inconsistent) over the placement of the insect order Strepsiptera within the homometabolous insects by inferring that the Strepsiptera is not the sister group of the Diptera (Rokas and Holland, 2000).

GENE FUSIONS

A gene fusion occurs when two or more open reading frames (ORFs) become a single ORF (Doolittle, 1995). They occur by three main molecular mechanisms: chromosomal translocation, interstitial deletion, and chromosomal inversion (Leonard, 2010). Gene fusions most often occur in genetic contexts where there are functionally related genes proximal to one another such as within operons (Conant and Wagner, 2005, Kummerfeld and Teichmann, 2005), multi-domain genes (Teichmann and Mitchison, 1999), and genes that encode separate components of multimeric protein complexes (Marcotte *et al.*, 1999). One of the more useful aspects of using gene fusions as SDCs is that due to their relative rarity as genomic events (Nakamura et al., 2007) if there is conservation in the

arrangement of domains within a fusion gene it suggests that the fusion gene is likely to be a product of single fusion event (Bashton and Chothia, 2002).

EUKARYOTIC PHYLOGENY

The current consensus state of the eukaryotic phylogeny is the division of the eukaryotic tree into six major groups (with a few cryptic *incertae sedis*[i] species): the Opisthokonta (e.g. the Fungi and Metazoa), Amoebozoa, Archaeplastida, Rhizaria, Chromalveolata and Excavata (Brinkmann and Philippe, 2007, Simpson and Roger, 2004, Stechmann and Cavalier-Smith, 2003, Arisue et al., 2005, Rodriguez-Ezpeleta et al., 2007, Rodriguez-Ezpeleta et al., 2005, Archibald et al., 2003, Richards et al., 2006, Hampl et al., 2009). Stechmann and Cavalier-smith have proposed that these six groups form into two major divisions: the unikonts (containing the Opisthokonta and Amoebozoa) and the bikonts containing most other eukaryotic taxa with the root placed between these supergroups (Stechmann and Cavalier-Smith, 2003). Much evidence has been cited for elements of this bifurcation:

- Dihydrofolate reductase-thymidylate synthase (DHFR-TS) gene fusion (Stechmann and Cavalier-Smith, 2002) only within the bikonts.
- Carbamoyl phosphate synthase II, dihydroorotase, aspartate carbamoyltransferase (CPSII-DHO-ACT) triple pyrimidine synthesis gene fusion found only within the unikonts (Nara et al., 2000). However, the recent discovery of the CPSII-DHO-ACT tri-fusion in the bikont red alga *Cyanidioschyzon merolae* suggests that the fusion may have occurred before the unikont-bikont bifurcation and subsequently lost in many 'bikont' taxa (Nozaki et al., 2005, Matsuzaki et al., 2004).
- Distribution of specific myosin domain families and gene architectures (Richards and Cavalier-Smith, 2005).

[i] Of uncertain placement

- Rogozin *et al*'s genome-wide analysis of RGCs associated with sections of conserved amino acids have demonstrated the monophyly of the unikonts but not the bikonts (with the authors hypothesising a root between the photosynthetic and non-photosynthetic eukaryotes) (Rogozin *et al*., 2009).

Ultimately, identified SDC datasets relevant to the ancient relationships within the eukaryotes are slightly inconsistent and thus more data is required to test the relative branching order of the major eukaryotic groups.

FOLATE BIOSYNTHESIS PATHWAY FUSION

Tetrahydrofolate (referred to as folate) is a tripartite molecule comprising pterin, p-aminobenzoate and glutamate moieties which acts as an essential metabolite in single carbon transfer reactions such as those involved in the biosynthesis of a range of important molecules: the purine adenine base, the pyrmidine thymidine base, methionine and histidine amino acids, and formyl-tRNA (Brown, 1971). Folate is vital to all 3 domains of life. Most organisms synthesise folate via a biosynthetic pathway, however, some higher eukaryotes such as the vertebrates have lost this pathway and instead rely on acquiring folate from the environment through specialised transporters such as the proton-coupled folate transporter (Zhao and Goldman, 2007). There is considerable interest in the mechanisms of the folate biosynthesis pathway as its loss in the higher animals and its high degree of conservation makes it a potent anti-pathogen drug target (Lawrence *et al.*, 2005).

In the folate biosynthetic pathway there are 3 key subsequent synthesis enzymes: dihydroneopterin aldolase (DHNA, FOLB, EC 4.1.2.25) which catalyses the conversion of dihydroneopterin to 6-hydroxymethyl-7,8-dihydropterin; 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase (HPPK, FOLK, EC2.7.6.3) which catalyses the conversion of 6-hydroxymethyl-7,8-dihydropterin to 6-hydroxymethyl-7,8-dihydropterin pyrophosphate; and dihydropteroate synthase (DHPS, FOLP, EC 2.5.1.15) which catalyses the condensation of 6-hydroxymethyl-7,8-dihydropterin pyrophosphate with p-aminobenzoate (PAB) to produce 7,8-dihydropteroate (Lopez and Lacks, 1993, Lawrence et al., 2005) (as shown in Figure 2).

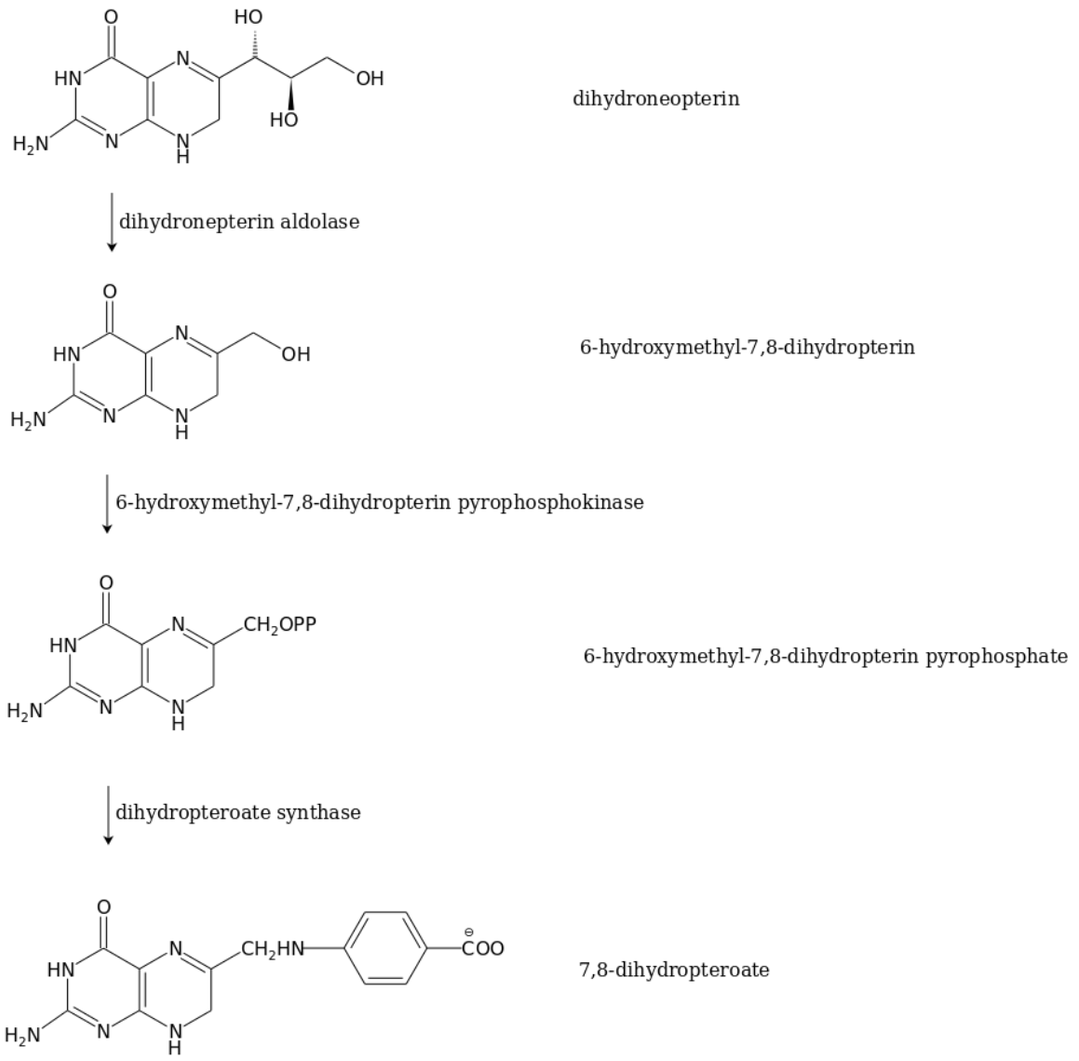

FIGURE 2 – SECTION OF THE FOLATE SYNTHESIS PATHWAY (XIAO ET AL., 2001, PRIBAT ET AL., 2009)

The product of these 3 enzyme catalysed reactions subsequently undergoes glutamate attachment by dihydrofolate synthetase (DHFS, EC 6.3.2.12) to form dihydrofolate which is reduced by dihydrofolate reductase (DHFR, EC 1.5.1.3) to form tetrahydrofolate (de Crecy-Lagard et al., 2007). Derivatives of tetrahydrofolate such as 5,10-methylenetetrahydrofolate are then used as carbon donors in the reactions such as the thymidylate synthase (TS, EC 2.1.1.45) catalysed reduction of deoxyuridine monophosphate (dUMP) to deoxythymidine monophosphate (Zhang et al., 2010). Enzymes involved in this pathway have been found to exist both as the individual mono-functional enzymes and as a variety of fused forms, where the transcribed and translated proteins are composed of

multiple domains each individually homologous to the separate genes. These include bi-functional DHNA-HPPK, HPPK-DHPS, and DHFR-TS fusions and tri-functional DHNA-HPPK-DHPS double fusions (Lawrence et al., 2005).

It has been suggested that fusion of enzyme domains may represent an adaptation to substrate channelling. Substrate channelling involves an intermediate being directly transferred from one domain's active site to next without free diffusion into the solution (Liang and Anderson, 1998) and has been detected in the case of the DHFR-TS fusion (Meek et al., 1985). Channelling has also been speculated for the DHNA, HPPK, and DHPS domain fusions. However, investigation of the kinetics of the HPPK-DHPS bi-functional fusion found in the Archaeplastida has discovered that the 6-hydroxymethyl-7,8-dihydropterin intermediate is able to equilibrate with the external media implying no substrate channelling (Mouillon et al., 2002). However, this doesn't necessarily discount a limited diffusion pathway channelling from taking place in the bi-fusion (Mouillon et al., 2002) or that tri-fusion form conducts substrate channelling. While the channelling hypothesis is very appealing for these fused consecutive biosynthesis enzymes more work is needed to determine the true extent of the kinetic associations between these three domains including the identification of how their active sites interact across the wider protein structure and where the pathway is localised relative to the plastid organelle in Archaeplastida and Chromalveolata (Basset et al., 2004) or the cytosol in 'unikonts'.

The DHNA-HPPK-DHPS triple domain fusion is potentially a useful and phylogenetically informative SDC as it is most likely the product of two distinct fusion events. This means it is the product of two rare genomic changes and therefore the probability of multiple incidences of the triple domain fusion is relatively low compared to single RGC fusion characters such as DHFR-TS fusion (Stechmann and Cavalier-Smith, 2002) and is therefore less likely to independently arise in different taxa. The objective of this study is to investigate the evolutionary history of the DHNA, HPPK, DHPS elements of the folate biosynthetic pathway and assess the suitability of this fusion character as a phylogenetically informative SDC. To do this I will sample additional folate biosynthesis genes from

transcriptome sequencing of the amoebozoan *Acanathamoeba castellanii* Neff (Neff, 1957) and through this identify the distribution of the folate biosynthesis fusion gene characters among the Amoebozoa. I will then use comparative genomics and phylogenetics to test the hypothesis that the fusion domain states are monophyletic. Finally, I will use these data to identify whether the DHNA-HPPK-DHPS tri-fusion is a reliable synapomorphy for the unikonts.

MATERIALS AND METHODS

CONFIRMATION OF THE GENE FUSION FROM TRANSCRIPTOME SEQUENCING

To investigate the distribution of the folate biosynthesis gene fusions relative to the unikont hypothesis it is important to identify the presence of this character within the Amoebozoa. This is problematic because of the paucity of high coverage assembled amoebozoan genome sequences (10 sequencing projects representing only 2 groups in GenBank: Mycetozoa (e.g. several *Dictyostelium* species and *Polysphondylium pallidum)* and Archamoebae (e.g. several *Entamoeba* species). Furthermore, the entire folate biosynthesis pathway is absent from the three *Entamoeba* genomes sampled as corroborated by Loftus *et al* (Loftus et al., 2005). However, I had identified a third amoebozoan taxon with a putative DHNA-HPPK-DHPS tri-gene fusion in the *Acanthamoeba castellani* low coverage genome assembly (publicly available at Baylor College of Medicine (BCM) genome data facility – http://www.hgsc.bcm.tmc.edu/microbial-detail.xsp?project_id=163) (Loftus, 2008). It was therefore important to test if these three protein domains are encoded as a single transcript. This was necessary because *Acanthamoeba* has an intron-rich gene architecture containing many intron-like sequences composed of tandem and triplicate repeats (Anderson et al., 2005). This characteristic makes it very difficult to predict the ORF structure of these gene/s and therefore making it difficult to discount the possibility that this was a gene cluster (Keller et al., 2005) and not a true gene fusion.

ACANTHAMOEBA CDNA PRODUCTION

The *Acanthamoeba castellanii* complementary DNA (cDNA) was provided by Dr Fiona Henriquez[ii]. *Acanthamoeba* was grown axenically in a modified M11 defined media (Shukla et al., 1990) (see Figure 13 for full media composition) and without folates in order to encourage the transcription of the folate biosynthesis pathway. Cells were collected and suspended in 1ml of TRIzol reagent (Invitrogen) and RNA extracted using the single-step acid

[ii] University of The West of Scotland

guanidinium thiocyanate-phenol-chloroform protocol as described in (Chomczynski and Sacchi, 1987). The cDNA was then synthesised using the AffinityScript kit with random hexamers (Stratagene).

PCR AMPLIFICATION OF THE FOLATE BIOSYNTHESIS GENES USING MULTIPLE PRIMERS IN ORDER TO IDENTIFY GENE DOMAIN STRUCTURE

In order to test whether the folate biosynthesis gene in *Acanthamoeba* was arranged as a triple-domain gene, primers were designed to target different domain sections and used in combination to confirm the protein domain architecture of the transcribed gene in the cDNA. These primers were designed using the preliminary *Acanthamoeba castellani* sequencing data from BCM and the PerlPrimer tool (Marshall, 2004). PCR amplification was conducted using Master Mix (Promega, containing 3mM $MgCl_2$, 400µM of each dNTP, and 50 U/ml of Taq DNA polymerase) to create a 25 µl PCR reaction mix (12.5 µl of Master Mix, 1 µl of each primer (10 $pM\mu l^{-1}$), 9.5 µl of Milli-Q pure water (Millipore), and 1 µl of template DNA) (Jones, 2007). The provided cDNA was initially diluted to ~10ng/µl using spectrophotometery (NanoDrop ND-1000) but owing to poor results this was increased to a concentration of ~100ng/µl. An MJ Mini Personal Thermal Cycler (Bio-Rad Laboratories) was used with a program of a 95°C 5 minute initialisation temperature, 35 cycles of a 95°C 1 minute denaturation step, a range of temperatures between 55-65°C (55.8°C, 59°C, 63°C, 65°C) were held for 1 minute to anneal the primers (as empirically determined optimum annealing temperatures had not been assessed), and a 72 °C elongation step was held for variable time depending on the primer pair used and the expected size of the amplified fragment according to the Taq DNA polymerase specification of 1kb per minute (for instance 40 seconds was used for the reaction of FOLB18-F and FOLK620-R). Finally, a 72°C 5 minute final elongation period of was held followed by reduction of the reaction to a holding temperature of 15°C.

The samples were then tested for successful PCR via gel electrophoresis (1x TBE buffer with a 1% agarose gel). The bands sizes were assessed using HyperLadder I (BioLine) ladder and those of the predicted size (from the primers used) and of adequate concentration (assessed via comparative fluorescence level with the ladder) were then

excised on a UV transilluminator (AutoChemiSystem) using a sterile scalpel. Seven sets of excised bands were purified using the Wizard SV Gel and PCR Clean-Up kit (Promega) and then the DNA concentration assessed using spectrophotometery (NanoDrop ND-1000). Two sets of bands were rejected at this stage due to failure of the purification protocol.

Primer	Primer direction	Sequence (5'-3')	Successful?
FOLB8	Forward	CAAGGATCTGATGGTGCAGGCC	No
FOLB18	Forward	GGATCTGATGGTGCAGGCCATCC	Yes (paired with FOLP1662)
FOLB39	Forward	CCTGGGCGTCAACAAGGAGGA	No
FOLB98	Forward	TCTTCCACGACATCAAGCAGGCC	No
FOLB168	Forward	CAAGGCCGTCGTGGCCTACA	Yes (paired with FOLK620, FOLP1662)
FOLB247	Forward	TGCGTTCAGTTCGGCGCCGCC	No
FOLK617	Reverse	GGCCGAGGTGCTGATGATGTCGC	No
FOLK618	Forward	CGACATCATCAGCACCTCGGCCC	No
FOLK618	Reverse	GCTGTAGTAGTCGTGGAGCCGGG	No
FOLK620	Reverse	AGAGGGCCGAGGTGCTGATGATGT	Yes (paired with FOLB168)
FOLK623	Forward	TCATCAGCACCTCGGCCCTCTACC	Yes (paired with FOLP1151)
FOLP1151	Reverse	GGGCGTCGCGTTGATGATGC	Yes (paired with FOLK623)
FOLP1662	Reverse	CCGCCAGCTGAGCCGAATCG	Yes (paired with FOLB18, FOLB168)

FIGURE 3 – FOLBKP PRIMERS USED TO INVESTIGATE THE FOLATE BIOSYNTHESIS GENE ARCHITECTURE.

CLONING OF CDNA PCR PRODUCTS

The five sets of successfully purified bands from the gel were cloned using TA-cloning (Holton and Graham, 1991) (PCR StrataClone Cloning Kit (Agilent Technologies) following the manufacturer's specifications). Blue-white screening (where the vector lacZ' cassette complements the lacZΔM15 mutation in the genome of the competent cells in successfully transformed cells) on 2% X-gal Ampicillin LB-agar was conducted to assess the success of vector integration into the competent *Escherichia coli* cells (Agilent Technologies, 2006). The Wizard Plus SV Miniprep DNA Purification System (Promega) was then used to extract and purify the cloned vectors from the StrataClone competent *E. coli* colonies. Five colonies were selected from each plate to produce 25 samples. These purified preparations were tested using spectrophotometery (NanoDrop ND-1000) to assess the success of purification. At

this point three samples were rejected due to low DNA concentration and apparent rubber contamination from the broth culture bottles.

SEQUENCING OF PLASMID INSERTS

The 22 successfully purified samples had their insert size verified and were prepared for sequencing via PCR with M13/pUC primers. M13/pUC primers are commonly used sequencing primers that bind to elements of the lacZ on the purified vector. The protocol and reaction mixture used for this was the same as that used for the previous PCR reactions with the exception of a fixed 50 °C annealing step, 40 second 72 °C elongation step, 20 cycles, a final elongation step of 10 minutes (at 72 °C) and the use of single transformed white colony sample as the DNA template. All 22 samples were of correct size and adequate concentrations for sequencing and 16 were subsequently sequenced using M13F primers with a further six duplicate plasmid samples were sequenced in the reverse orientation using the M13R. Sequencing was undertaken externally via Sanger sequencing at the Cogenics Beckman-Coulter sequencing service.

CONTIG ASSEMBLY OF DNA SEQUENCES

The flanking vector sequences were removed; the sequences trimmed to areas of high chromatograph quality and ambiguously defined bases corrected. The overlapping contigs were then assembled using the GeneCodes Sequencher™ version 4.10.1 program (http://www.genecodes.com/) (GeneCodes, 2010). Four of sequences were rejected as being of low quality. By assembling the 18 contigs a high confidence consensus sequence for the cDNA was produced and the reading frame determined.

PHYLOGENETIC ASSESSMENT OF MONOPHYLY IN THE GENE FUSIONS

TAXONOMIC SAMPLING

GenBank (Benson et al., 2010), the Joint Genome Institute (http://genome.jgi-psf.org/) (JGI, 2010), TBestDB (O'Brien *et al.*, 2007), and the Broad Institute (http://www.broadinstitute.org/) (Broad, 2010) genome databases

were sampled (as of August 2010) using the Protein Basic Local Alignment Search Tool (BLASTp) (Altschul et al., 1990) and seed sequences for each of the three separate folate biosynthesis domains from *Bacillus cereus* (DHNA- NP 829975.1, HPPK- ZP 03233543.1, DHPS- ZP 07056868.1). Care was taken to survey the major eukaryotic, archaeal and bacterial groups.

ALIGNMENT

The amino acid sequences gathered for each domain were run through the REFGEN script (Leonard et al., 2009) to transform the JGI and GenBank sequence labels into labels readable by phylogenetic analysis software. The sequences acquired from TBestDB and Broad Institute were manually relabelled. The MUltiple Sequence Comparison by Log-Expectation (MUSCLE) program (v3.8.31) (Edgar, 2004) was utilised to produce a multiple sequence alignment (MSA) for each domain (DHNA, HPPK, and DHPS). MUSCLE was chosen because of its increased or equal accuracy (in the BAliBASE, SABmark, SMART and PREFAB benchmark tests) and faster run-time compared to other popular MSA algorithms (e.g. CLUSTALW, MAFFT and T-COFFEE) (Edgar, 2004).

MASKING

All alignments were inspected for misalignments errors and then manually corrected in the SeaView (version 4.2.4) (Gouy et al., 2010) application. Sites that were present in all or most taxa were selected (masked) and sequences that caused an unacceptable loss of putatively informative sites (due to the sequence non-alignment or not masking well) were removed.

SUBSTITUTION MODEL SELECTION

As I used maximum-likelihood methods for phylogenetic analysis it was necessary to select a model of amino-acid substitution for evolutionary changes across the MSAs. Different substitution model selection demonstrably leads to very different phylogenies being reproduced, therefore an objective and statistically supported method of model selection was vital in the reconstruction of phylogenies (Keane et al., 2006). All data matrices were analysed

using the MODELGENERATOR (version 0.85) (Keane et al., 2006) program. This program tests the data matrices against 96 different substitution models (12 models and the 8 combinations of: invariant sites (I), gamma distribution with 8 discrete rate categories (Γ), and observed amino acid frequencies (F) and assesses their appropriateness for selection via a pair of penalised-model selection criteria (Kuha, 2004): Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Keane et al., 2006). See Figure 14 in the appendix for the full list of PhyML parameters predicted by MODELGENERATOR for each phylogeny.

Long branch attraction (LBA) (Felsenstein, 1978) artefacts are a phenomenon in phylogenetic reconstruction in which the long branches of fast evolving taxa falsely infer other long branches as being closely related. This is due to the increased chance of the same sites arising in two sets of rapidly evolving taxa (Philippe *et al*., 2005). The effects of LBA were minimised by: increased taxonomic sampling in order to 'fill in the gaps' and thus break up the long branches, utilisation of the complex parameter rich LG substitution model, use of an α-parameter to identify the gamma distribution (which allows weighting of characters depending on where they fit on the probability curve), and in some cases removal of long branching prokaryote sequences from the alignment. Due to the computational problems of calculating a continuous gamma distribution an 8 category discrete model was used as an approximation (Yang, 1994). Invariant sites were also taken into account in several of the phylogenetic reconstructions (in accordance with MODELGENERATOR recommendations). Invariant sites are those alignment positions which remain constant across the alignment therefore if unaccounted for will lead to under-correction for site changes in the phylogenetic reconstruction (Leonard, 2010).

TREE RECONSTRUCTION

The phylogenies were calculated using the PhyML (version 3.0) (Guindon and Gascuel, 2003) program running through the TITAN cluster of the University of Oslo's BioPortal (Kumar *et al.*, 2009). PhyML is a maximum-likelihood-based tree constructing algorithm utilising a hill-climbing algorithm to minimise the number of required

iterations between the initial neighbour-joining tree, serial maximum-likelihood re-evaluation steps, and the final tree produced (Guindon and Gascuel, 2003). This approach allows for gamma distribution and invariant site rate heterogeneity parameters and is generally less susceptible to LBA than some other methods e.g. maximum parsimony (Bergsten, 2005). All topologies were evaluated with 1000 bootstrap replicates. Bootstraps are a reliability test in which masked alignment sites are randomly re-sampled in order to generate a 'pseudo-replicate' dataset and the tree recalculated for each replicate. If a branch is maintained in the a replicate tree it is given a score of 1 and if not 0, this process was repeated for 1000 replicates for each analysis (Felsenstein, 1985).

The program TREENAMER was then run on the resulting tree files (Leonard et al., 2009) in order to restore the correct taxa names from the REFGEN tags used during phylogenetic processing. Duplicate taxa entries and closely related taxa (usually in the same genus) which formed a clade were removed from the collected sequences. The entire phylogenetic process pipeline was then re-run with these modified data sets to produce new sets of phylogenies with redundancies and long branches removed.

CONCATENATION OF FOLATE BIOSYNTHESIS DOMAINS

To further assess support for the relative placement of different gene fusion characters, I conducted a series of concatenated analyses. Concatenated alignment analyses can increase the availability of the sequence characters and therefore improve phylogenetic resolution and further test key branching relationships (Gadagkar et al., 2005). The masked sites for each domain were concatenated into four new masked alignments (DHNA-HPPK, HPPK-DHPS, DHNA-DHPS, and DHNA-HPPK-DHPS) based on the overlapping presence of taxa in each individual domain alignment. The same MODELGENERATOR and PhyML procedure outlined above was performed to produce concatenated phylogenies. The concatenation process was aided via the creation of two custom built Perl scripts: ConcatenatedTreeList (see Figure 18 for source code) which produces lists of taxa found in each combination of

domains and Renamer (see Figure 19 for source code) which renamed one set of masked sequences to match the name of the masked sequences from the same taxa in the other dataset to allow concatenation in SeaView.

LABELLING OF PHYLOGENETIC TREES

A profile hidden Markov model (Krogh et al., 1994) (profile HMMs) was used against the PFAM hidden Markov model training dataset via the HMMER program (Durbin, 2010) to predict the conserved domains found within the collected sequences. These predictions were then plotted onto the phylogenies. A script originally created by Bill Wickstead (Wickstead et al., 2010) and modified by Guy Leonard (Leonard, 2010) was used to automate this process. The DHNA domain is part of a wider domain family called the T-Fold domain superfamily and as the T-Fold HMM produced higher identity scores in the domain comparison analysis this HMM was used for consistency in place of the DHNA HMM. Bootstrapped phylogenies were viewed and prepared using the NJPlot (Perriere and Gouy, 1996) and Dendroscope applications (Huson et al., 2007).

RESULTS

I aim to determine that *Acanathamoeba castellanii* expresses DHNA, HPPK, and DHPS folate biosynthesis domains as a single fused transcript, that the fusion domain states are monophyletic and that the DHNA-HPPK-DHPS tri-fusion is a synapomorphy for the unikonts. Through this I intended to present an evolutionary scenario for the evolution of these biosynthetic gene fusions and evaluate the three domain gene fusion as a reliable phylogenetically informative SDC.

GENE FUSION CONFIRMATION

Acanthamoeba castellanii represents a distant relative of the mycetozoan slime moulds within the Amoebozoa. Currently, the DHNA-HPPK-DHPS three domain fusion has only been found in the mycetozoan slime moulds among the Amoebozoa. Therefore, identifying the presence of this gene fusion in *Acanthamoeba* represents an important piece of data demonstrating the presence of the fusion in two diverse branches of the Amoebozoa and thus confirming the character across the Amoebozoa as a whole. Figure 4 demonstrates multiple overlapping reads in forward and reverse (green and red respectively) confirming the cDNA sequence amplified (see Figure 17 for raw sequencing data). The results demonstrate one large ORF but do not identify a stop codon, which is consistent with the location of the PCR primers which were proximal to the end of the DHPS domain (see Figure 15 for DNA sequence and Figure 16 for translated protein sequence in the appendices). Together these data demonstrate that the folate biosynthesis domains are transcribed as one long gene transcript encoding all three protein domains.

CONTIG ASSEMBLY

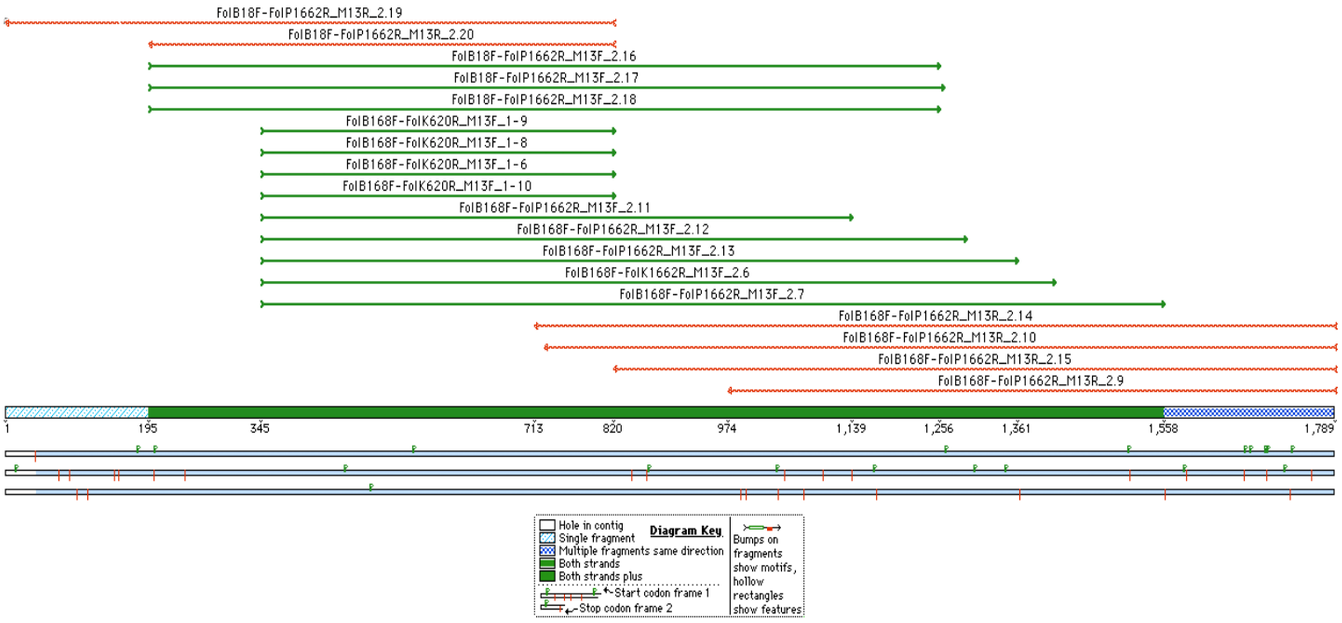

FIGURE 4 – OVERLAPPING TRIMMED SEQUENCING READS FROM THE PUTATIVE *ACANTHAMOEBA* TRI-FUSION CDNA ASSEMBLED IN SEQUENCHER. THIS RESULT DEMONSTRATES ONE LONG 1,789 BP OPEN READING FRAME SAMPLED FROM THE ACANTHAMOEBA TRANSCRIPTOME. LOCATION OF THE DOMAIN-SPECIFIC PRIMERS STRONGLY SUGGESTS THAT ALL THREE FOLATE BIOSYNTHESIS DOMAINS ARE PRESENT IN THE TRANSCRIPT.

PROTEIN SEQUENCE ANALYSIS

When the cDNA sequencing assembled contig is translated in the +1 reading frame and analysed using the RPS-BLAST against the CDD (Marchler-Bauer et al., 2009) (Figure 5) it shows hit for DHPS, HPPK, and DHNA domains. This therefore indicates with a high degree of confidence that the DHNA-HPPK-DHPS tri-fusion protein in *A. castellanii* is transcribed as a single fusion gene and is not merely a series of tandemly arrayed genes in the genome. Furthermore, this analysis confirms that the domain architecture of the tri-fusion is arranged in the same pattern as that of the Mycetozoa and is essentially the same architecture as that of fungi (with some fungi possessing a tandem duplication of the DHNA domain).

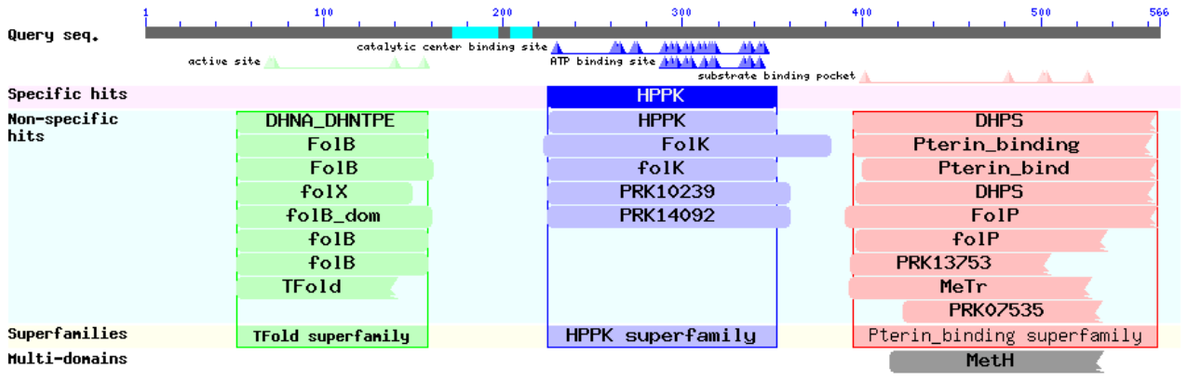

FIGURE 5 – CONSERVED DOMAINS IDENTIFIED USING THE RPS-BLAST OF THE NCBI CONSERVED DOMAIN DATABASE WITHIN THE *ACANTHAMOEBA* PUTATIVE DHNA-HPPK-DHPS TRANSCRIPT. DHNA (CD00651) SCORE- 41, E-VALUE- $1E^{-7}$; HPPK (CD00483) SCORE- 156, E-VALUE- $1E^{-42}$; DHPS (CD00423) SCORE- 180, E-VALUE- $1E^{-49}$.

EVOLUTIONARY ANALYSIS OF THE FOLATE SYNTHESIS GENES

To investigate the ancestry of the fusion genes I performed an evolutionary analysis of the DHNA, HPPK and DHPS folate synthesis genes separately and then produced a series of concatenated phylogenies based on the four combinations of these domain alignments. I used these phylogenies to investigate the branching order of the different fusion architectures enabling one to map when and where the fusion events occur relative to the eukaryotic tree and whether they represent monophyletic derived characters that are useful as SDCs. The single-domain analyses haven't been included in this section due to having relatively low bootstrap support throughout the trees compared to the concatenated analyses but can be viewed in the appendices (see Figure 22 to Figure 24).

COMPARATIVE GENOMIC AND MOLECULAR BIOLOGICAL SAMPLING OF THE FOLATE SYNTHESIS GENES

From surveying a wide-range of eukaryotic genomes representing the 6 eukaryotic supergroups (Yoon et al., 2008) (Excavata, Amoebozoa, Opisthokonta, Rhizaria, Chromaleveolata, Archaeplastida) I discovered the distribution of the folate biosynthetic genes. None of the three folate fusion gene homologues were identified in any of the Excavata (except a DHNA encoding gene present in *Naegleria gruberi*) or Rhizaria (except all 3 unfused homologues

in *Paulinella chromatophora*). Within the Chromalveolata the HPPK-DHPS bi-fusion gene was consistently found in the stramenopiles, within the dinoflagellate alveolate *Perkinsus marinus* and within the apicomplexan alveolate *Toxoplasma gondii.* However, an unfused form of the DHPS enzyme was encountered in *Plasmodium falciparum* suggesting a possible reversion of the HPPK-DHPS bi-fusion in this species. The Archaeplastida were more varied with none of the folate synthesis gene homologues being identified in the Rhodophyta or Glaucophyta and the unfused putative DHNA homologue and HPPK-DHPS bi-fusion being identified in the Embryophyta (land plants) and Cholrophyta. The chlorophyte *Volvox carteri f. Sp. nagriensis* was identified as having three unfused homologues suggesting a possible reversion of the fusion in this taxon. In the Amoebozoa the triple fusion was encountered within the mycetozoan slime moulds and the Centramoebida *A. castellanii*. Finally, within the Opisthokonta folate biosynthesis gene homologues were indentified in a few cases within the Metazoa – a DHNA-HPPK bi-fusion in *Branchistoma floridae* and *Nematostella vectensis*, and an unfused copy of the DHNA gene in *Trichoplax adherens.* All the main four fungal phyla possess the four domain gene architecture of DHNA-DHNA-HPPK-DHPS with a possible reversion in the basidiomycetes *Malassezia globosa* (unfused DHNA and HPPK-DHPS bi-fusion) and *Laccaria bicolor* (tandem DHNA-DHNA and HPPK-DHPS bi-fusion). *Capsaspora owczarki,* an independent opisthokont lineage related to the fungi (Ruiz-Trillo et al., 2004), also represented a possible reversion possessing DHNA-DHNA encoding gene and the HPPK-DHPS bi-fusion.

The bacteria lacked the tri-fusion but each of the unfused synthesis gene homologues were consistently identified across a wide range of different phyla. The bacteria also contained instances of the DHNA-HPPK bi-fusion as well as a putative duplication and fusion of the HPPK homologues (Figure 23). All three phyla of the Archaea (Crenarcheota, Euryarcheota, Korarcheota) possessed members with the unfused DHPS but seemed to lack the DHNA or HPPK domain. However, the Crenarchaeota *Sulfolobus acidocaldarius* did possess the HPPK gene homologues in a putative DHNA-HPPK bi-fusion with high sequence similarity to the fusion identified in the Metazoa suggesting a possible eukaryote to archaea HGT (Figure 8).

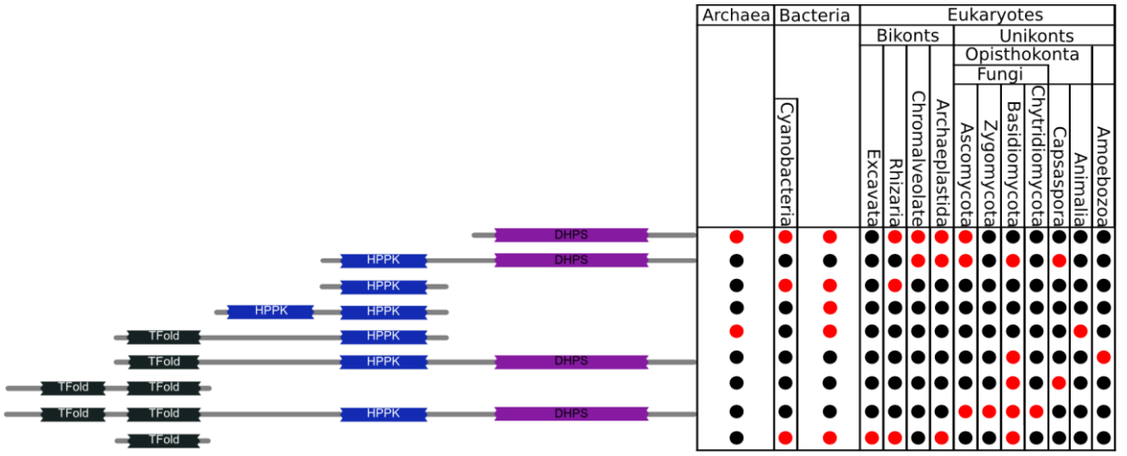

FIGURE 6 – FOLATE BIOSYNTHESIS GENE FUSION STATES ACROSS TAXA

While most of the archaea were identified as having an unfused copy of the DHPS homologues a unique bi-fusion between DHPS and a folylpolyglutamate synthase gene (FPGS) was discovered in the euryarcheal Halobacteria species. While some of the Euryarchaea did also possess a putative unfused DHPS paralogue it consistently demonstrated a lower identity match to the DHPS seed sequence suggesting it was a highly variant form or, alternatively, undergoing degradation and loss-of-function compared to highly conserved DHPS in the FPGS-DHPS bi-fusion. Multiple species such as *Plasmodium falciparum* (Salcedo et al., 2001) and *Pseudomonas aeruginosa* (Murata et al., 2000) contained a putative bi-functional fusion protein known as FolC containing a FPGS and DHFS functionality. What is intriguing about FPGS existing in fusions with two different members of the folate synthesis pathway is that this domain is activated by the binding of the end-product of this pathway- folate (Sun et al., 2001) suggesting another possible incidence of biochemical channelling in the folate pathway. In other Euryarchaeal species, such as the Methanococci, a possible DHPS domain protein was identified but was demonstrated to possess low identity when compared to the seed sequences so therefore it can only be definitively stated as being a member of the same pterin-binding domain superfamily. There is an indication that this is in fact part of the TIGR00284 gene, a gene of unknown function believed to be an archaeal alternative to DHPS (Marchler-Bauer et al., 2011).

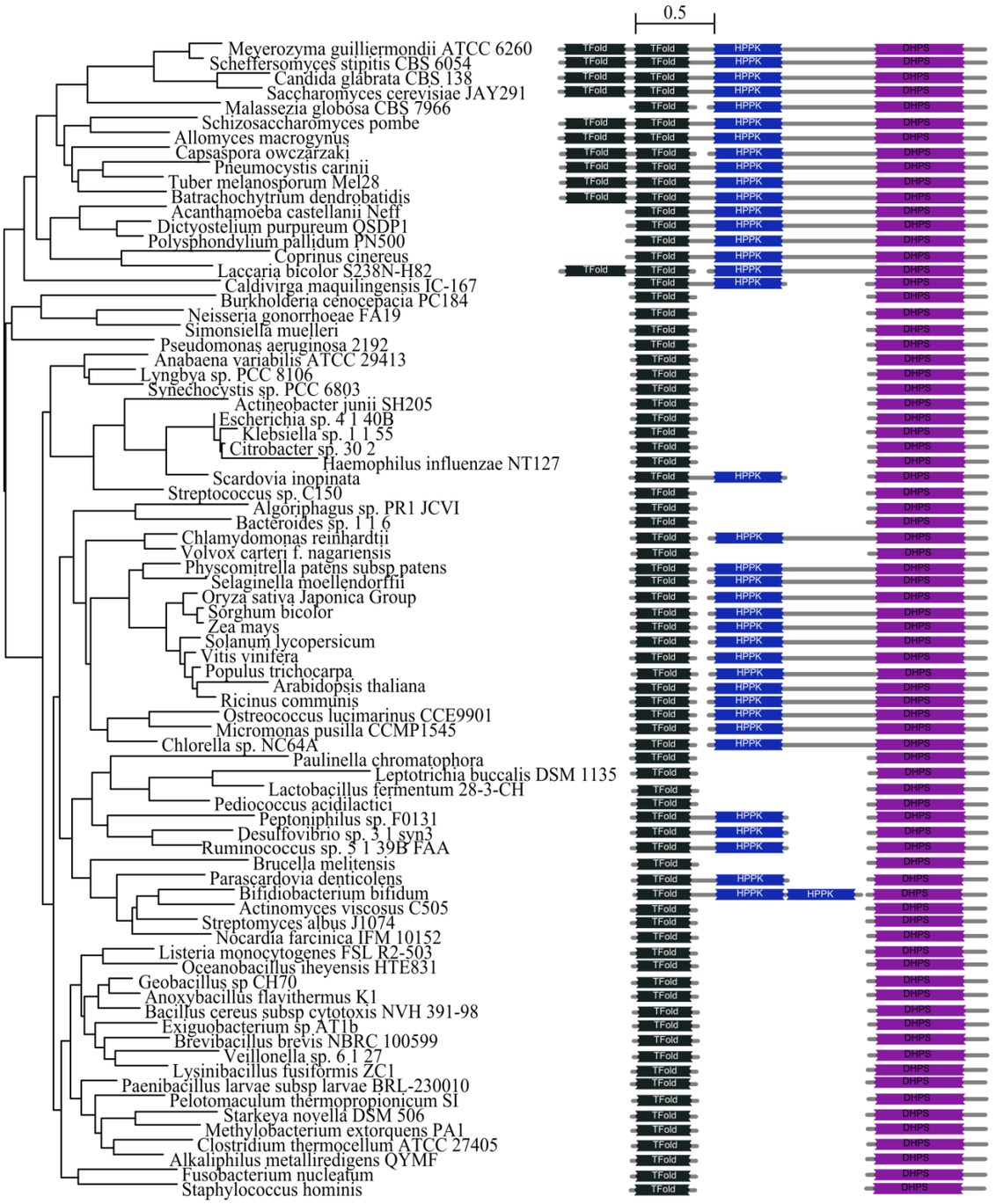

FIGURE 7 – DHNA-DHPS - PHYML TREE RECONSTRUCTED USING 78 SEQUENCES AND 173 CHARACTERS SHOWING MONOPHYLY OF THE UNIKONTS AND ARCHAEPLASTIDA AS WELL AS A POSSIBLY HGT BRANCH POSISTION FOR *PAULINELLA*

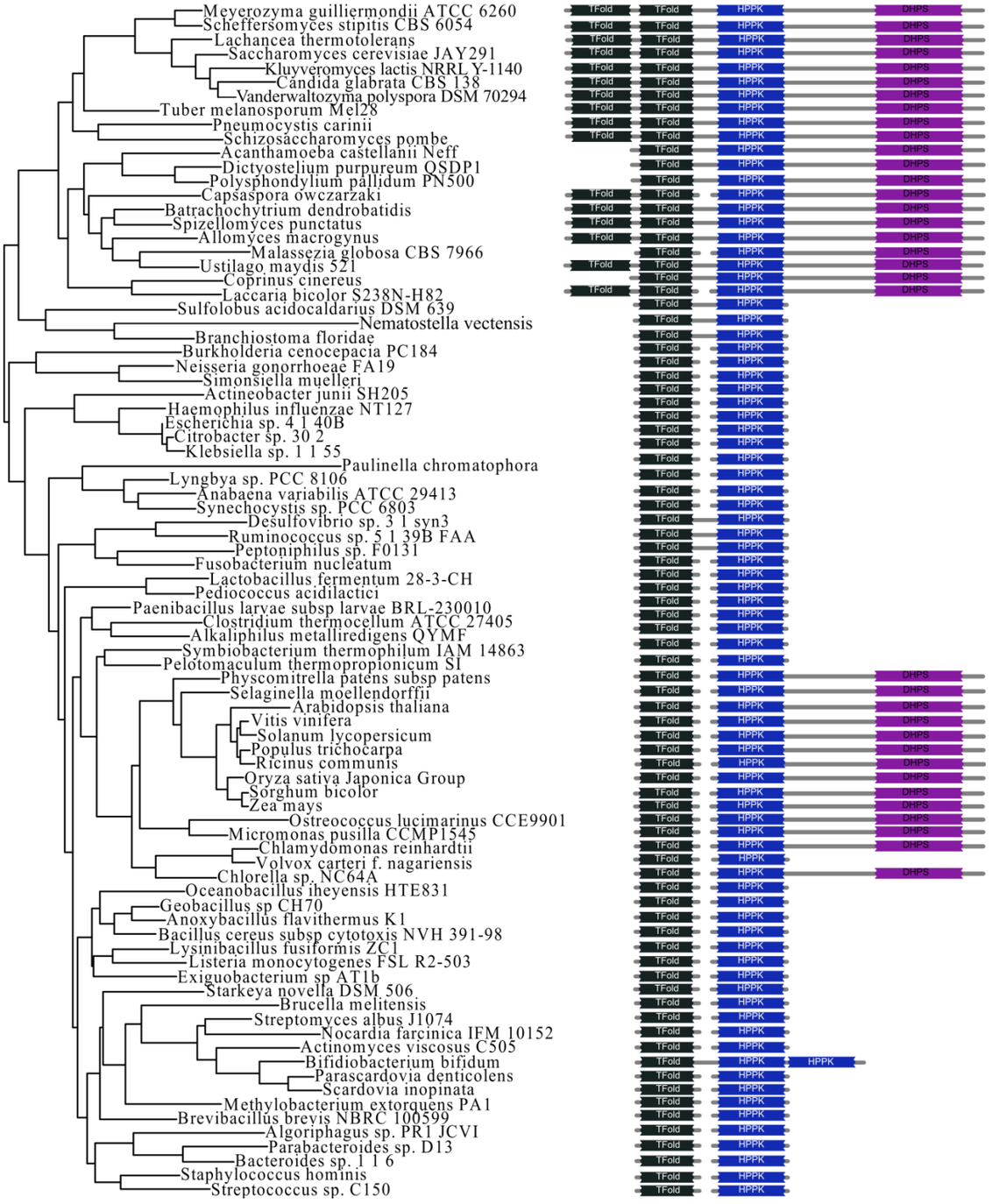

FIGURE 8 – DHNA-HPPK - PHYML TREE RECONSTRUCTED USING 84 SEQUENCES AND 177 CHARACTERS SHOWING MONOPHYLY OF THE UNIKONTS AND THE PUTATIVE METAZOA TO *SULFOLOBUS* HGT

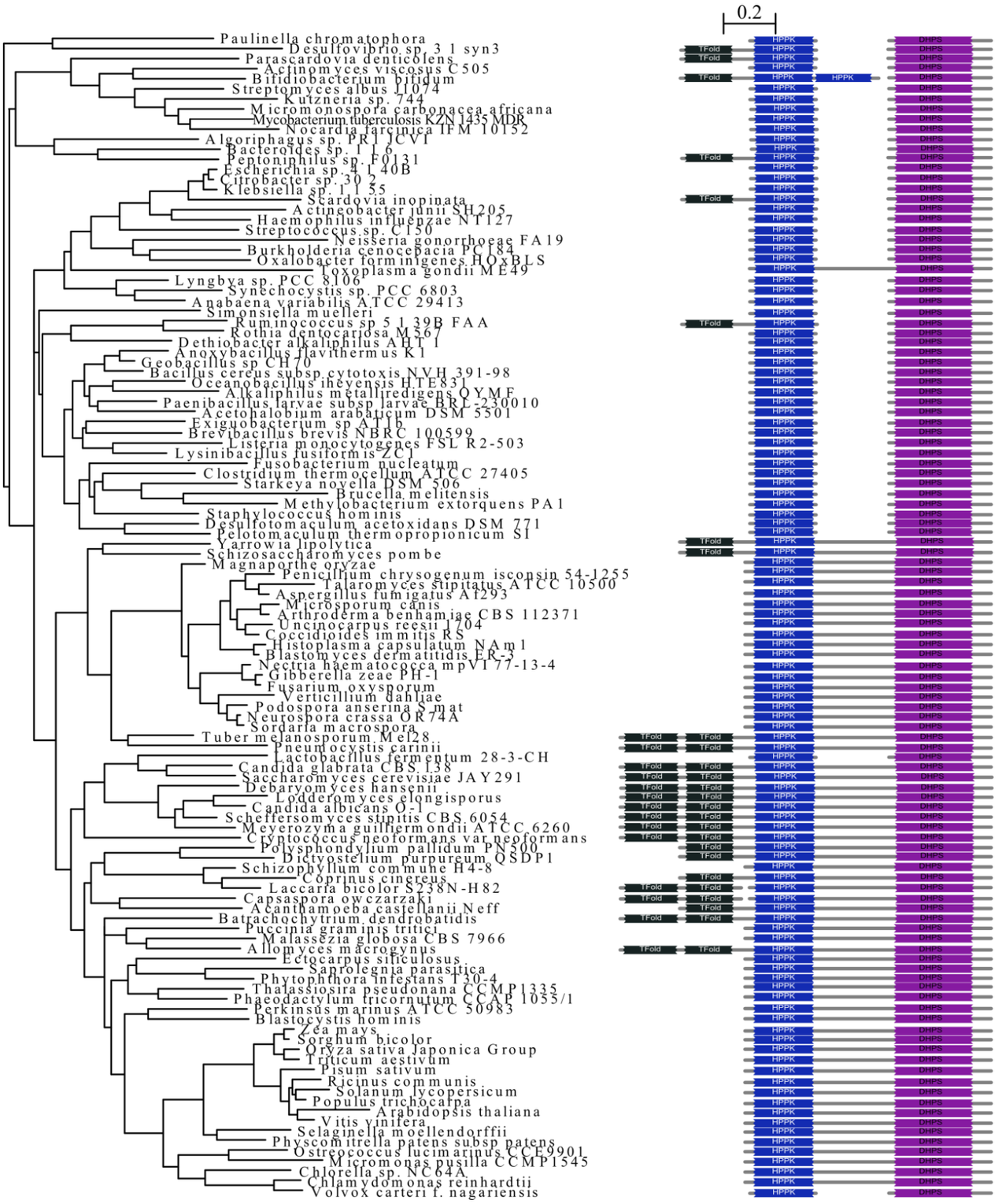

FIGURE 9 – HPPK-DHPS PHYML TREE RECONSTRUCTED USING 116 SEQUENCES AND 166 CHARACTERS SHOWING THE MONOPHYLY OF THE ARCHAEPLASTIDA

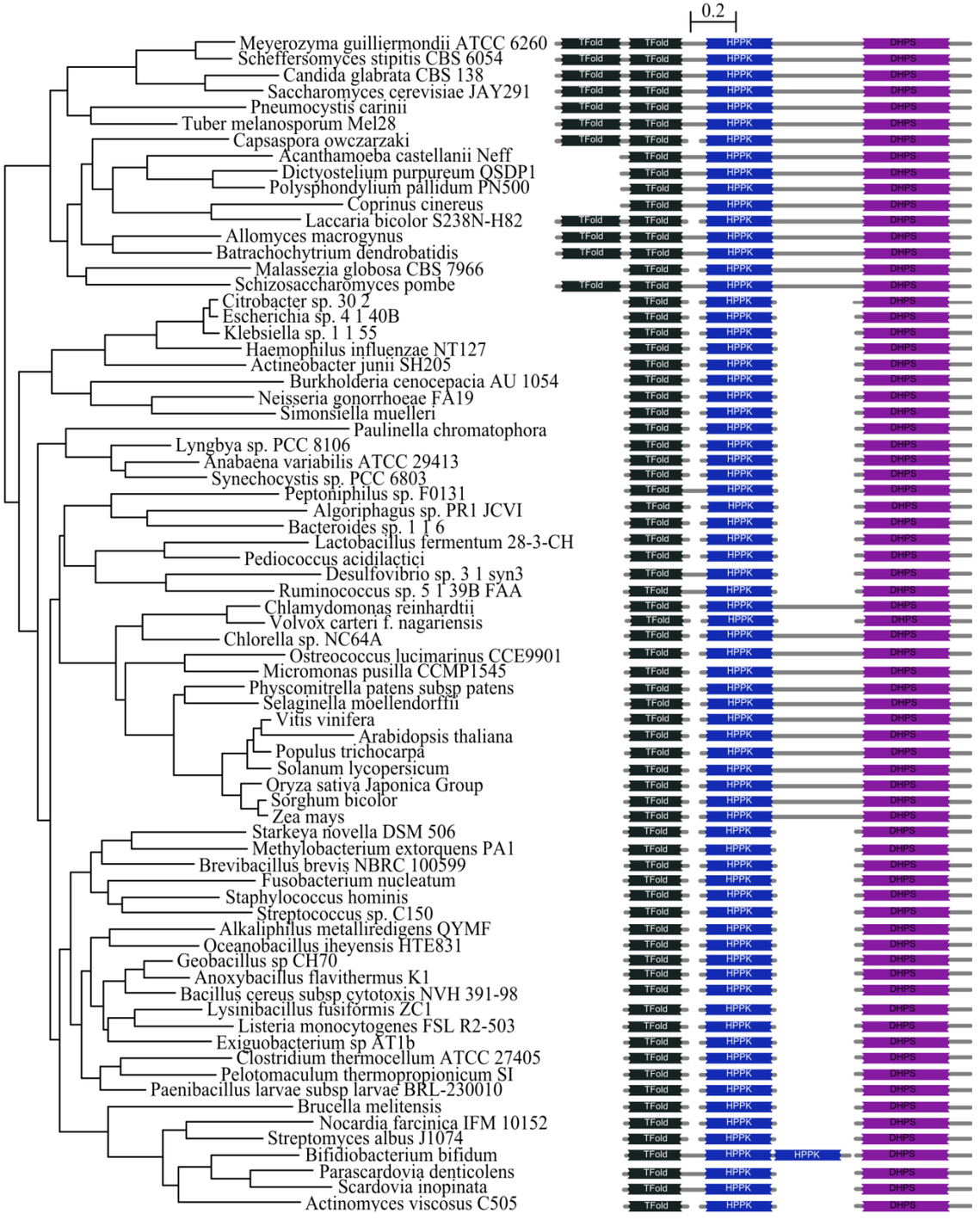

FIGURE 10 – DHNA-HPPK-DHPS - PHYML TREE RECONSTRUCTED USING 73 SEQUENCES AND 213 CHARACTERS SHOWING THE MONOPHYLY OF THE UNIKONTS AND ARCHAEPLASTIDA AS WELL AS POSSIBLE *PAULINELLA* HGT IN BRANCHING OF *PAULINELLA* WITH THE CYANOBACTERIA

MONOPHYLY OF KEY EUKARYOTE TAXA

The concatenated phylogenetic analysis confirmed with reasonable support the monophyly of several key eukaryotic taxa, in particular the Archaeplastida (98.8% in the DHNA-HPPK-DHPS phylogeny) and the unikonts (97.1% in the DHNA-HPPK-DHPS phylogeny). Full bootstrapped phyogenies can be found in the appendix (Figure 25 to Figure 31) however, key bootstrap values are highlighted in Figure 11 below. While the monophyly of the eukaryotes was not the highest supported topological relationship the bootstrap values for the bacterial clades grouping between the eukaryotic taxa tended to have very low support values (for instance in the DHNA-HPPK-DHPS tree - 3.2% for the Firmicutes grouping with the Archaeplastida, and 21.9% for the unikonts branching as an outgroup of the Archaeplastida, Chromalevolata and bacteria) meaning this can be regarded as unresolved. Another possible explanation for this is that the Archaeplastida (with primary plastids) and the chromalveolates (with secondary plastids) have obtained their folate biosynthesis enzymes by separate sources (e.g. photosynthetic endosymbiosis). This explanation is unlikely; however, as the bi-fusion architecture links it to the eukaryotes suggesting it is an ancient eukaryotic character (weakly supported in the HPPK-DHPS tree (Figure 9) which would best test this relationship).

HORIZONTAL GENE TRANSFER

Two main putative Prokaryote-Eukaryote HGTs were identified- *Paulinella chromatophora* and *Sulfolobus acidocaldarius*. The Crenarchaea *S. acidocaldarius* was discovered to branch with the 2 metazoan species (*B. floridae, N. vectensis*) with weak support (29.5% in HPPK phylogeny Figure 23 and 52.2% in the DHNA-HPPK concatenation (Figure 8)) forming an out-group to the rest of unikonts (43.7%) in the DHNA-HPPK phylogeny. This combined with the shared DHNA-HPPK found in this out-group provides tentative support for an HGT from the metazoans, or a close relative, to *S. acidocaldarius*. The cercozoan *Paulinella chromatophora* was found to group with the cyanobacteria in the DHNA-HPPK concatenation and the DHNA-HPPK-DHPS concatenation (Figure 10) (with 29.5% and 29.2% support respectively). A compelling result as *P. chromatophora* is believed to have relatively

recently undergone endosymbiosis with a relative of *Synechococcus* as an independent primary photosynthetic endosymbiosis (Marin et al., 2005). It should be noted that *P. chromatophora* does branch with *Desulfovibrio* sp. 3.1 syn3 in the HPPK-DHPS phylogeny (Figure 9) (39.4%) and in HPPK phylogeny (Figure 23) (15.2%). As *P. chromtophora* was ancestrally a serial phagotroph (Nowack et al., 2008) it could have potentially acquired this gene from food bacterium through Doolittle's 'gene transfer ratchet' which proposes phagotrophy acts a major route of gene transfer into eukaryotic genomes (Doolittle, 1998).

	DHNA	HPPK	DHPS	DHNA-HPPK	HPPK-DHPS	DHNA-DHPS	DHNA-HPPK-DHPS
Embryophyta clade	88.5%	65.4%	X	98.7%	91.5%	90.3%	99.1%
Archaeplastida clade	X	X	X	X	78.1%	60.5%	98.8%
Amoeba clade	86.1%	X	X	84.5%	X	89.5%	89.4%
Metazoa clade	X	60.2%	/	75.3%	/	/	/
Unikonta clade	14.7%	X	X	43.7%	X	32.8%	97.1%
Sulfolobus acidocaldarius HGT	X	29.5%	/	52.2%	/	/	/
Paulinella chromatophora HGT	X	X	X	29.5%	X	X	29.2%

FIGURE 11 – BOOTSTRAP SUPPORT VALUES OF SPECIFIC CLADES AND PUTATIVE HGTS IN ALL THE PHYLOGENETIC ANALYSES CONDUCTED IN THIS STUDY (X REPRESENTING TOPOLOGY IS NOT THE MOST SUPPORTED AND / REPRESENTING THE ABSENCE OF THIS BRANCH IN THESE PHYLOGENIES

FUSION STATES

In the DHNA-HPPK-DHPS concatenation there is strong support (97.1%) for the Archaeplastida clade with the HPPK-DHPS bi-fusion (apart from *Volvox cateri f. Nagariensis*). Similarly, there is a well supported (98.8%) unikont clade containing all the identified incidences of the triple DHNA-HPPK-DHPS fusion. Interestingly, it appears all the fungi with the exception of the tri-fusion of *Coprinus cinereus* and the unfused DHNA and HPPK-DHPS bi-fusion of *Malassezia globosa* have a duplication of the DHNA domain. As this tandem duplication is mostly incorporated into

the tri-fusion (with the exception of probably fusion reversions in *Malassezia globosa* and *Laccaria bicolor*) the most parsimonious explanation is that this duplication occurred after the unikont-bikont bifurcation. There were no identified duplications of the DHPS domain in a fusion and there was a single recorded HPPK duplication in the *Bifidiobacterium bifidum* DHNA-HPPK-HPPK tri-fusion. The lower E-value for the second copy of the HPPK domain ($1e^{-8}$ versus $2.51e^{-34}$) is indicative of a duplication then degradation as evolutionary constraints are relaxed on the duplicate domain.

DISCUSSION

The bootstrap support for the phylogenies was generally weak, an inevitable consequence of an analysis covering an ancient and divergent evolutionary history and with a limited number (compared to the 5000 characters predicted as being required to robustly resolve the Archaeplastida (Rodriguez-Ezpeleta et al., 2005)) of phylogenetically informative sites available for each protein domain alignments (see Figure 21 for summary). This means it is difficult to rely on phylogenies alone to reconstruct the evolutionary history of the folate biosynthesis pathway and therefore phylogenetics must be combined with comparative genomics of the gene fusion characters.

The use of the DHNA-HPPK-DHPS tri-fusion as an SDC is limited by four potential evolutionary phenomenon: HGT, hidden paralogy, reversions (including both loss and fissions), and convergent evolution (Richards, 2005, Simpson and Roger, 2002). Firstly, from the phylogenies produced I can pin down at least two putative HGTs (as described previously) as well as another unlikely but possible HGT from bacteria to the parasitic protist *Toxoplasma gondii* (with very weak support- 5.2%). Fortunately, there are no apparent HGTs in which a fused form is replaced by unfused form. However, due to incomplete taxon sampling and phylogenetic uncertainty it is currently not possible to identify all the HGTs which may affect the interpretation of the phylogenetic support and comparative genomics.

Hidden paralogy has long been acknowledged as a problem for phylogenies (Richards et al., 2003), it occurs when a duplication event has been followed by the differential loss of paralogues (Gribaldo and Philippe, 2002, Daubin et al., 2001). This can lead to the formation of incongruous trees as it means the gene is no longer directly tracking the evolution of the species (Richards, 2005). Even standard cases of paralogy can pose a problem to the construction of phylogenies especially if there is inconsistent paralogue sampling. This is a potential source of error especially considering that many of the sequences used in these reconstructions were acquired from incomplete genome sequencing projects. Fortunately, only in very few cases were paralogues actively identified during the genome sampling (See Figure 20 for examples) but again it doesn't discount the possibility of further cases of

unidentified hidden paralogy. When paralogues were encountered the fusion form always had a higher domain identity and was thus chosen for phylogenetic reconstruction but it is still a possible source of error in the phylogenetic resolution of the bacteria. In particular, there is some evidence the DHNA in many bacteria has been replaced with putatively functionally equivalent 6-pyruvoyltetrahydropterin synthase (PTSP) (Pribat et al., 2009). This could also account for the fact that many of the bacterial putative DHNA were only identifiable as divergent members of the T-fold domain superfamily.

The third obfuscating factor in identifying the evolutionary ancestry of this gene fusion is the prevalence of reversion events. Unaccounted reversion of the fusion state can severely confuse attempts to use comparative genomics in order to discern phylogenetic synapomorphies. The literature is divided when it comes to estimates of the relative rate of gene fusions and with some groups finding that gene fissions are more common than fusions (in 17 prokaryotic species) (Snel et al., 2000). The broadest study conducted so far by Kummerfeld and Teichmann shows that gene fissions occur at roughly ¼ the rate of gene fusions (Kummerfeld and Teichmann, 2005) an appealing figure when one considers the genetic mechanisms behind these processes. In particular, the fact that for a fusion to take place a terminal region from one gene and the initial regulatory region is required to be lost whereas fission would require the gain of these features. Genetically a loss is a lot simpler than the gain of not one but two specific elements (Stechmann and Cavalier-Smith, 2002). On the basis of the triple concatenated analysis of the domains there were four identifiable putative reversions of the tri-fusion (*Capsaspora owczarzaki, Laccaria bicolor, Malassezia globosa,* and *Volvox carteri f. sp. nagariensis*). The most common reversion appears to be between the DHNA domain (duplicated or not) and the HPPK domain with three examples in the fungi. In the DHNA-HPPK-DHPS fusions there also only appears to be incidences of terminal fission (i.e. the fusion is lost between a domain at one end of the ORF rather than a central domain). However, owing to the relatively few representatives of several taxa (as a result of the breadth of the study and limited genome sequencing data) it is conceivable that all represented taxa have reversions thus masking the true fusion character of the clade. For

instance, if all non-unikont eukaryotes sampled had a reversion from the tri-fusion, it would indicate that the tri-fusion possibly occurred ancestrally to the eukaryotes. However, this is a less parsimonious scenario, as it requires numerous more fission or loss events than if the tri-fusion occurred in the last common ancestor of the unikonts. The most important reversions to explain if we are to use the tri-fusion as a unikont synapomorphy is the DHNA-HPPK bi-fusion discovered in the two metazoan species (*B. floridae* and *N. vectensis*) and the loss in the sampled Archaemoebae. There are several possible explanations behind the presence of the folate synthesis genes in these two Metazoa (in order of decreasing parsimony):-

- A partial loss of the unikont tri-fusion (loss of DHPS domain from DHNA-HPPK-DHPS to form the DHNA-HPPK bi-fusion
- The HPPK-DHPS bi-fusion undergoing fission and loss of the DHPS domain followed by the fusion of the HPPK with DHNA.
- Loss of this pathway and then the secondary recovery of the DHNA-HPPK bi-fusion via a HGT from species possessing the DHNA-HPPK bi-fusion

The unfused DHNA in *Trichoplax adherens* was very poorly resolved so I cannot speculate on to its origin.

The absence of the folate biosynthesis genes in the *Entamoeba* is most likely as a by-product of adaptation to parasitism in this organism which has led to the stream-lining and reduction of its genome (Loftus et al., 2005) and doesn't discount the theory that the tri-fusion is an ancestral character to the Amoebozoa.

The final potential problem with using this DHNA-HPPK-DHPS gene fusion as an SDC is that there is a risk of homoplasies i.e. multiple evolutionary innovations of same character leading to identification of false synapomorphies. However, this appears relatively unlikely in this dataset due to:

- An absence of distinct strongly supported polyphyletic clades of similar fusion states being rooted with an unfused or less fused group.

- The tri-fusion most likely being the product of two separate fusion events.

- Conservation of domain order in the fusion character (Bashton and Chothia, 2002).

Having accounted for all the above factors as much as possible with currently available datasets I can combine using the fusion states as SDCs with the most supported clades in the phylogenies – i.e. the Archaeplastida, the unikonts, the Amoebozoa, and the Chromalveolata (as seen in Figure 12). This suggests that the DHNA-HPPK-DHPS tri-fusion is most likely to have originated from the additional fusion of DHNA with HPPK-DHPS (and not the entire fusion occurring de-novo) in the last common ancestor of the unikonts and that the HPPK-DHPS bi-fusion is a genetic synapomorphy for the eukaryotes (with secondary losses).

Therefore, I can propose a model of how the folate biosynthesis fusion states evolved (Figure 12) and declare that these results provide tentative support for the holophyly of the unikonts.

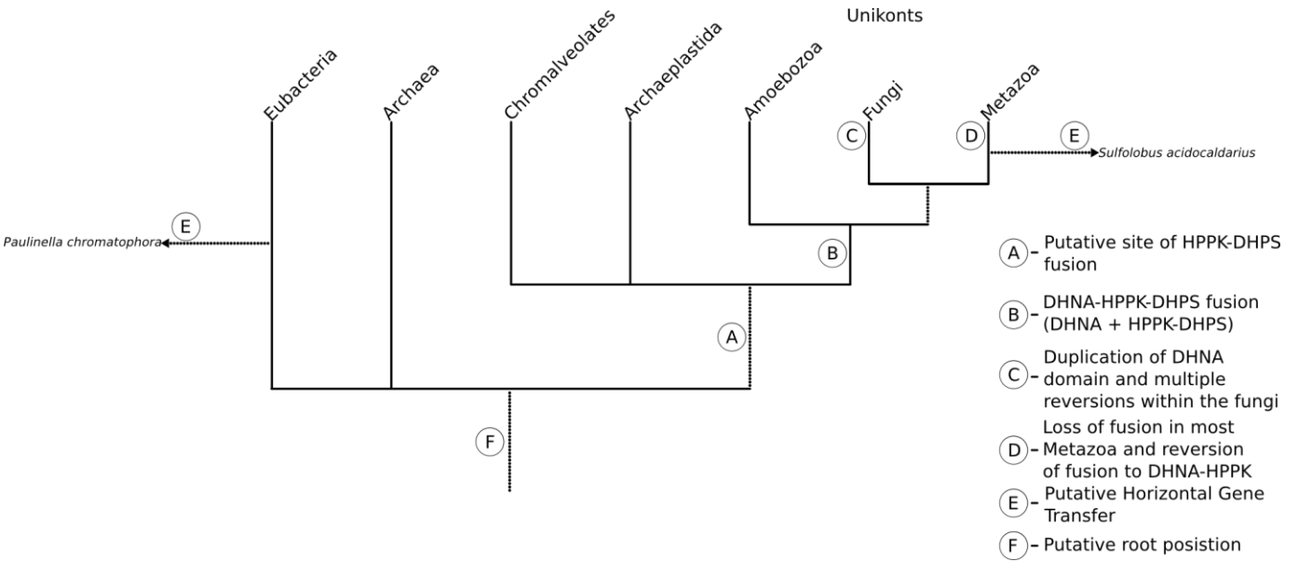

FIGURE 12 – EVOLUTIONARY RELATIONSHIPS PREDICTED FROM FOLATE BIOSYNTHESIS GENES WITH DOTTED LINES REPRESENTING THE MORE SPECULATIVE RELATIONSHIPS

ACKNOWLEDGEMENTS

I am grateful to CEEM for help and supervision.

REFERENCES

ALTSCHUL, S., GISH, W., MILLER, W., MYERS, E. & LIPMAN, D. 1990. Basic Local Alignment Search Tool. *Journal of Molecular Biology***,** 403-410.

ANDERSON, I., WATKINS, R., SAMUELSON, J., SPENCER, D., MAJOROS, W., GRAY, M. & LOFTUS, B. 2005. Gene discovery in the Acanthamoeba castellanii genome. *Protist***,** 203-214.

ARCHIBALD, J. M., LONGET, D., PAWLOWSKI, J. & KEELING, P. J. 2003. A novel polyubiquitin structure in Cercozoa and Foraminifera: Evidence for a new eukaryotic supergroup. *Molecular Biology and Evolution,* 20**,** 62-66.

ARISUE, N., HASEGAWA, M. & HASHIMOTO, T. 2005. Root of the eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. *Molecular Biology and Evolution***,** 409-420.

BAPTESTE, E., BOUCHER, Y., LEIGH, J. & DOOLITTLE, W. 2004. Phylogenetic reconstruction and lateral gene transfer. *Trends in Microbiology***,** 406-411.

BASHTON, M. & CHOTHIA, C. 2002. The geometry of domain combination in proteins. *Journal of Molecular Biology***,** 927-939.

BASSET, G. J. C., QUINLIVAN, E. P., RAVANEL, S., RÉBEILLÉ, F., NICHOLS, B. P., SHINOZAKI, K., SEKI, M., ADAMS-PHILLIPS, L. C., GIOVANNONI, J. J., GREGORY, J. F. & HANSON, A. D. 2004. Folate synthesis in plants: The p-aminobenzoate branch is initiated by a bifunctional PabA-PabB protein that is targeted to plastids. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 1496-1501.

BENSON, D., KARSCH-MIZRACHI, I., LIPMAN, D., OSTELL, J. & SAYERS, E. 2010. GenBank. *Nucleic Acids Research***,** D46-D51.

BERGSTEN, J. 2005. A review of long-branch attraction. *Cladistics***,** 163-193.

BRINKMANN, H. & PHILIPPE, H. 2007. The diversity of eukaryotes and the root of the eukaryotic tree. *Eukaryotic Membranes and Cytoskeleton: Origins and Evolution*, 20-37.

BROAD 2010. Broad Institute Data. *Broad Institure.*

BROWN, G. M. 1971. The biosynthesis of pteridines. *Advan. Enzymol. Relat. Areas Mol. Biol.,* 35, 35-77.

CAVALIER-SMITH, T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. *International Journal of Systematic and Evolutionary Microbiology*, 297-354.

CHOMCZYNSKI, P. & SACCHI, N. 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate phenol choloform extraction. *Analytical Biochemistry*, 156-159.

CONANT, G. & WAGNER, A. 2005. The rarity of gene shuffling in conserved genes. *Genome Biology*, -.

DARWIN, C. 1859. *On the Origin of Species by Means of Natural Selection*.

DAUBIN, V., GOUY, M. & PERRIÈRE, G. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Inform Ser Workshop Genome Inform,* 12, 155-164.

DE CRECY-LAGARD, V., EL YACOUBI, B., DE LA GARZA, R., NOIRIEL, A. & HANSON, A. 2007. Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. *Bmc Genomics*, -.

DOOLITTLE, R. F. 1995. The Origins and Evolution of Eukaryotic Proteins. *Philosophical Transactions: Biological Sciences,* 349, 235-240.

DOOLITTLE, W. E. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics,* 14, 307-311.

DURBIN 2010. HMMER 3.

EDGAR, R. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics*, 1-19.

FELSENSTEIN, J. 1978. Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Biology,* 27, 401-410.

FELSENSTEIN, J. 1985. Confidence-limits on phylogenies - an approach using the bootstrap. *Evolution*, 783-791.

GADAGKAR, S., ROSENBERG, M. & KUMAR, S. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*, 64-74.

GAFFNEY, E. 1977. *The side-necked turtle family Chelidae: A theory of relationships using shared derived characters*, American Museum of Natural History (New York).

GENECODES 2010. GeneCodes.

GOUY, M., GUINDON, S. & GASCUEL, O. 2010. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution*, 221-224.

GRIBALDO, S. & PHILIPPE, H. 2002. Ancient phylogenetic relationships. *Theoretical Population Biology,* 61, 391-408.

GUINDON, S. & GASCUEL, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 696-704.

HAMPL, V., HUG, L., LEIGH, J. W., DACKS, J. B., LANG, B. F., SIMPSON, A. G. B. & ROGER, A. J. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Sciences of the United States of America,* 106, 3859-3864.

HENNIG, W. 1966. *Phylogenetic Systematics*.

HOLTON, T. A. & GRAHAM, M. W. 1991. A simple and efficient method for direct cloning of PCR products using DDT-tailed vectors. *Nucleic Acids Research*, 1156-1156.

HUSON, D., RICHTER, D., RAUSCH, C., DEZULIAN, T., FRANZ, M. & RUPP, R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. *Bmc Bioinformatics*, -.

JENKINS, C. & FUERST, J. 2001. Phylogenetic analysis of evolutionary relationships of the planctomycete division of the domain bacteria based on amino acid sequences of elongation factor Tu. *Journal of Molecular Evolution*, 405-418.

JGI 2010. JGI Genome Portal. *Joint Genome Institute.*

JONES, M. D. M. 2007. *Oilfield microbiology and the effects of nitrate injection on bacterial communities.* PhD, University of Exeter.

KEANE, T., CREEVEY, C., PENTONY, M., NAUGHTON, T. & MCLNERNEY, J. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *Bmc Evolutionary Biology*, -.

KELLER, N., TURNER, G. & BENNETT, J. 2005. Fungal secondary metabolism - From biochemistry to genomics. *Nature Reviews Microbiology*, 937-947.

KROGH, A., BROWN, M., MIAN, I., SJOLANDER, K. & HAUSSLER, D. 1994. Hidden Markov-models in computational biology - application to protein modelling. *Journal of Molecular Biology*, 1501-1531.

KUHA, J. 2004. AIC and BIC - Comparisons of assumptions and performance. *Sociological Methods & Research*, 188-229.

KUMAR, S., SKJAEVELAND, A., ORR, R., ENGER, P., RUDEN, T., MEVIK, B., BURKI, F., BOTNEN, A. & SHALCHIAN-TABRIZI, K. 2009. AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. *Bmc Bioinformatics*, -.

KUMMERFELD, S. & TEICHMANN, S. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends in Genetics*, 25-30.

LAWRENCE, M., ILIADES, P., FERNLEY, R., BERGLEZ, J., PILLING, P. & MACREADIE, I. 2005. The three-dimensional structure of the bifunctional 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase/dihydropteroate synthase of Saccharomyces cerevisiae. *Journal of Molecular Biology*, 655-670.

LEONARD, G. 2010. Development of fusion and duplication finder BLAST (fdfBLAST): systematic tool to detect differentially distributed gene fusions and resolve trifurcations in the tree of life. PhD Thesis, University of Exeter.

LEONARD, G., STEVENS, J. & RICHARDS, T. 2009. REFGEN and TREENAMER: Automated Sequence Data Handling for Phylogenetic Analysis in the Genomic Era. *Evolutionary Bioinformatics*, 1-4.

LIANG, P. & ANDERSON, K. 1998. Substrate channeling and domain - Domain interactions in bifunctional thymidylate synthase - Dihydrofolate reductase. *Biochemistry*, 12195-12205.

LOFTUS, B., ANDERSON, I., DAVIES, R., ALSMARK, U., SAMUELSON, J., AMEDEO, P., RONCAGLIA, P., BERRIMAN, M., HIRT, R., MANN, B., NOZAKI, T., SUH, B., POP, M., DUCHENE, M., ACKERS, J., TANNICH, E., LEIPPE, M., HOFER, M., BRUCHHAUS, I., WILLHOEFT, U., BHATTACHARYA, A., CHILLINGWORTH, T., CHURCHER, C., HANCE, Z., HARRIS, B., HARRIS, D., JAGELS, K., MOULE, S., MUNGALL, K., ORMOND, D., SQUARES, R., WHITEHEAD, S., QUAIL, M., RABBINOWITSCH, E., NORBERTCZAK, H., PRICE, C., WANG, Z., GUILLEN, N., GILCHRIST, C., STROUP, S., BHATTACHARYA, S., LOHIA, A., FOSTER, P., SICHERITZ-PONTEN, T., WEBER, C., SINGH, U., MUKHERJEE, C., EL-SAYED,

N., PETRI, W., CLARK, C., EMBLEY, T., BARRELL, B., FRASER, C. & HALL, N. 2005. The genome of the protist parasite Entamoeba histolytica. *Nature*, 865-868.

LOFTUS, B. J. 2008. Acanthamoeba castellanii sequencing white paper. *Baylor College of Medicine Human Genome Sequencing Center.*

LOPEZ, P. & LACKS, S. 1993. A bifunctional protein in the folate biosynthetic-pathway of Streptococcus pneuomiae with dihydroneopterin aldolase and hydroxymethyldihydropterin pyrophosphokinase activities. *Journal of Bacteriology*, 2214-2220.

MARCHLER-BAUER, A., ANDERSON, J., CHITSAZ, F., DERBYSHIRE, M., DEWEESE-SCOTT, C., FONG, J., GEER, L., GEER, R., GONZALES, N., GWADZ, M., HE, S., HURWITZ, D., JACKSON, J., KE, Z., LANCZYCKI, C., LIEBERT, C., LIU, C., LU, F., LU, S., MARCHLER, G., MULLOKANDOV, M., SONG, J., TASNEEM, A., THANKI, N., YAMASHITA, R., ZHANG, D., ZHANG, N. & BRYANT, S. 2009. CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Research*, D205-D210.

MARCHLER-BAUER, A., LU, S., ANDERSON, J. B., CHITSAZ, F., DERBYSHIRE, M. K., DEWEESE-SCOTT, C., FONG, J. H., GEER, L. Y., GEER, R. C., GONZALES, N. R., GWADZ, M., HURWITZ, D. I., JACKSON, J. D., KE, Z., LANCZYCKI, C. J., LU, F., MARCHLER, G. H., MULLOKANDOV, M., OMELCHENKO, M. V., ROBERTSON, C. L., SONG, J. S., THANKI, N., YAMASHITA, R. A., ZHANG, D., ZHANG, N., ZHENG, C. & BRYANT, S. H. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research,* 39, D225-D229.

MARCOTTE, E. M., PELLEGRINI, M., NG, H.-L., RICE, D. W., YEATES, T. O. & EISENBERG, D. 1999. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. *Science,* 285, 751-753.

MARIN, B., NOWACK, E. & MELKONIAN, M. 2005. A plastid in the making: Evidence for a second primary endosymbiosis. *Protist*, 425-432.

MARSHALL, O. 2004. PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics*, 2471-2472.

MATSUZAKI, M., MISUMI, O., SHIN-I, T., MARUYAMA, S., TAKAHARA, M., MIYAGISHIMA, S., MORI, T., NISHIDA, K., YAGISAWA, F., YOSHIDA, Y., NISHIMURA, Y., NAKAO, S., KOBAYASHI, T., MOMOYAMA, Y., HIGASHIYAMA, T., MINODA, A., SANO, M., NOMOTO, H., OISHI, K., HAYASHI, H., OHTA, F., NISHIZAKA, S., HAGA, S., MIURA, S., MORISHITA, T., KABEYA, Y., TERASAWA, K., SUZUKI, Y., ISHII, Y., ASAKAWA, S., TAKANO, H., OHTA, N., KUROIWA, H., TANAKA, K., SHIMIZU, N., SUGANO, S., SATO, N., NOZAKI, H., OGASAWARA, N., KOHARA, Y. & KUROIWA, T. 2004. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature*, 653-657.

MEEK, T. D., GARVEY, E. P. & SANTI, D. V. 1985. Purification and characterization of the bifunctional thymidylate synthetase dihydrofolate reductase from methotrexate-resistant leishmania-tropica. *Biochemistry,* 24, 678-686.

MOUILLON, J., RAVANEL, S., DOUCE, R. & REBEILLE, F. 2002. Folate synthesis in higher-plant mitochondria: coupling between the dihydropterin pyrophosphokinase and the dihydropteroate synthase activities. *Biochemical Journal*, 313-319.

MURATA, T., BOGNAR, A., HAYASHI, T., OHNISHI, M., NAKAYAMA, K. & TERAWAKI, Y. 2000. Molecular analysis of the folC gene of Pseudomonas aeruginosa. *Microbiology and Immunology*, 879-886.

NAKAMURA, Y., ITOH, T. & MARTIN, W. 2007. Rate and polarity of gene fusion and fission in Oryza sativa and Arabidopsis thaliana. *Molecular Biology and Evolution*, 110-121.

NARA, T., HSHIMOTO, T. & AOKI, T. 2000. Evolutionary implications of the mosaic pyrimidine-biosynthetic pathway in eukaryotes. *Gene*, 209-222.

NEFF, R. J. 1957. Purification, axenic cultivation, and description of a soil amoeba, Acanthaemoeba sp. *Journal of Protozoology*, 176-182.

NOWACK, E., MELKONIAN, M. & GLOCKNER, G. 2008. Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. *Current Biology*, 410-418.

NOZAKI, H., MATSUZAKI, M., MISUMI, O., KUROIWA, H., HIGASHIYAMA, T. & KUROIWA, T. 2005. Phylogenetic implications of the CAD complex from the primitive red alga Cyanidioschyzon merolae (Cyanidiales, Rhodophyta). *Journal of Phycology*, 652-657.

O'BRIEN, E., KOSKI, L., ZHANG, Y., YANG, L., WANG, E., GRAY, M., BURGER, G. & LANG, B. 2007. TBestDB: a taxonomically broad database of expressed sequence tags (ESTs). *Nucleic Acids Research*, D445-D451.

PERRIERE, G. & GOUY, M. 1996. WWW-Query: An on-line retrieval system for biological sequence banks. *Biochimie*, 364-369.

PHILIPPE, H., ZHOU, Y., BRINKMANN, H., RODRIGUE, N. & DELSUC, F. 2005. Heterotachy and long-branch attraction in phylogenetics. *Bmc Evolutionary Biology*, -.

PRIBAT, A., JEANGUENIN, L., LARA-NUNEZ, A., ZIEMAK, M., HYDE, J., DE CREY-LAGARD, V. & HANSON, A. 2009. 6-Pyruvoyltetrahydropterin Synthase Paralogs Replace the Folate Synthesis Enzyme Dihydroneopterin Aldolase in Diverse Bacteria. *Journal of Bacteriology*, 4158-4165.

RICHARDS, T. 2005. *Horizontal Gene Transfer and the Evolution of the Eukaryotes,* D. Phil Thesis, University of Oxford.

RICHARDS, T. & CAVALIER-SMITH, T. 2005. Myosin domain evolution and the primary divergence of eukaryotes. *Nature*, 1113-1118.

RICHARDS, T. A., DACKS, J. B., JENKINSON, J. M., THORNTON, C. R. & TALBOT, N. J. 2006. Evolution of filamentous plant pathogens: Gene exchange across eukaryotic kingdoms. *Current Biology,* 16, 1857-1864.

RICHARDS, T. A., HIRT, R. P., WILLIAMS, B. A. P. & EMBLEY, T. M. 2003. Horizontal gene transfer and the evolution of parasitic protozoa. *Protist,* 154**,** 17-32.

RODRIGUEZ-EZPELETA, N., BRINKMANN, H., BUREY, S. C., ROURE, B., BURGER, G., LOFFELHARDT, W., BOHNERT, H. J., PHILIPPE, H. & LANG, B. F. 2005. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. *Current Biology,* 15**,** 1325-1330.

RODRIGUEZ-EZPELETA, N., BRINKMANN, H., BURGER, G., ROGER, A. J., GRAY, M. W., PHILIPPE, H. & LANG, B. F. 2007. Toward resolving the eukaryotic tree: The phylogenetic positions of jakobids and cercozoans. *Current Biology,* 17**,** 1420-1425.

ROGOZIN, I., BASU, M., CSUROS, M. & KOONIN, E. 2009. Analysis of Rare Genomic Changes Does Not Support the Unikont-Bikont Phylogeny and Suggests Cyanobacterial Symbiosis as the Point of Primary Radiation of Eukaryotes. *Genome Biology and Evolution***,** 99-113.

ROKAS, A. & HOLLAND, P. W. H. 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution,* 15**,** 454-459.

RUIZ-TRILLO, I., INAGAKI, Y., DAVIS, L., SPERSTAD, S., LANDFALD, B. & ROGER, A. 2004. Capsaspora owczarzaki is an independent opisthokont lineage. *Current Biology***,** R946-R947.

SALCEDO, E., CORTESE, J. F., PLOWE, C. V., SIMS, P. F. G. & HYDE, J. E. 2001. A bifunctional dihydrofolate synthetase-folylpolyglutamate synthetase in Plasmodium falciparum identified by functional complementation in yeast and bacteria. *Molecular and Biochemical Parasitology,* 112**,** 239-252.

SHUKLA, O., KAUL, S. & MEHLOTRA, R. 1990. Nutritional studies on Acanthamoeba culbertsoni and development of chemically defined medium. *Journal of Protozoology***,** 237-242.

SIMPSON, A. G. B. & ROGER, A. J. 2002. Eukaryotic evolution: Getting to the root of the problem. *Current Biology,* 12**,** R691-R693.

SIMPSON, A. G. B. & ROGER, A. J. 2004. The real 'kingdoms' of eukaryotes. *Current Biology,* 14**,** R693-R696.

SNEL, B., BORK, P. & HUYNEN, M. 2000. Genome evolution - gene fusion versus gene fission. *Trends in Genetics***,** 9-11.

STECHMANN, A. & CAVALIER-SMITH, T. 2002. Rooting the eukaryote tree by using a derived gene fusion. *Science,* 297**,** 89-91.

STECHMANN, A. & CAVALIER-SMITH, T. 2003. The root of the eukaryote tree pinpointed. *Current Biology,* 13**,** R665-R666.

SUN, X., CROSS, J., BOGNAR, A., BAKER, E. & SMITH, C. 2001. Folate-binding triggers the activation of folylpolyglutamate synthetase. *Journal of Molecular Biology***,** 1067-1078.

TEICHMANN, S. & MITCHISON, G. 1999. Making family trees from gene families. *Nature Genetics***,** 66-67.

WICKSTEAD, B., GULL, K. & RICHARDS, T. 2010. Patterns of kinesin evolution reveal a complex ancestral eukaryote with a multifunctional cytoskeleton. *Bmc Evolutionary Biology***,** -.

WOESE, C. R. 1996. Phylogenetic trees: Whither microbiology? *Current Biology,* 6**,** 1060-1063.

XIAO, B., SHI, G., GAO, J., BLASZCZYK, J., LIU, Q., JI, X. & YAN, H. 2001. Unusual conformational changes in 6-hydroxymethyl-7,8-dihydropterin pyrophosphokinase as revealed by X-ray crystallography and NMR. *Journal of Biological Chemistry***,** 40274-40281.

YANG, Z. H. 1994. MAXIMUM-LIKELIHOOD PHYLOGENETIC ESTIMATION FROM DNA-SEQUENCES WITH VARIABLE RATES OVER SITES - APPROXIMATE METHODS. *Journal of Molecular Evolution,* 39**,** 306-314.

YOON, H., GRANT, J., TEKLE, Y., WU, M., CHAON, B., COLE, J., LOGSDON, J., PATTERSON, D., BHATTACHARYA, D. & KATZ, L. 2008. Broadly sampled multigene trees of eukaryotes. *Bmc Evolutionary Biology*, -.

ZHANG, Y., YANG, S., LIU, M., SONG, C., WU, N., LING, P., CHU, E. & LIN, X. 2010. Interaction between Thymidylate Synthase and Its Cognate mRNA in Zebrafish Embryos. *Plos One*, -.

ZHAO, R. & GOLDMAN, I. 2007. The molecular identity and characterization of a Proton-Coupled Folate Transporter-PCFT; biological ramifications and impact on the activity of pemetrexed. *Cancer and Metastasis Reviews*, 129-139.

APPENDICES

Amino acids	**Conc. (mg/litre)**	**Trace elements**	**Conc. (mg/litre)**
L – Arginine	825	$ZnSO_4.7H_2O$	1
L – Methionine	300	$MnCl_2.4H_2O$	2.3
L – Leucine	900	$(NH_4)_6Mo_7O_{24}.4H_2O$	0.4
L – Isoleucine	600	$CoCl_2$	0.017
L – Valine	700	$CuSO_4.5H_2O$	0.0033
Glycine	1500	H_3BO_3	0.1
L - Lysine. HCl	1250	EDTA	0.01
L – threonine	500		
Salts	**Conc. (mg/litre)**	**Vitamins**	**Conc. (mg/litre)**
$MgSO_4.7H_2O$	985	Biotin	0.25
$CaCl_2.2H_2O$	58.8	B12	0.00125
$(NH_4)_2SO_4FeSO_4.6H_2O$	19.6	Thiamine HCl	1.25
$Na_2HPO_4.2H_2O$	445		
KH_2PO_4	340		
Na Citrate	1000	All reagents obtained from: (Sigma)	
Carbohydrates	**Conc. (mg/litre)**		
Glucose	36000		

FIGURE 13 – DEFINED MEDIA USED BY DR FIONA HENRIQUEZ FOR ACANTHAMOEBA CDNA EXTRACTION

Phylogeny	**Model Name**	**Number of substitution rate categories**	**Gamma distribution parameter (α-parameter)**	**Proportion of invariable sites**
DHNA	LG	8	1.210000	not selected
HPPK	LG	8	0.840000	0.100000
DHPS	LG	8	0.520000	not selected
DHNA-HPPK	LG	8	0.960000	0.050000
DHNA-DHPS	LG	8	0.880000	0.080000
HPPK-DHPS	LG	8	0.860000	0.190000
DHNA-HPPK-DHPS	LG	8	0.950000	0.130000

FIGURE 14 – MODELGENERATOR PARAMETERS

CGCCAGTTTCCGAGCGATGTGGTGAAACGCCCGGTGAGCGCGAACACCACCTATCGCAACTGGAGCAGCCGCGGCGCGGCGGCGCTGGAAGTGGATCC
GCCGGGCTGCAGCCCGATGTGGAACAGCCCGCTGGATCTGATGGTGCAGGCGATTCTGGGCGTGAACAAAGAAGAACGCGTGAAACGCCAGAACATTA
ACATTAGCATTGTGATTTTTCATGATATTAAACAGGCGGCGCTGCATGATGATGTGCGCCATACCATTAACTATAGCCTGGTGCGCAAAGCGGTGGTGGC
GTATACCGAAGATAGCCATCATTATACCCTGGAAAGCCTGTGCAGCGGCATTGCGAAAATTTGCTGCGTGCAGTTTGGCGCGGCGGAAGCGATTGTGCAT
GTGGAAAAACCGTGCGCGCTGAGCCTGGCGCGCTGCCCGGCGGTGGAAGTGCATCGCACCCGCGATTTTTTTCTGGTGCAGGAACCGCTGATGCTGCAG
CAGAGCGCGGCGCAGGCGCAGCATCAGCAGCAGCCGCAGCCGGCGGTGGCGGCGCCGGCGGTGGTGGCGAGCCCGGGCAAACCGAAAGATGATGCG
GCGGGCGTGGCGGAAGAAAGCAGCGGCGGCGCGGCGAAAAAAAGCAACAGCGTGCATACCGCGTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGCG
AAGATAACATTTATAAAGCGCTGAAAGTGCTGGCGGAAAGCTGCGATATTATTAGCACCAGCGCGCTGTATCAGACCCCGCCGGCGTATGTGGTGGATCA
GCCGGCGTTTCTGAACGCGGCGTGCAAAATTCGCACCAAACTGACCCCGGGCGAACTGCTGGATCTGGTGAAAAGCGTGGAACAGCGCCTGGGCCGCAC
CACCGGCGGCATTCGCTTTGGCCCGCGCTGCATTGATGTGGATATTCTGTTTTATGATAGCATTCATGTGCATCGCAGCGATGAAAGCCTGATTATTCCGC
ATAGCCGCATTCCGGAACGCGATTTTGTGCTGGGCCCGCTGAAAGATATTGCGGCGGATTATGTGCATCCGATTCTGAAAAAAACCATTGCGCAGCTGTTT
CATGAACTGCCGCAGCATAAACTGTATCGCGTGACCCCGATTCGCAGCCAGCTGTGGACCTGGAGCGAACGCACCTTTGTGATGGGCATTATTAACGCGA
CCCCGGATAGCTTTAGCGATGGCGGCGATTGCTTTGATATTGAAACCGCGGTGCGCCATGCGAAAGAACTGGTGGAAGGCGGCGCGCATATTCTGGATA
TTGGCGGCCAGAGCACCAACCCGCGCAGCACCCTGCTGAGCGCGGAAGAAGAAGTGAAACGCGTGGTGCCGCTGGTGCAGGCGCTGCGCAGCGAAGC
GAACTGCGCGTGGATGAAAAACATTCCGATTAGCATTGATACCTTTTATAGCAGCGTGGCGGAAGAAGCGATTAAAGTGGGCGCGGATGTGATTAACGA
TATTAGCGGCGGCGTGTATGATGAAAAAATTCTGAGCGTGGCGCATCATTATCAGGTGCCGATTGTGCTGATGCATATGCGCGGCACCCCGCAGACCATG
ATGCAGCCGGTGAACACCGATTATGGCGGCAAAATGCTGGATGAAGTGGCGCGCGTGCTGAAACAGCGCGCGGATAGCGCGCAGCTGGCG

FIGURE 15 – ASSEMBLED PUTATIVE DHNA-HPPK-DHPS CONTIG SEQUENCE

RQXFPSXDVVKXRPVSAXNTTXYRXNWSSRGAAALEXVDPPGCSPMWNSPLDLMVQAILGVNKEERVKRQNINISIVIFHDIKQAALHDDVRHTINYSLVRKAV
VAYTEDSHHYTLESLCSGIAKICCVQFGAAEAIVHVEKPCALSLARCPAVEVHRTRDFFLVQEPLMLQQSAAQAQHQQQPQPAVAAPAVVASPGKPKDDAAGV
AEESSGGAAKKSNSVHTAYLALGSNLGQREDNIYKALKVLAESCDIISTSALYQTPPAYVVDQPAFLNAACKIRTKLTPGELLDLVKSVEQRLGRTTGGIRFGPRCID
VDILFYDSIHVHRSDESLIIPHSRIPERDFVLGPLKDIAADYVHPILKKTIAQLFHELPQHKLYRVTPIRSQLWTWSERTFVMGIINATPDSFSDGGDCFDIETAVRHA
KELVEGGAHILDIGGQSTNPRSTLLSAEEEVKRVVPLVQALRSEANCAWMKNIPISIDTFYSSVAEEAIKVGADVINDISGGVYDEKILSVAHHYQVPIVLMHMRGT
PQTMMQPVNTDYGGKMLDEVARVLKQRADSAQLAX

FIGURE 16 – TRANSLATED DHNA-HPPK-DHPS AMINO ACID SEQUENCE (+1 READING FRAME)

Primer Pair	Sequencing Primer	DNA concentration of sample (ng μl^{-1})	Raw Sequencing Data
FOLB168F-FOLP1662R (6)	M13F	738.5	GNATNNTNGGCGATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATAGATCTGCTGCGTCCAGTTCG GCGCCGCCGAGGCCATCGTCCATGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTGCA GGAGCCGCTCATGCTCCAGCAGAGCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCAAGC CCAAGGACGACGCGGCGGGCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAACC TGGGCCAGCGGGAGGACAACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTACCAGACGCCTCCGGCCTACG TAGTGGACCAGCCGGCCTTCTTGAATGCCGCGTGCAAGATTCGAACGAAGCTCACACCCGGCGAGCTGCTCGACTTGGTCAAGAGCGTCGAACAGAGACTTG GACGCACGACCGGAGGGATCCGCTTTGGTCCGCGGTGCATTGACGTTGACATCCTGTTCTACGACTCGATCCACGTCCACCGCAGCGATGAATCGCTGATCAT CCCCCACTCGAGAATTCCTGAGCGCGACTTCGTGCTGGGTCCGCTGAAGGACATTGCTGCGGACTACGTGCACCCAATTTTGAAAAAGACGATCGCGCAGCTG TTCCATGAGCTGCCGCAGCACAAGCTGTACCGAGTTACGCCCATCCGCTCGCAGCTCTGGACTGGAGCGAACGCACGTTTGTCATGGGCATCATCANGCGAC GCCCGANAGCTTNAGCGATGGTGGCGANTGCTTNNATATCAAAACGCCGTCAGGATGCAAGGAATNGNTTAAGGCGGGCCCAATATGNAATNGGGGGNAT NNACAACCCAGTCANNTTTTTGCNNAGAGAGTNAGNGGTGGGCCCTTTGNACGNTAGAANAGCATGCCCTGATAAAATCCATTCATGACTTNATCTCNTGTA AAGCAAAATGGNCNATTNAAAAANNNGNGGTAAAAAAATTTNGGGTCNNTAAAGGCATNGTTTNATNAGNCCCAAAANNNCGNAAAAANGGAAATNNAA GGCCNTTAAAANATTTTTTTGGNGGATTNNNTTTTTTTTNNNGGGGGNNCTTTTTTNNAANGGGNAAAAATTTTTTNTN
FOLB168F-FOLP1662R (7)	M13F	529	GTGGAATTCGCCCTTCAAGGCCGTCGTGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGNGTCCAG TTCGGCGCCGCCGAGGCCATCGTCCATGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGT GCAGGAGCCGCTCATGCTCCAGCAGAGCGCCGCTCAGGCCCAACACCAGCAACAGNCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCA AGCCCAAGGACNACNCGGCGGGCGTGGCGGAANANAGCAGCGGTGGTGCGGCCAATAAGAGNAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCA ACCTGGGCCAGCGGGAGGACAACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTACCAGACGCCTCCGGCCT ACNTAGTGGATCAGCCGGCCTTCTTGAATGCCGCGTGCAAGATTNGAACGAAGCTCACACCCGGCGAGCTGCTCGACTTGGTCAAGAGCGTCGAACAGAGAC TTGGACGCACGACCGGANGGATCCGCTTTGGTCCGCGGTGCATTGACGTTGACATCNTGTTCTACGACTCGATNCACGTCCACCGCAGCGATGAATCNCTGAT CATCCCCCACTCGAGAATTCCTGANCGCGACTTCGTGCTGGGTCCGCTGAANGACATTGCTGCGGACTACGTGCACCCAATTTTGAAGAANACGATTCGCGCA NCTGTTCCNTGAGCTGCCGCAACANAAGCTGTACCAGTTNNCCTATNCGCTCGCAACTTTTGGANCTTGGANCGAANGCCNNGTTTTTTCTTGGGCATCATTT ANGCGAACGNCCGAAANCTTTCNNGATGGTGGTCGANTNNGGGGANTCNAACAANCNCNNGTTTCNNCTTTGCTNTNGGCNNAGNANAAGGTNANGGG GTNGGTGCCCTTCTTNAAANCCTNATNGGGCCNATNATTTTANCAATCCCGGCGGGGTTANAAAAAAAATTTTTNGGGGNTCCNNTAANGGNCATTTNTNG TANNNTGAANNNCCCAAAAAAATANCTNTNATAAANNGNNAANATTCCAANGNNCCCTTTTAAANNCTTTTTNTGGNAGGTTTTCGGTNNNTNTTTATNCC CNGGGGNCTCTTTTTTTGGTTTCCTGNNTNNNNTTTTATNTTTTTTNAAAAANANGGTGTTANATTTTTCTCTNAAANATAANAGNNTNTTT
FOLB168F-FOLP1662R (8)	M13F	174.6	ATTCGCCCTTCCGCCAGCTGAGCCGAATCGGCTCGCTGCTTCAAAACGCGGGCCACCTCGTCGAGCATCTTACCGCCATAGTCGGTATTGACCGGCTGCATCAT CGTCTGCGGCGTGCCTCACATGTGCATCAGCACGATCGGCACCTGGTAGTGGTGAGCCACCGACAAAATCTTCTCGTCGTACACGCCGCCGCTGATGTCGTTG ATCACATCGGCGCCCACTTTGATGGCTTCCTCAGCAACGGAGGAGTAGAAGGTGTCGATCGAAATGGGGATGTTCTTCATCCAGGCGCAGTTGGCCTCGCTCC TGAGCGNTTGCCAGTATGTGGGCGCCGNCCTCAACCANNCCCTTGGAATGCCTGANNGNGGANTCNATATCNAANNANTNNNCANNATCANNTGAANNT GNNNGNNGNCNNGANNATNATGNNNATGANNNANNNGNNNTNNNTNNANNNNNNNNANNTGNNANNGNANGNNNNNANTCNNANNNNNTNGNNNN GNNNNNNTNANGNANNNNNNTGNNNNANNNNNNNNNNNNNNNNNGNNNGNNNNNNNNNNNNNANNNNGNNNTNNNNNNNNNNNNNNNNNNNNANNN NNNNNNNNNNANNNNNNNNNGNNNNANNANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNANNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
FOLB168F-FOLP1662R (9)	M13R	415.5	GGGCATTACCCTCACTAAAGGGAAACAAAAGCTGGGTACCGGGCCCCCCCTCGAGGTCGACGGTATCGATAAGCTTGATATCCACTGTGGAATTCGCCCTTCC GCCAGCTGAGCCGAATCGGCTCGCTGCTTCAAAACGCGGGCCACCTCGTCGAGCATCTTACCGCCATAGTCGGTATTGACCGGCTGCATCATCGTCTGCGGCG TGCCTCGCATGTGCATCAGCACGATCGGCACCTGGTAGTGGTGAGCCACCGACAAAATCTTCTCGTCGTACACGCCGCCGCTGATGTCGTTGATCACATCGGC GCCCACTTTGATGGCCTCCTCAGCAACGGAGGAGTAGAAGGTGTCGATCGAAATGGGGATGTTCTTCATCCAGGCGCAGTTGGCCTCGCTCCTGAGCGCTTG CACGAGGGGCACCACCCGCTTGACCTCCTCCTCGGCCGAGAGCAGCGTGGACCTGGGGTTGGTCGACTGCCCGCCAATGTCCAGTATGTGGGCGCCGCCCTC AACCAGCTCCTTGGCATGCCTGACGGCGGTCTCGATATCGAAGCAGTCGCCACCATCGCTGAAGCTGTCGGGCGTCGCGTTGATGATGCCCATGACAAACGT GCGTTCGCTCCAGGTCCAGAGCTGCGAGCGGATGGGCGTAACTCGGTACAGCTTGTGCTGCGGCAGCTCATGGAACAGCTGCGCGATCGTCTTCTTCAAAAT TGGGTGCACGTAGTCCGCANCAATGTCCTTCAGCGGACCCAGCACGAAGTCGCGCTCAGGAATTCTCGAGTGGGGGATGATCAGCGATTCATCGCTGCGGTG GACGTGGATCNAGTCNTANAACAGGATGTCAACGTCAATGCACCGCGGANCAAANCGGATCCCTCCGGTCGTGNGTCCAANTNTCTGTTCNANNCTCTTGA CCAANTCNAGNANTCNNCGGGNGTGANNTTNNTTNNAATCTTGNNNNNNNNATTNNANAANGNNNGNTGATCNNTANNNANNNNNNANNNNNNNNGGN AAANGNNNAANNNNTNANAANNNNNANAANNNNNNNNNNANNTNNNNNNNNTNNNNANNTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAA NNNNNNNNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
FOLB168F-FOLP1662R (10)	M13R	553.7	GGCATTACCCTTCACTAAAGGGACAAAAGCTGGGTACCGGGCCCCCCCTCGAGGTCGACGGTATCGATAAGCTTGATATCCACTGTGGAATTCGCCCTTCCGC CAGCTGAGCCGAATCGGCTCGCTGCTTCAAAACGCGGGCCACCTCGTCGAGCATCTTAGTGGTGAGCCACCGACAAAATCTTCTCGTCGTACACGCCGCCGCT GATGTCGTTGATCACATCGGCGCCCACTTTGATGGCCTCCTCAGCAACGGAGGAGTAGAAGGTGTCGATCGAAATGGGGATGTTCTTCATCCAGGCGCAGTT GGCCTCGCTCCTGAGCGCTTGCACGAGGGGCACCACCCGCTTGACCTCCTCCTCGGCCGAGAGCAGCGTGGACCTGGGGTTGGTCGACTGCCCGCCAATGTC CAGTATGTGGGCGCCGCCCTCAACCAGCTCCTTGGCATGCCTGACGGCGGTCTCGATATCGAAGCAGTCGCCACCATCGCTGAAGCTGTCGGGCGTCGCGTT GATGATGCCCATGACAAACGTGCGTTCGCTCCAGGTCCAGAGCTGCGGCGGATGGGCGTAACTCGGTACAGCTTGTGCTGCGGCAGCTCATGGAACAGCTGC GCGATCGTCTTCTTCAAAATTGGGTGCACGTAGTCCGCAGCAATGTCCTTCAGCGGACCCAGCACGAAGTCGCGCTCAGGAATTCTCGAGTGGGGGATGATC AGCGATTCATCGCTGCGGTGAACGTGGATCGAGTCGTAAAACAGGATGTCAACGTCAATGCACCGCGGACCAAAGCGGATCCCTCCGGTCGTGCGTCCAAGT CTCTGTTCGACGCTCTNGACCAAGTCGAGCAGCTCGCCGGGTGTGANCTTCGTTCGAATCTTGCACGCGGCATTCAAAAAGNCGGCTGATCNACTACNTAGN CNGAAGCGTCTGGTAAANGNCCANGTGCTGANAAGTCCCAAGAATCGGCAAGAACTTGANCGCNTNGTAANGTNGTCTCCGCTGGCCAAGTTGCTGCCAAC NCNAAAGNCNNGGNACCTGTTGNTTTTTGGCGNACANGTGTTTTTTCGCAGCCGCNGGTNTCTNGGTTGNGGANTCCNCACCCGACCCANGCNTTNGTTTT TNGNTGNCTAANNNNTTTTGAAANANGTCGNCAAAAATCGGGGGGACNNNGGGNTNNNNGNANGNNGTTNNNNAANNNNNNCAANANNAATTNNNNA AAAANNGNNGGNNNNNNNANTNANANNTNNNNGNNNNNNNNNNNNNTNNNNANNNNNNTNNNGNNNNNNNNNNNNNNNNNNNNTTNNNNNNNNNN NNNNNNNNNNNCTCNTNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

FOLB168F-FOLP1662R (11)	M13F	543.3	GAAATNTATGGGCGAATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTCAAGGTCGTCG TGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGCGTCCAGTTCGGCGCCGCCGAGGCCATCGTCCAT GTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAGCAGA GCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCAAGCCCAAGGACGACGCGGCAGGCGT GGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGACAACAT CTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTACCAGACGCCTCCGGCCTACGTAGTGGATCAGCCGGCCTTCTTG AATGCCGCGTGCAAGATTCGAACGAAGCTCACACCCGGCGAGCTGCTCGACTTGGTCAAGAGCGTCGAACAGAGACTTGGACGCACGACCGGAGGGATCCG CCTTGGTCTCNAGAATTCCTGANCGCGACTTCNTGCTGGGTCCGCTGAAGGAATTGCTGCGGACTACNTGCACCCAATTTTGAANAANANNATNNNNCANTG TTNCATGANTGNNNNANNANAANNTGTANNANTTANNCCATCCNNTCNNANTNTGGANTGNANNNAANNNNTTGTNTGGNNTNTNANNNNANNCCNA AANNTTNNNNATGNGNNANTGNTTNANTNNNNANNNNNNGNNTNNNNNANTNGNTNANNNNNNNNNNTNNGGAANTNNNNNNNNNNNNANNNN NNNNNNNNNTTTNNNNANANANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
FOLB168F-FOLP1662R (12)	M13F	264.6	GNAATNTNTGGGCGAATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTCAAGGCCGTCG TGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGCGTCCAGTTCGGCGCCGCCGAGGCCATCGTCCAT GTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAGCAGA GCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCAAGCCCAAGGACGACGCGGCGGGCGT GGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGACAACAT CTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTACCAGACGCCTCCGGCCTACGTAGTGGATCAGCCGGCCTTCTTG AATGCCGCGTGCAAGATTCGAACGAAGCTCACACCCGGCGAGCTGCTCGACTTGGTCAAGAGCGTCGAACAGAGACTTGGACGCACGACCGGAGGGATCCG CTTTGGTCCGCGGTGCATTGACGTTGACATCCTGTTCTACGACTCGATCCACGTCCACCGCAGCGATGAATCGCTGATCATCCCCCACANGATCGCGCAGCTGT TCCATGAGCTGCCGCANCACAAGCTGTACCGAGTNACGCCCATCCGCTCGCAGCTCTGGACTGGAGCGAANGCACGTTTGTCATGGGCATCATCAANGCGAC GCCCGANAGNTNNNCGATGGTGGNANTGCTTCATATCAAAACGCCGTCAGGATGCNAGGAATTGNTNANGGNGGNCCNNATATGNAATTGGGGGANTNA ANANCCAGTCAGNTGTTTNGNNAAAAGAGTAAGGGNGGGGCCTNNGNANNNTANAANAGNCANTGNCTGNANAAAATCCATTTATNAACTTTATCTNNT NTNAANCNNAANGGNCNANTTAAANTNNGGGGGTAAAAAATTNNGGGNNNNNNNNNNNNNNNNNATNNNNNNAAAAANNNNNAANAATGNAANNN AANGCNTTNAANNNTTNTTNGNNNNTNNNNNTTNNNNNNGGGNNNNTTTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN NNNNNNNNGNATNNTNGGCGATTGGAGCTCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTCAAGG CCGTCGTGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGCGTCCAGTTCGGCGCCGCCGAGGCCAT CGTCCATGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCC AGCAGAGCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCAAGCCCAAGGACGACGCGGC GGGCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGA CGACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTACCAGACGCCTCCGGCCTACGTAGTGGATCAGCCGGCC TTCTTGAATGCCGCGTGCAAGATTCGAACGAAGCTCACACCCGGCGAGCTGCTCGACTTGGTCAAGAGCGTCGAACAGAGACTTGGACGCACGACCGGAGG GATCCGCTTTGGTCCGCGGTGCATTGACGTTGACATCCTGTTCTACGACTCGATCCACGTCCACCGCAGCGATGAATCGCTGATCATCCCCCACTCGAATCGCG CAGCTGTTCCATGAGCTGCCGCANCACAAGCTGTACCGAGTTACGCCCATCCGCTCGCAGCTCTGGACTGGAGCGAACGCACGTTTGTCATGGGCATCATCAA NGCGACGCCCGANAGNTNANCGATGGTGGCGACTGCTTCAATATCAAAACGCCGTCAGNATGCAAGGAANTGNTNAAGGGGGNCCCNNTATTGGAATNG GGGGNATCGACAACCCAGTCANGNTGTTTGGCNAAAAGAGTNAGGGGNGGNCCTNTGAANGTNAGAANAGCATGNCTGATNAAATTCCNTTNTTAACTTT TTTNNNTNTNAAGCTNANNGGNCNATNTAAATNNGGGGGTAAAAAATTTNGGGTCCNANNGNATNNNNAANGNGNCCNAAAANNCGNAANATNGNAA TNNANGNCNTTAANNNNTTTTNNGNGNTTNGNANTTTTNNCNGGGGNNNTTTTTTNNNNNGGNNNNNTTNNTTNNNNANNNAANNNNNAAANNNNN NNNNAAAAAAAAAAATTTTTNNNNTTTTTTTTNNNTN
FOLB168F-FOLP1662R (13)	M13F	578.2	GNATNNTNGGCGATTGGAGCTCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTCAAGGCCGTCGTGG CCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGCGTCCAGTTCGGCGCCGCCGAGGCCATCGTCCATGTC GAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAGCAGAGCG CCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCAAGCCCAAGGACGACGCGGCGGGCGTGGC GGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGACGACATCTAC AAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTACCAGACGCCTCCGGCCTACGTAGTGGATCAGCCGGCCTTCTTGAATG CCGCGTGCAAGATTCGAACGAAGCTCACACCCGGCGAGCTGCTCGACTTGGTCAAGAGCGTCGAACAGAGACTTGGACGCACGACCGGAGGGATCCGCTTT GGTCCGCGGTGCATTGACGTTGACATCCTGTTCTACGACTCGATCCACGTCCACCGCAGCGATGAATCGCTGATCATCCCCCACTCGAGAATTCCTGAGCGCGA CTTCGTGCTGGGTCCGCTGAAGGACATTGCTGCGGACTACGTGCACCCAATTTTGAAAAAAACGATCGCGCAGCTGTTCCATGAGCTGCCGCANCACAAGCTG TACCGAGTTACGCCCATCCGCTCGCAGCTCTGGACTGGAGCGAACGCACGTTTGTCATGGGCATCATCAANGCGACGCCCGANAGNTNANCGATGGTGGCG ACTGCTTCAATATCAAAACGCNAAAAGAGTNAGGGGNGGNCCTNTGAANGTNAGAANAGCATGNCTGATNAAATTCCNTTNTTAACTTTTTTNNNTNTNAA GCTNANNGGNCNATNTAAATNNGGGGGTAAAAAATTTNGGGTCCNANNGNATNNNNAANGNGNCCNAAAANNCGNAANATNGNAATNNANGNCNTTA ANNNNTTTTNNGNGNTTNGNANTTTTNNCNGGGGNNNTTTTTTNNNNNGGNNNNNTTNNTTNNNNANNNAANNNNNAAANNNNNNNNNAAAAAAAA AAATTTTTN
FOLB168F-FOLP1662R (14)	M13R	484.9	GGCATTANCCTACTAAAGGGAACAAAAGCTGGGTACCGGGCCCCCCCCTCGAGGTCGACGGTATCGATAAGCTTGATATCCACTGTGGAATTCGCCCTTCCGCC AGCTGAGCCGAATCGGCTCGCTGCTTCAAAACGCGGGCCACCTCGTCGAGCATCTTACCGCCATAGTCGGTATTGACCGGCTGCATCATCGTCTGCGGCGTGC CTCGCATGTGCATCAGCACGATCGGCACCTGGTAGTGGTGAGCCACCGACAAAATCTTCTCGTCGTACACGCCGCCGCTGATGTCGTTGATCACATCGGCGCC CACTTTGATGGCCTCCTCAGCAACGGAGGAGTAGAAGGTGTCGATCGAAATGGGGATGTTCTTCATCCAGGCGCAGTTGGCCTCGCTCCTGAGCGCTTGCAC GAGGGGCACCACCCGCTTGACCTCCTCCTCGGCCGAGAGCAGCGTGGACCTGGGGTTGGTCGACTGCCCGCCAATGTCCAGTATGTGGGCGCCGCCCTCAAC CAGCTCCTTGGCATGCCTGACGGCGGTCTCGATATCGAAGCAGTCGCCACCATCGCTGAAGCTGTCGGGCGTCGCGTTGATGATGCCCATGACAAACGTGCG TTCGCTCCAGGTCCAGAGCTGCGAGCGGATGGGCGTAACTCGGTACAGCTTGTGCTGCGGCAGCTCATGGAACAGCTGCGCGATCGTCTTCTTCAAAATTGG GTGCACGTAGTCCGCAGCAATGTCCTTCAGCGGACCCAGCACGAAGTCGCGCTCAGGAATTCTCGAGTGGGGGATGATCAGCGATTCATCGCTGCGGTGAAC GTGGATCGAGTCGTAGAACAGGATGTCAACGTCAATGCACCGCGGACCAAAGCGGATNCCTCCGGTCGTGCGTCCAAGTCTCTGTTCGACGCTCTTGACCAA GTCGAGCANCTCGCCGGGTGTGANCTTCGTTCGAATCTNGCACGCGGCATTCAANAAGGCCGGCTGATCCACTACGTANGCCGGANGCGTCTGGTAAAAGG CCAAGTGCTGATNATGTCCCAGGANTCGNCAGAACTNGAACGCCTTGTAANGTNGTCTCCGCTGNCNAGTTGNTNCAACGCNAANAGCNTGNNACCTNTGT NTTTTGGCGACACGNTGTTTTTCGCAGCCGCNGTTTCTTGGTTGCGGAATCCACCCCCGACCCAATCCNCTGGNTTTTTGNTTGCNNANGCCTTTNGAAAAAN GNTCCNCAAAAAANCGGGGGGANTNANGGGGNTNGNAANANGNAGNTNTANNAANGTCNGGNCAAAANAAATTNNNNAAAAAANNGGNGGGNNCNN NNNNNNNTNNAANATTTTNNGGNAATNNNGNNGGNNNNNAAAAAANNNGNTTTTNGNNANNNNTNNNNCNNNNAAAAAAAN
FOLB168F-FOLP1662R (15)	M13R	786.7	NNGGCATTANCCTACTAAAGGGAACAAAAGCTGGGTACCGGGCCCCCCCTCGAGGTCGACGGTATCGATAAGCTTGATATCCACTGTGGAATTCGCCCTTCC GCCAGCTGAGCCGAATCGGCTCGCTGCTTCAAAACGCGGGCCACCTCGTCGAGCATCTTACCGCCATAGTCGGTATTGACCGGCTGCATCATCGTCTGCGGCG TGCCTCGCATGTGCATCAGCACGATCGGCACCTGGTAGTGGTGAGCCACCGACAAAATCTTCTCGTCGTACACGCCGCCGCTGATGTCGTTGATCACATCGGC GCCCACTTTGATGGCCTCCTCAGCAACGGAGGAGTAGAAGGTGTCGATCGAAATGGGGATGTTCTTCATCCAGGCGCAGTTGGCCTCGCTCCTGAGCGCTTG CACGAGGGGCACCACCCGCTTGACCTCCTCCTCGGCCGAGAGCAGCGTGGACCTGGGGTTGGTCGACTGCCCGCCAATGTCCAGTATGTGGGCGCCGCCCTC AACCAGCTCCTTGGCATGCCTGACGGCGGTCTCGATATCGAAGCAGTCGCCACCATCGCTGAAGCTGTCGGGCGTCGCGTTGATGATGCCCATGACAAACGT GCGTTCGCTCCAGGTCCAGAGCTGCGAGCGGATGGGCGTAACTCGGTACAGCTTGTGCTGCGGCAGCTCATGGAACAGCTGCGCGATCGTCTTCTTCAAAAT TGGGTGCACGTAGTCCGCAGCAATGTCCTTCAGCGGACCCAGCACGAAGTCGCGCTCAGGAATTCTCGAGTGGGGGATGATCAGCGATTCATCGCTGCGGTG GACGTGGATCGAGTCGTAGAACAGGATGTCAACGTCAATGCACCGCGGACCAAAGCGGATCCCTCCGGTCGTGCGTCNAGTCTCTGTTCGACGCTCTTGACN AGTCGAGCAGCTCGCCGGGTGTGAGCTTCGTTCGANTCTNGCACGCGGCATTCAANAAGGCGGCTGATCACTACNTAGNCNGAAGCGTCTGGTAAAGGGCC

			AAGTGCTGANATGTCCCAGANTCGNCAGAACTNGACGCNTGTAANGTTGTCCTCCCCTGGCCAGTTGTGCCAACNCAANAGCNNGNGACCTGTGTNTTTNG GCGACACGNTGTTTTTCGCAGCCGCGGTNTCTNGGTTGCGGATCCACCCCCGACCNATCCNNTGNTTTTTGGTTGCCNANGCTTTTGAANANGTNCGAAAA AANCGGGGGGGACNAGGGGANGGAGNAGNAGTTNAAGAAGCNGGNCAAGANAAATGAACNAAANACNGGNGGGACNNNANAANTTANATTTGNNCG GAATTAGCCNGNNNNCNANANNNNGGTTNGGGAGNNNTTTCTTNTNANCTTTATNNTTTTTNTTTTTT
FOLB18F-FOLP1662R (16)	M13F	628.8	NGATNTNTGGGCGAATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTGGATCTGATGGT GCAGGCCATCCTGGGCGTCAACAAGGAGGAACGCGTGAAGCGGCAGAACATCAACATCAGCATCGTCATCTTCCACGACATCAAGCAGGCCGCCCTACACGA CGACGTGCGGCACACCATCAACTACAGCCTTGTCCGCAAGGCCGTCGTGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATC GCCAAGATCTGCTGCGTTCAGTTCGGCGCCGCCGAGGCCATCGTCCACGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACC GCACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAGCAGAGCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGG CGGTGGTGGCGAGTCCCGGCAAGCCCAAGGACGACGCGGCGGGCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGG CCTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGACAACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCC TCTAAGGGCGAATTCCACAGTGGATATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGC GCGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGGAGCATAAAGTGTAANGCCTGG GGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGNCCGCTTTNCAGTCGGGAAANCTGTCGTGCCAGCTGCATTAATGANTCGGCNAN GCGCGGGAAAAGGCGGTTTGCGTATNGGGGGNTNTTCGCTTCTCGTTAATGAATNGTTGNNCTNGGTCGTTCGGTGNGGGNANGGTATAANTNATNNAAG GGGNNAAACGNTNTCNNAAATCGGGGANAANCAGNAAAAANTGGNANNAAAGNCANAAAGGCNGAACGTAAANGNNNGTNNTGGGTTTTCAAGGTCN CCCCTGAANNATNAAAATNNNNTANTNAAGGGGAACCNNGGATAAAAAAANGGTTNCCGNAANCNNGGGNTTTTTTAACGGNTTANNAANNNNCTTTTN TTNGANGGGNTTTAANNNNNGANNNNTGGNNGNTNCNNGGGGNGNANCCNTNCNNNNNTTTNNTTTTTNNNAAANTTNNNNNNNAAAAAANNNGGN NTTNGNNNNAAAAANNNNNNNNAANNNNNNNNNTTNAAAAANTNNANANATATTTTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
FOLB18F-FOLP1662R (17)	M13F	474.9	NATNTNTGGGCGAATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTGGATCTGATGGTG CAGGCCATCCTGGGCGTCAACAAGGAGGAACGCGTGAAGCGGCAGAACATCAACATCAGCATCGTCATCTTCCACGACATCAAGCAGGCCGCCCTACACGAC GACGTGCGGCACACCATTAACTACAGCCTTGTCCGCAAGGCCGTCGTGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCG CCAAGATCTGCTGCGTTCAGTTCGGCGCCGCCGAGGCCATCGTCCACGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCG CACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAGCAGAGCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGC GGTGGTGGCGAGTCCCGGCAAGCCCAAGGACGACGCGGCGGGCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGC CTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGACAACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCT CTAAGGGCGAATTCCACAGTGGATATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCG CGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGGAGCATAAAGTGTAAAGCCTGGG GTGCCTAATGAGTGAGCTAACTCACATTAATNGCGTTGCGCTCACTGNCCGCTTTNCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGANTCGGCNANG CGCGGGGANAGGCGGTTTGNGTATTGGGNGCTNTTCGCTTCTCGTTNATGATNGNTGNNCTNGGTNGTTCGGTGNGGGANGGNATCANTCATNNAAGGG GTAAACGNTNNCNNNANTNGGGNAAANCAGNAAAAATGGGACAAAGGCACAANGCAGANCGAAAGGCNGTNTTGGNTTTCNAGGTCCCCCTGAAAATNA AAATNANTTANTNANGGGNANCNANGATAAAAANNGNTTCCGNAANCTNGGNTTTTTNANCNNNTAANAANTNCTTTTTTNGANGGGNTTNAANNNGNN NNNNTNGGNGNTNCNGGGGGNANCCNNCNNNNNTTTNTTTTTNNNNAAATTNNNNNNNAAAAANNGGGNNNNGGNNNNAAANNNNNNNNNNNNN NNNNNNNNNNNANNNGGAAANTTTTTATTNNNNNNNNNNNNNNNNNTN
FOLB18F-FOLP1662R (18)	M13F	669.4	NATNTNTNGGCGAATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTGGATCTGATGGTG CAGGCCATCCTGGGCGTCAACAAGGAGGAACGCGTGAAGCGGCAGAACATCAACATCAGCATCGTCATCTTCCACGACATCAAGCAGGCCGCCCTACACGAC GACGTGCGGCACACCATCAACTACAGCCTTGTCCGCAAGGCCGTCGTGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCG CCAAGATCTGCTGCGTTCAGTTCGGCGCCGCCGAGGCCATCGTCCACGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCG CACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAGCAGAGCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGC GGTGGTGGCGAGTCCCGGCAAGCCCAAGGACGACGCGGCGGGCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGC CTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGACAACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCT CTAAGGGCGAATTCCACAGTGGATATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCG CGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGGAGCATAAAGTGTAANGCCTGGG GTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGNCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGANTCGGCNANGC GCGGGGAAAAGNGGTTTGNGTATTGGGNGNTNTNCGNTTCTCGTTAATGATNGNTGNNTNGGTGTTCGGTGNGGGAGGGTANANNTATNAAGGGGGAA ANGNTNTCNNAAATNGGGANACNAGAAAAAATGGANNAAGGCNNAAGGCAGAACGNAAGGCNGTTGTGGTTTTNAAGNTCNCCCTGAAAATNAAATNN NTTANTAANGGGAACNAGNATNAAAAAGGNTCCCGNAATCNNGGGTTTTTAANNNNTANNAAGNCNTTTNTTGANGGGTTTAANNNNNNNNNNNTGGNN NTNCNNGGGGGNANCCTNCNNNNTTNNNTTTNNNNAANTTNNNNNNNAAAAANNNGNNTTGGNNNAAANTNNNNNNANNNNNNNNNTNNNAATTNN ATNTTNATTTNTTTTTNNNNNNNNNNNNNNNNNNNNNN
FOLB18F-FOLP1662R (19)	M13R	620.7	GGGCATTANCCTCACTAAAGGGAACAAAAGCTGGGTACCGGGCCCCCCCTCGAGGTCGACGGTATCGATAAGCTTGATATCCACTGTGGAATTCGCCCTTAG AGGGCCGAGGTGCTGATGATGTCGCAGGACTCGGCCAGGACCTTGAGCGCCTTGTAGATGTTGTCCTCCCGCTGGCCCAGGTTGCTGCCCAGCGCCAGATAG GCCGTGTGCACGCTGTTGCTCTTCTTGGCCGCACCACCGCTGCTCTCTTCCGCCACGCCCGCCGCGTCGTCCTTGGGCTTGCCGGGACTCGCCACCACCGCCGG AGCCGCGACTGCCGGCTGTGGCTGTTGCTGGTGTTGGGCCTGAGCGGCGCTCTGCTGGAGCATGAGCGGCTCCTGCACCAGGAAGAAGTCGCGGGTGCGGT GGACCTCGACGGCGGGGCATCGGGCGAGGGAGAGGGCGCAGGGCTTCTCGACGTGGACGATGGCCTCGGCGGCGCCGAACTGAACGCAGCAGATCTTGGC GATGCCGGAGCAGAGGGACTCGAGGGTGTAGTGGTGCGAGTCCTCGGTGTAGGCCACGACGGCCTTGCGGACAAGGCTGTAGTTGATGGTGTGCCGCACG TCGTCGTGTAGGGCGGCCTGCTTGATGTCGTGGAAGATGACGATGCTGATGTTGATGTTCTGCCGCTTCACGCGTTCCTCCTTGTTGACGCCCAGGATGGCCT GCACCATCANATCCAAGGGCGAATTCCACATTGGGCTGCANCCCGGGGGATCCACTANTTCTAGAGCGGCCGCACCGCGGGAGCTCCAATTCNCCCTATANT GAGTCGTATTACNCGCGCTCACTGGCCGTCNTTTTACAACGTCNTGACTGGGAAAANCCTGGCGTTACCCAACTNAATCNCCTTGNANNACATCCCCCTTTCN CCANNTGGNGTAANANNNAAAANGCCCNCACCNATNAANTTTGGTCATGAAATTATCAAAAAGNATNTTCNCCTAAATCNTTTTAANTTAAAATGAANTTTT AAATNANTNTAANNANNTNTAANNAAATTGGNNTGAANNNNNAAAAANTCNNNAAAANNNNAAAAANGNNANNNNTNNNANNNNNANNGNNANNC NNNAANNNNAANAAANNNNNNNCCNTTNCCNCNAATTTTNNNNAANTNNNNNAANCAANNTNTNNNNNAAANNNNCNCNNNCNNNNNNNNANNNN AAAANNNAAAANNNNNTTTCNNNNANNTNNNAANNNNNNNNNNNNNNAAAAANTNNNNNNNNNNNNNNNTNNNNNNNAAANNNNNNNNNNNCN NNNTTTNNNNANNNNNNAAAANNNNNNNNAANNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNANNNNNNNNNNNNNNNNNNNNNTTNNNNNNNN AAAANNNNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTNNTATTTTTNNNNNA TTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
FOLB18F-FOLP1662R (20)	M13R	575.5	GGGCATTANCCTCACTAAAGGGAACAAAAGCTGGGTACCGGGCCCCCCCTCGAGGTCGACGGTATCGATAAGCTTGATATCCACTGTGGAATTCGCCCTTAG AGGGCCGAGGTGCTGATGATGTCGCAGGACTCGGCCAGGACCTTGAGCGCCTTGTAGATGTTGTCCTCCCGCTGGCCCAGGTTGCTGCCCAGCGCCAGATAG GCCGTGTGCACGCTGTTGCTCTTCTTGGCCGCACCACCGCTGCTCTCTTCCGCCACGCCCGCCGCGTCGTCCTTGGGCTTGCCGGGACTCGCCACCACCGCCGG AGCCGCGACTGCCGGCTGTGGCTGTTGCTGGTGTTGGGCCTGAGCGGCGCTCTGCTGGAGCATGAGCGGCTCCTGCACCAGGAAGAAGTCGCGGGTGCGGT GGACCTCGACGGCGGGGCATCGGGCGAGGGAGAGGGCGCAGGGCTTCTCGACGTGGACGATGGCCTCGGCGGCGCCGAACTGAACGCAGCAGATCTTGGC GATGCCGGAGCAGAGGGACTCGAGGGTGTAGTGGTGCGAGTCCTCGGTGTAGGCCACGACGGCCTTGCGGACAAGGCTGTAGTTGATGGTGTGCCGCACG TCGTCGTGTAGGGCGGCCTGCTTGATGTCGTGGAAGATGACGATGCTGATGTTGATGTTCTGCCGCTTCACGCGTTCCTCCTTGTTGACGCCCAGGATGGCCT GCACCATCAGATCCAAGGGCGAATTCCACATTGGGCTGCAGCCCGGGGGATCCACTAGTTCTAGAGCGGCCGCACCGCGGGAGCTCCAATTCGCCCTATAGT GAGTCGTATTACGCGCGCTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAANCCTGGCGTTACCCAACTTAATCGCCTTGCANCACATCCCCCTTTCGC CAGCTGGCGTAATAACGAAAAAGCCCGCACCGATNAATTTTGGTCATGAAATATCAAAAAGGATCTTCACCTAAATCNTTTNAATTAAAAATGAAGTTTTAATC AATCTAAAGTATATATGANTNAAACTNGGTCTGANANTCAAAAAANTCNTCAAAANGCNANAAAAGGNATGCCCTNCNANTCGGAACGGNAANCCNNAAA NANNANAAACGGTCACCCATTCCCNCCANTTNTNANNAAATNNGGGNACCAANCTTTTTCTGAAAGGGNCCCCANNCCNCCGGCNNATNAAAANCCAAAA ANGNCNTTTNCCNNNANTTNGAAAAAGNNNCCNTGGTAAAAAAATTNNCCNNGNNNNTNNTTNANNGNAAAATTNNGGGGNACCNNNTTTTTNNAANN NNNNAAAAANTTNNCAAAANNNNNNNNANTTTTNNGNNANNNNNNNNAAAAANNCCCCNTTNNNNANANTTTNNNNNNANAAAAAANNNNNNNAAAN

			NNTNTTNNAAAANAANNNNNNNNNNNNNNTTTTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTTTTTTTTTTTTTTTTTTNTTTTNTNNNNNNNNNNNNNNNN
FOLB168F-FOLP620R (6)	M13F	315.6	GGGAGACTAGTATAGGGCGATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTCAAGGCC GTCGTGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGCGTCCAGTTCGGCGCCGCCGAGGCCATCG TCCATGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAG CAGAGTGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCAAGCCCAAGGGCGACGCGGCGG GCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGACA ACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTAAGGGCGAATTCCACAGTGGATATCAAGCTTATCGATACC GTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATT GTTATCCGCTCACAATTCCACACAACATACGAGCCGGGAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTC ACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCT CGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCACGGGATAACGCAG AAGAACATGTGAGCAAAAGGCCAGCAAAGGCCAGGACCGTAAAAAGCCGCGTTTGCTGGCGTTTTTCCATAGCTCGCCCCCTGACGAGCATCACAAAAATCG ACGCTCAAGTCAGAGTGCGAAACCCGGACAGAACTATAGAATACAAGCTTTCCCCTGGAAGCTCTCGTGCGCCTCTCCTGTTCGACCCTTGCCCGCTTAACCGG AATACCGTGTC
FOLB168F-FOLP620R (8)	M13F	354.2	TGGAATTCGCCCTTCAAGGCCGTCGTGGCCTACACCGAGGACTCGCACCATTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGCGTCCAGTT CGGCGCCGCCGAGGCCATCGTCCATGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTG CAGGAGCCGCTCATGCTCCAGCAGAGCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCAA GCCCAAGGACGACGCGGCGGGCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAA CCTGGGCCAGCGGGAGGACAACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTAAGGGCGAATTCCACAGTG GATATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCA TAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGGAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAAC TCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCG TATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCAC AGAATCAGGGATAACGCAGAAGAACATGTGAGCAAAGGCCAGCAAAGGCCAGGAACCGTAAAAAGCCGCGTTTGCTGGCGTTTTTCATAGCTCCGCCCCCCT GACGAGCATCACAAAAATTCGACGCTCAAGTCAGAGTGCGAAACCGACCGAACTATAAGAATACAGCGTTCCATGAAGCTCCTGTGGCCTCTCTGTCCGACCC TGGCCCGCTTACGGAATATCCGT
FOLB168F-FOLP620R (9)	M13F	361.3	TGGCGACTATATAGGGCGATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTCAAGGCCG TCGTGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGCGTTCAGTTCGGCGCCGCCGAGGCCATCGT CCACGTCGAGAAGCCCTGCGCCCTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAG CAGAGCGCCGCTCAGGCCCAACACCAGCAACAGCCACAGCCGGCAGTCGCGGCTCCGGCGGTGGTGGCGAGTCCCGGCAAGCCCAAGGACGACGCGGCGG GCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAACCTGGGCCAGCGGGAGGACA ACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTAAGGGCGAATTCCACAGTGGATATCAAGCTTATCGATACC GTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATT GTTATCCGCTCACAATTCCACACAACATACGAGCCGGGAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTC ACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCC TCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCA GAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGCCGCGTTGCTGGCGTTTTTCCATAGCTCGCCCCCTGACGAGCATCACAAAAA TCGACGCTCAAGTCAGAGGGTGGCGAAACCCGGACAGAACTATAAGAATACCAGGCGTTTCCCCTTGGAGCTTCTTCTGCGCCTCCTCTGTTCGACCCTGACC GCCTTAACCGGAATACCTGGTC
FOLB168F-FOLP620R (10)	M13F	274.1	GGAGACTCTATAGGGCGATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTCAAGGCCGT CGTGGCCTACACCGAGGACTCGCACCACTACACCCTCGAGTCCCTCTGCTCCGGCATCGCCAAGATCTGCTGCGTCCAGTTCGGCGCCGCCGAGGCCATCGTC CATGTCGAGAAGCCCTGCGCCTTCTCCCTCGCCCGATGCCCCGCCGTCGAGGTCCACCGCACCCGCGACTTCTTCCTGGTGCAGGAGCCGCTCATGCTCCAGCA GAGCGCCGCTCGCGGCGGGCGTGGCGGAAGAGAGCAGCGGTGGTGCGGCCAAGAAGAGCAACAGCGTGCACACGGCCTATCTGGCGCTGGGCAGCAACCT GGGCCAGCGGGAGGACAACATCTACAAGGCGCTCAAGGTCCTGGCCGAGTCCTGCGACATCATCAGCACCTCGGCCCTCTAAGGGCGAATTCCACAGTGGAT ATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAG CTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGGAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCA CATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTAT TGGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGGTATCCAC AGAATCAGGGATAACGCAGAAGACATGTGAGCAAAAGGCCAGCAAAGGCCAGGACCGTAAAAGCCGCGTTTGCTGGCGTTTTCATAGCTCGCCCCCTGACG AGCATCACAAAATCGACGCTCAGTCAGAGTTGCGAAGCGGACCGACTATAAGATACAAGCGTTTGCCATGAGCTCCTCGGTGCGCCTCCCTGATCCGAACTGG CCCGCTTTACCGAATACCTGTGATCG
FOLK623F-FOLK1151R (11)	M13F	345.5	GGGAGACTCCTATAGGGCGATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTGGGCGTC GCGTTGATGATGCCCATGACAAACGTGCGTCCGCTCCAGGTCCAGAGCTGCGAGCGGATGGGCGTAACTCGGTACAGCTTGTGCTGCGGCAGCTCATGGAAC AGCTGCGCGATCGTCTTCTTCAAAATTGGGTGCACGTAGTCCGCAGCAATGTCCTTCAGCGGACCCAGCACGAAGTCGCGCTCAGGAATTCTCGAGTGGGGG ATGATCAGCGATTCATCGCTGCGGTGGACGTGGATCGAGTCGTAGAACAGGATGTCAACGTCAATGCACCGCGGACCAAAGCGGATCCCTCCGGTCGTGCGT CCAAGTCTCTGTTCGACGCTCTTGACCAAGTCGAGCAGCTCGCCGGGTGTGAGCTTCGTTCGAATCTTGCACGCGGCATTCAAGAAGGCCGGCTGATCCACTA CGTAGGCCGGAGGCGTCTGGTAGAGGGCCGAGGTGCTGATGAAAGGGCGAATTCCACAGTGGATATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGG CCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCC ACACAACATACGAGCCGGGAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCG GGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGC GCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCACGGGATAACGCAGGAAGACATGTGAGCAAA AGGCCAGCAAAAGGCCAGGACCGTAAAAGGCCGCGTTGCTGCGTTTTTCCATAGGCTCCGCCCCCTGACGAGCATCACAAAATCGACGCTCAGTCAGAGTGC GAACCCGACAGACTATAAGAAACAGCGTTCCCCTTGTAGCTCCCTCGTGGCCTCTCTGATCCGAGTGACGCTAACGATACATGTCCGCCTTTCTCTCCCTTCCGG A
FOLK623F-FOLK1151R (12)	M13F	361.4	GTGGAATTCGCCCTTGGGCGTCGCGTTGATGATGCCCATGACAAACGTGCGTTCGCTCCAGGTCCAGAGCTGCGAGCGGATGGGCGTAACTCGGTACAGCTT GTGCTGCGGCAGCTCATGGAACAGCTGCGCGATCGTCTTCTTCAAAATTGGGTGCACGTAGTCCGCAGCAATGTCCTTCAGCGGACCCAGCACGAAGTCGCG CTCAGGAATTCTCGAGTGGGGGATGATCAGCGATTCATCGCTGCGGTGAACGTGGATCGAGTCGTAGAGCAGGATGTCAACGTCAATGCACCGCGGACCAA AGCGGATCCCTCCGGTCGTGCGTCCAAGTCTCTGTTCGACGCTCTTGACCAAGTCGAGCAGCTCGCCGGGTGTGAGCTTCGTTCGAATCTTGCACGCGGCATT CAAGAAGGCCGGCTGATCCACTACGTAGGCCGGAGGCGTCTGGTAGAGGGCCGAGGTGCTGATGAAAGGGCGAATTCCACAGTGGATATCAAGCTTATCGA TACCGTCGACCTCGAGGGGGGGCCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGA AATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGGAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGC GCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCG CTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAATCATGGGATAA CGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCTGACGAGCAT CACAAAATCGACGCTCAGTCAGAGTGCGAACCCGACAGACTATAAGATACAGGCGTTTCCCCCTGTGCTCCCTCGTGCGCTCTCTGCTCGACTGACGCTACGAT ACTGTCGACTTTTCTCCCTTCGGAAGCCATTGC

FOLK623F-FOLK1151R (13)	M13F	389.2	GGGCTGACTATATAGGGCGATTGGAGCTCCCGCGGTGCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGCCCAATGTGGAATTCGCCCTTGGGCGTC GCGTTGATGATGCCCATGACAAACGTGCGTTCGCTCCAGGTCCAGAGCTGCGAGCGGATGGGCGTAACTCGGTACAGCTTGTGCTGCGGCAGCTCATGGAAC AGCTGCGCGATCGTCTTCTTCAAAATTGGGTGCACGTAGTCCGCAGCAATGTCCTTCAGCGGACCCAGCACGAAGTCGCGCTCAGGAATTCTCGAGTGGGGG ATGATCAGCGATTCATCGCTGCGGTGGACGTGGATCGAGTCGTAGAACAGGATGTCAACGTCAATGCACCGCGGACCAAAGCGGATCCCTCCGGTCGTGCGT CCAAGTCTCTGTTCGACGCTCTTGACCAAGTCGAGCAGCTCGCCGGGTGTGAGCTTCGTTCGAATCTTGCACGCGGCATTCAAGAAGGCCGGCTGATCCACCA CGTAGGCCGGAGGCGTCTGGTAGAGGGCCGAGGTGCTGATGAAAGGGCGAATTCCACAGTGGATATCAAGCTTATCGATACCGTCGACCTCGAGGGGGGG CCCGGTACCCAGCTTTTGTTCCCTTTAGTGAGGGTTAATTGCGCGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCC ACACAACATACGAGCCGGGAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCG GGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGC GCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAAGGCGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGC AAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAGGCCGCGTTTGCTGGCGTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAATCGACGCTCAAGG TCAGAGTTGCGAACCGACAGGAACTATAGATACAGCGTTCCCCTGAAGCTCCTCGTGGCCTCTCGGATCGACTTACGCTACCGGATACCTGTTCCGCCGTC

FIGURE 17 – RAW SEQUENCING DATA

```
#!usr/bin/perl

#Author: Finlay Maguire

#Title: ConcatenatedTreeList

#Date : 18/08/2010


use strict;

use warnings;


#main

&double;

&triple;

&duplicates;

#end


sub double {


#open and read species lists into arrays

my $DHNAdata_file="DHNAremainnames.txt";

open(DHNA, $DHNAdata_file);

my @DHNAraw_data=<DHNA>;

close(DHNA);


my $DHPSdata_file="DHPSremainnames.txt";

open(DHPS, $DHPSdata_file);

my @DHPSraw_data=<DHPS>;

close(DHPS);


my $HPPKdata_file="HPPKremainnames.txt";
```

```
open(HPPK, $HPPKdata_file);
my @HPPKraw_data=<HPPK>;
close(HPPK);

#open output files
open (DHPS_DHNA, ">DHPS_DHNAnames.txt");
open (DHPS_HPPK, ">DHPS_HPPKnames.txt");
open (DHNA_HPPK, ">DHNA_HPPKnames.txt");

#check for sequences with 2 domain combinations
#DHPS-DHNA combination
foreach $DHPSraw_data (@DHPSraw_data){                    #Loop through each DHPS species
        foreach $DHNAraw_data (@DHNAraw_data){                    #Check for matches in DHNA by looping
through DHNA for each DHPS entry
                if ($DHNAraw_data eq $DHPSraw_data) {             #If they match print sequence data
                        print DHPS_DHNA $DHPSraw_data;
                }
        }
}

#DHPS-HPPK combination
foreach $HPPKraw_data (@HPPKraw_data){
        foreach $DHPSraw_data (@DHPSraw_data){
                if ($HPPKraw_data eq $DHPSraw_data){
                        print DHPS_HPPK $DHPSraw_data;
                }
        }
}
```

```
#DHNA-HPPK combination
foreach $HPPKraw_data (@HPPKraw_data) {
        foreach $DHNAraw_data (@DHNAraw_data){
                if ($HPPKraw_data eq $DHNAraw_data){
                        print DHNA_HPPK $HPPKraw_data;
                }
        }
}

#close output files
close DHPS_DHNA;
close DHPS_HPPK;
close DHNA_HPPK;

}


sub triple {

#open output file
open (DHPS_HPPK_DHNA, ">DHPS_HPPK_DHNA.txt");

#open input files (output files from pervious sub) and read into arrays
my $DHPS_DHNAdata_file="DHPS_DHNAnames.txt";
open (DHPS_DHNA, $DHPS_DHNAdata_file);
my @DHPS_DHNAraw_data=<DHPS_DHNA>;
close DHPS_DHNA;
```

```
my $DHPS_HPPKdata_file="DHPS_HPPKnames.txt";
open(DHPS_HPPK, $DHPS_HPPKdata_file);
my @DHPS_HPPKraw_data=<DHPS_HPPK>;
close(DHPS_HPPK);



#check for sequences with triple fusion
foreach $DHPS_DHNAraw_data (@DHPS_DHNAraw_data) {
        foreach $DHPS_HPPKraw_data (@DHPS_HPPKraw_data){
                if ($DHPS_HPPKraw_data eq $DHPS_DHNAraw_data){
                        print DHPS_HPPK_DHNA $DHPS_HPPKraw_data;
                }
        }
}

#close output and input files
close DHPS_DHNA;
close DHPS_HPPK;
close DHNA_HPPK;
close DHPS_HPPK_DHNA;
}

#Delete duplicate entries in output files
sub duplicates {

#open output files
open (DHNAHPPKDHPS, ">DHNAHPPKDHPSnames.txt");
```

```
open (DHNAHPPK, ">DHNAHPPKnames.txt");
open (DHNADHPS, ">DHNADHPSnames.txt");
open (HPPKDHPS, ">HPPKDHPSnames.txt");

#search files for duplicate entries and delete duplicate
#DHNA-HPPK duplicates
my $file = 'DHNA_HPPKnames.txt';
my %seen = ();
my @ARGV = ($file);
  while(<>){
    $seen{$_}++;
    next if $seen{$_} > 1;
    print DHNAHPPK;
  }

#DHNA-DHPS duplicates
$file = 'DHPS_DHNAnames.txt';
%seen = ();
   @ARGV = ($file);
  while(<>){
    $seen{$_}++;
    next if $seen{$_} > 1;
    print DHNADHPS;
  }

#HPPK-DHPS duplicates
$file = 'DHPS_HPPKnames.txt';
%seen = ();
```

```
  @ARGV = ($file);
 while(<>){
   $seen{$_}++;
   next if $seen{$_} > 1;
   print HPPKDHPS;
 }

#DHNA-HPPK-DHPS duplicates
$file = 'DHPS_HPPK_DHNAnames.txt';
%seen = ();
@ARGV = ($file);
while(<>){
        $seen{$_}++;
        next if $seen{$_} > 1;
        print DHNAHPPKDHPS;
}

#close output files
close DHNAHPPKDHPS;
close DHNAHPPK;
close DHNADHPS;
close HPPKDHPS;

}
```

FIGURE 18 – CONCATENATEDTREELIST SOURCE CODE

```
#!usr/bin/perl

#Author: Finlay Maguire

#Title: Renamer

#Date : 18/08/2010


use strict;

use warnings;


#rename files for concatenation

my $ChangeFile="*.phy"; #open masked sites file

my $NewNames="newnames.txt"; #open list of names from target concatenation file

my $OldNames="oldnames.txt"; #open list of names from phylip file


#open and read phylip file into a single string and then close

open(Change, $ChangeFile);

my $ChangeFileString=do { local($/);<Change>};

close(Change);


#read list of target concatenation sequence names into an array

open(New, $NewNames);

chomp( my @NewNamesArray=<New> );  #remove newline character

close(New);


#read list of phylip file names into an array

open(Old, $OldNames);

chomp( my @OldNamesArray=<Old> );  #remove newline character

close(Old);
```

```
#change names in phylip for concatenation

for my $count ( 0 .. $#OldNamesArray ){ #do loop until the last name is processed

        $ChangeFileString =~ s/\Q$OldNamesArray[$count]/$NewNamesArray[$count]/g; #search for old sequence name in phylip and

replace with new name

}


#output updated renamed phylip file

print $ChangeFileString;


#end
```

FIGURE 19 – RENAMER SOURCE CODE

Species	**Putative Paralogue character**
Branchistoma floridae	Low identity unfused T-fold superfamily
Puccinia graminis	Low identity unfused pterin-binding superfamily
Physcomitrella patens	Unfused copy DHPS
Haloterrigena turkmenica	Copy of FPGS-DHPS fusion

FIGURE 20 – IDENTIFIED PARALOGUE CHARACTERS

Phylogeny	Total sites	Sites without polymorphism	Proportion of uninformative sites
DHNA	92	1	1.09%
HPPK	85	6	7.06%
DHPS	81	1	1.23%
DHNA-DHPS	173	10	5.78%
DHNA-HPPK	177	7	3.95%
HPPK-DHPS	166	17	10.24%
DHNA-HPPK-DHPS	258	20	7.75%

FIGURE 21 – INFORMATIVE SITES PER PHYLOGENETIC ANALYSIS

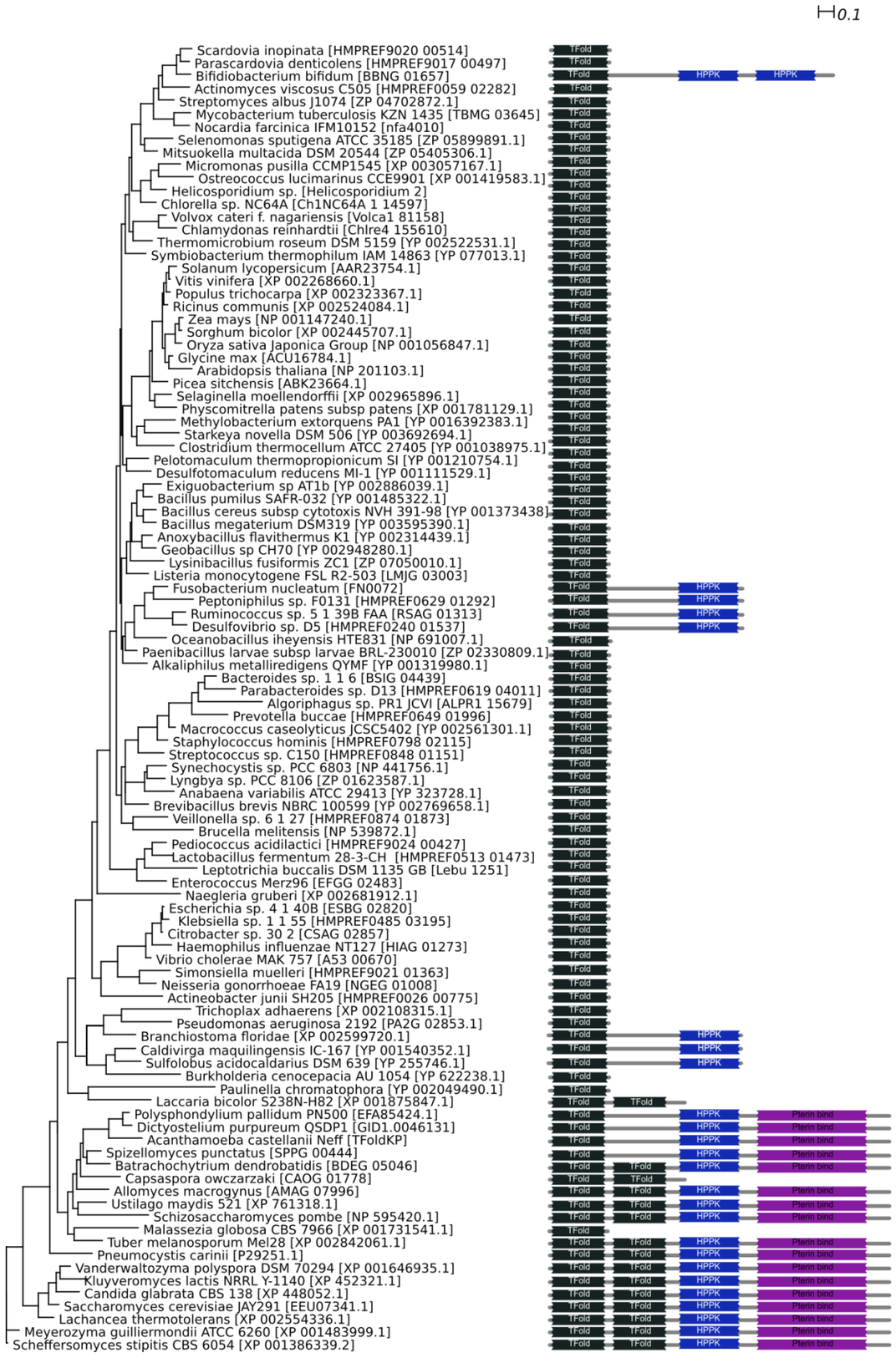

FIGURE 22 – DHNA PHYML TREE RECONSTRUCTED WITH 103 SEQUENCES AND 93 CHARCTERS.

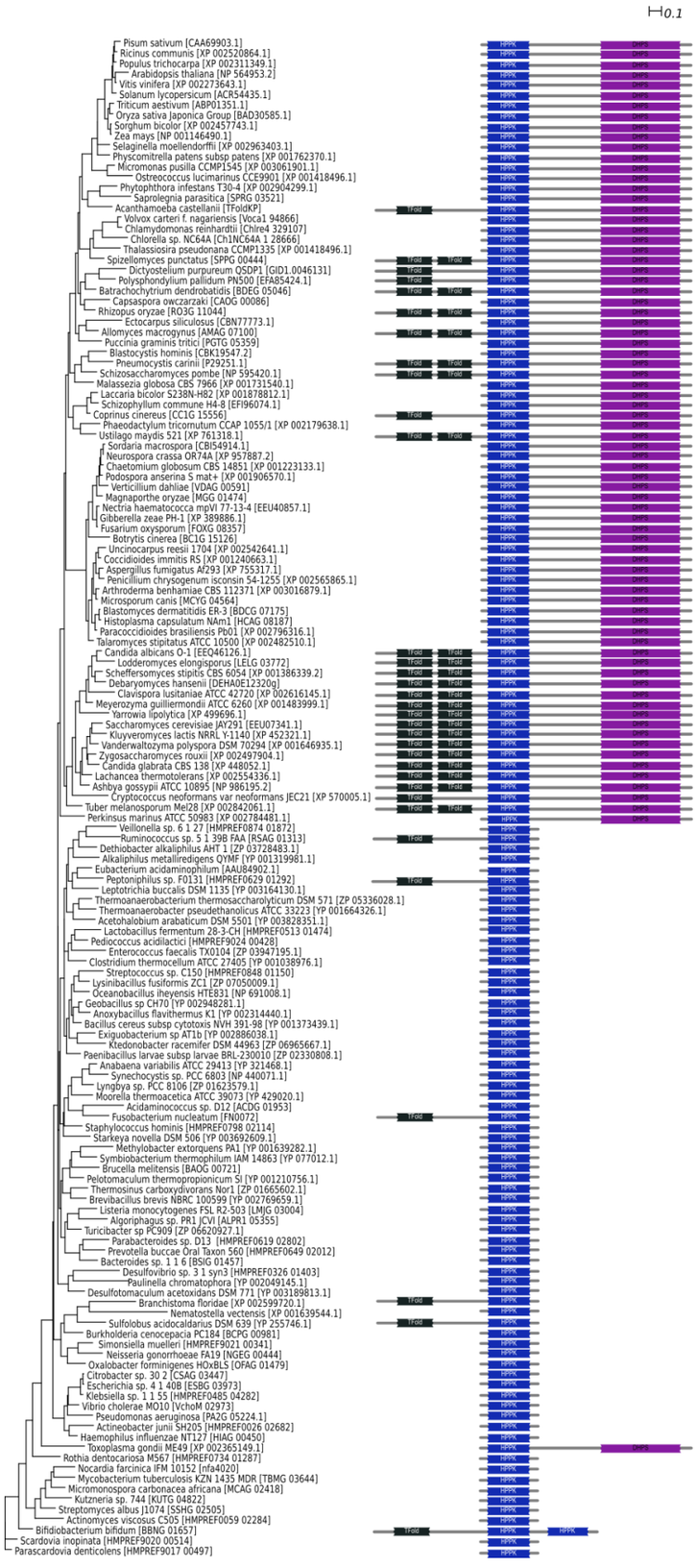

FIGURE 23 – HPPK PHYML TREE RECONSTRUCTED WITH 147 SEQUENCES AND 87 CHARCTERS.

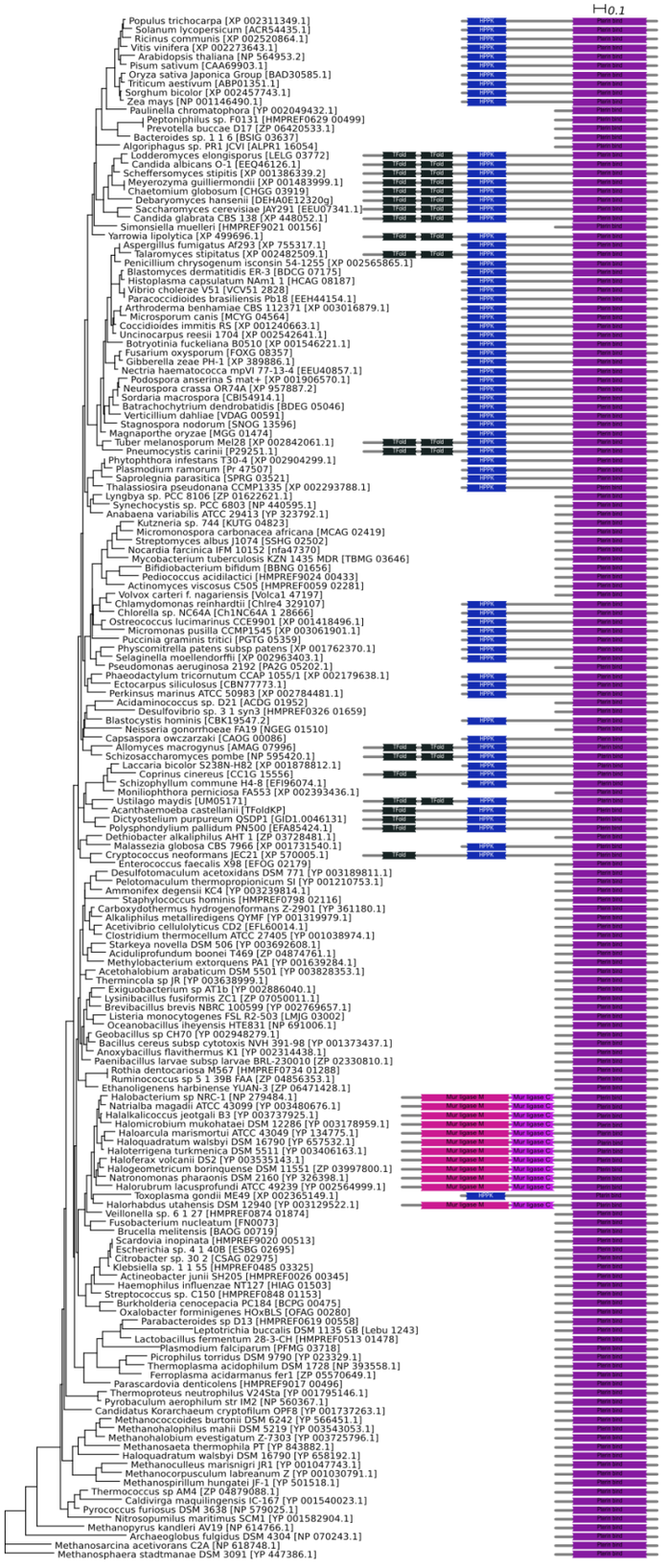

FIGURE 24 – DHPS PHYML TREE RECONSTRUCTED WITH 172 SEQUENCES AND 81 CHARCTERS.

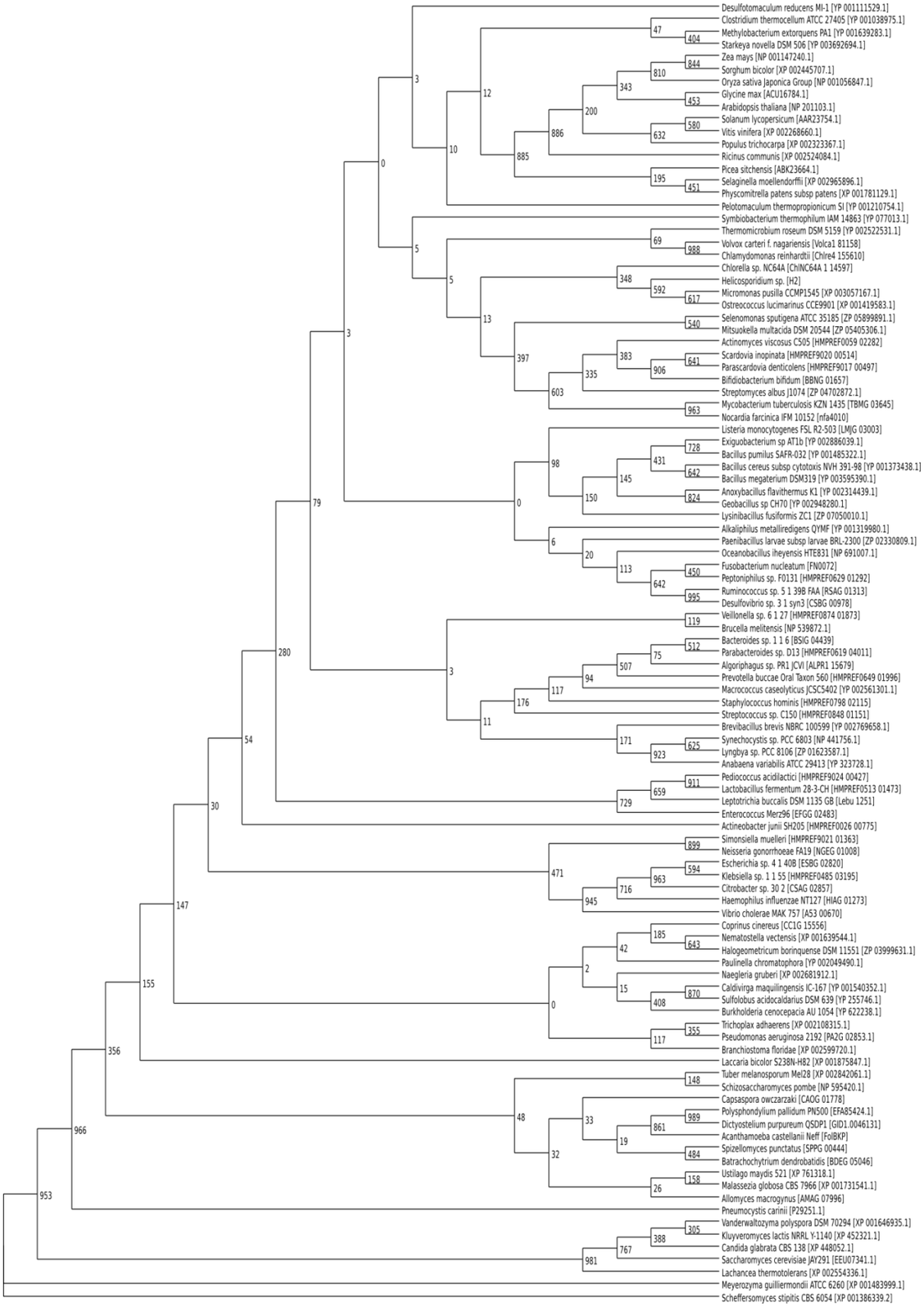

FIGURE 25 – DHNA BOOTSTRAP SUPPORT VALUES

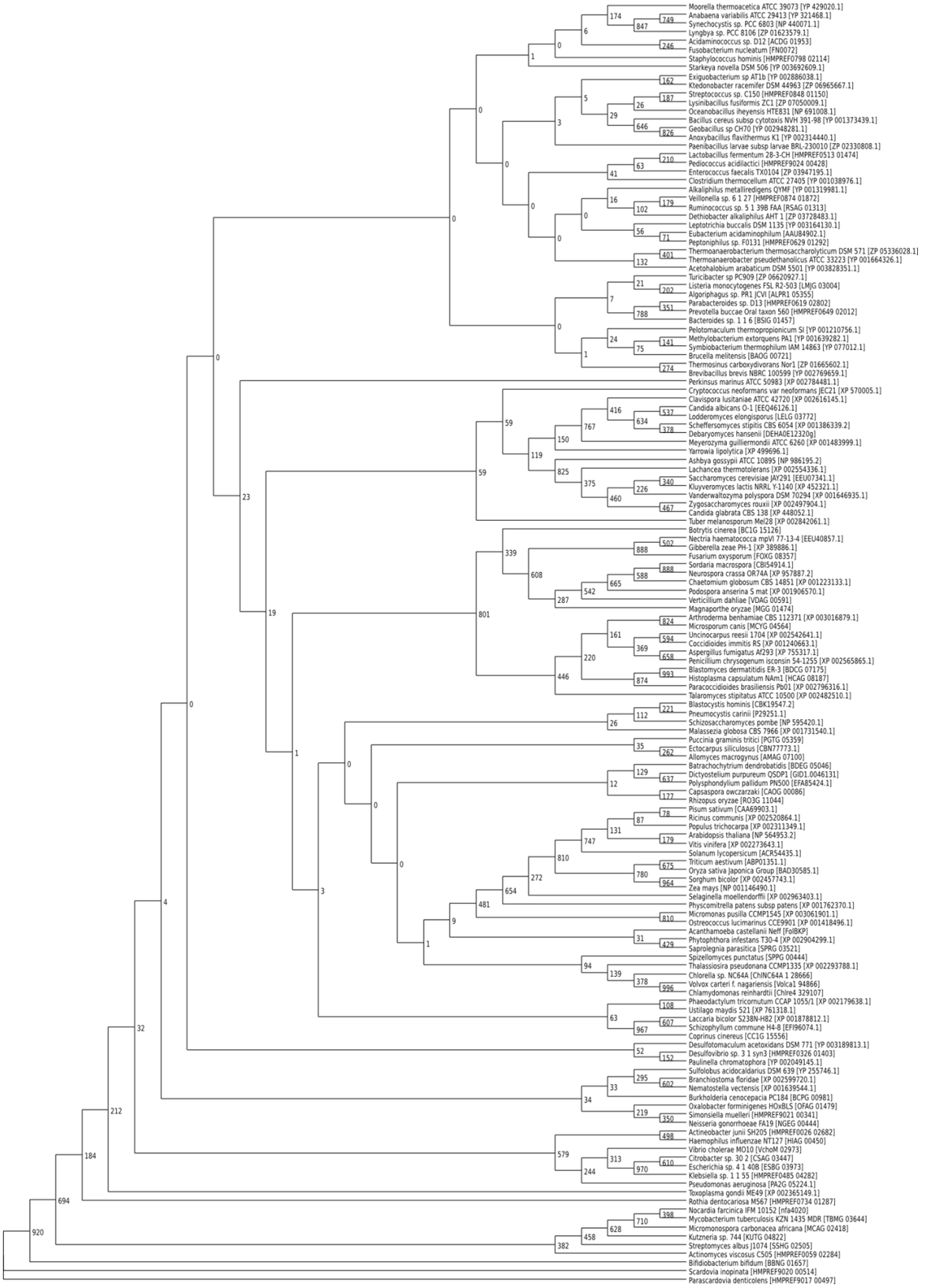

FIGURE 26 – HPPK BOOTSTRAP SUPPORT VALUES

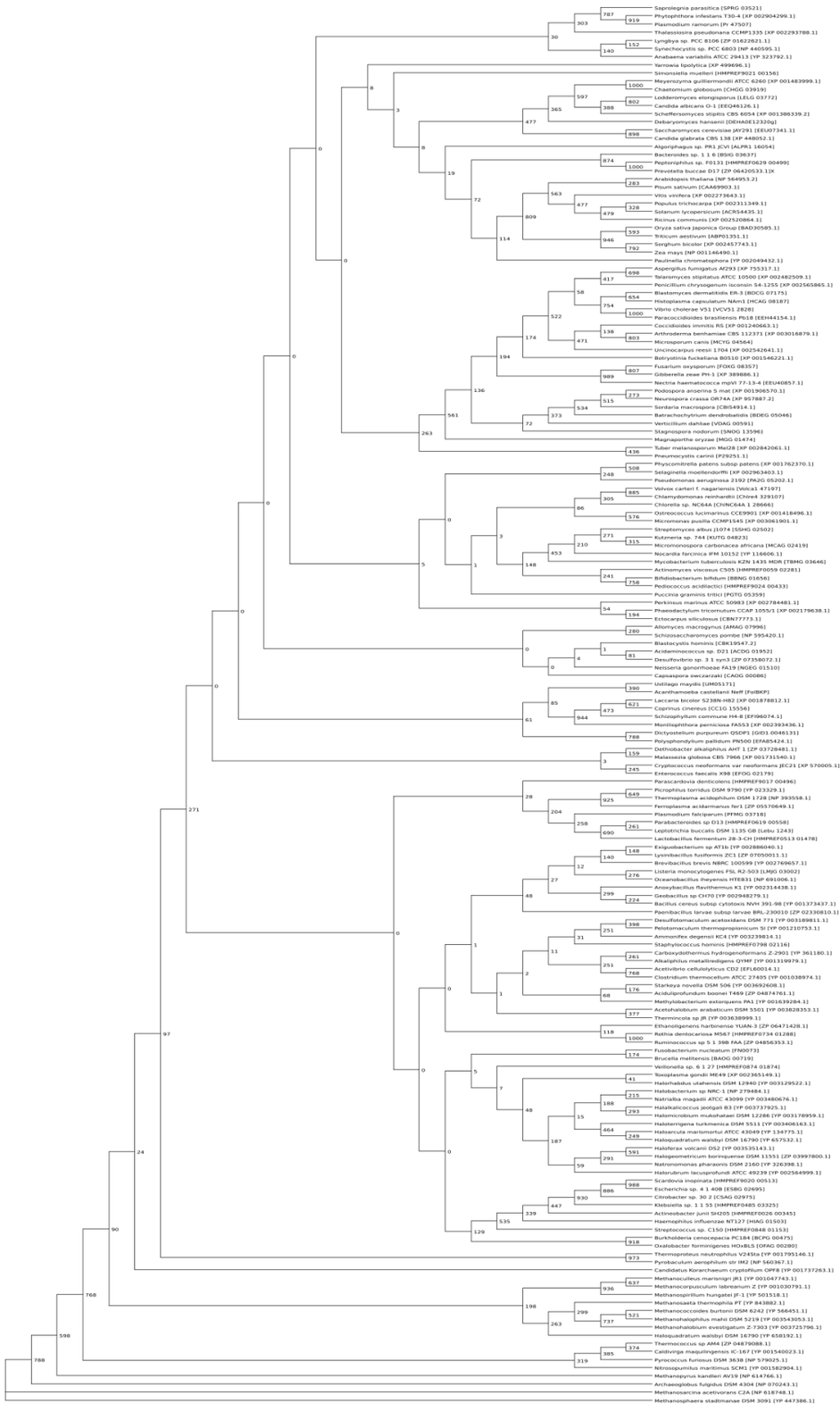

FIGURE 27 – DHPS BOOTSTRAP SUPPORT VALUES

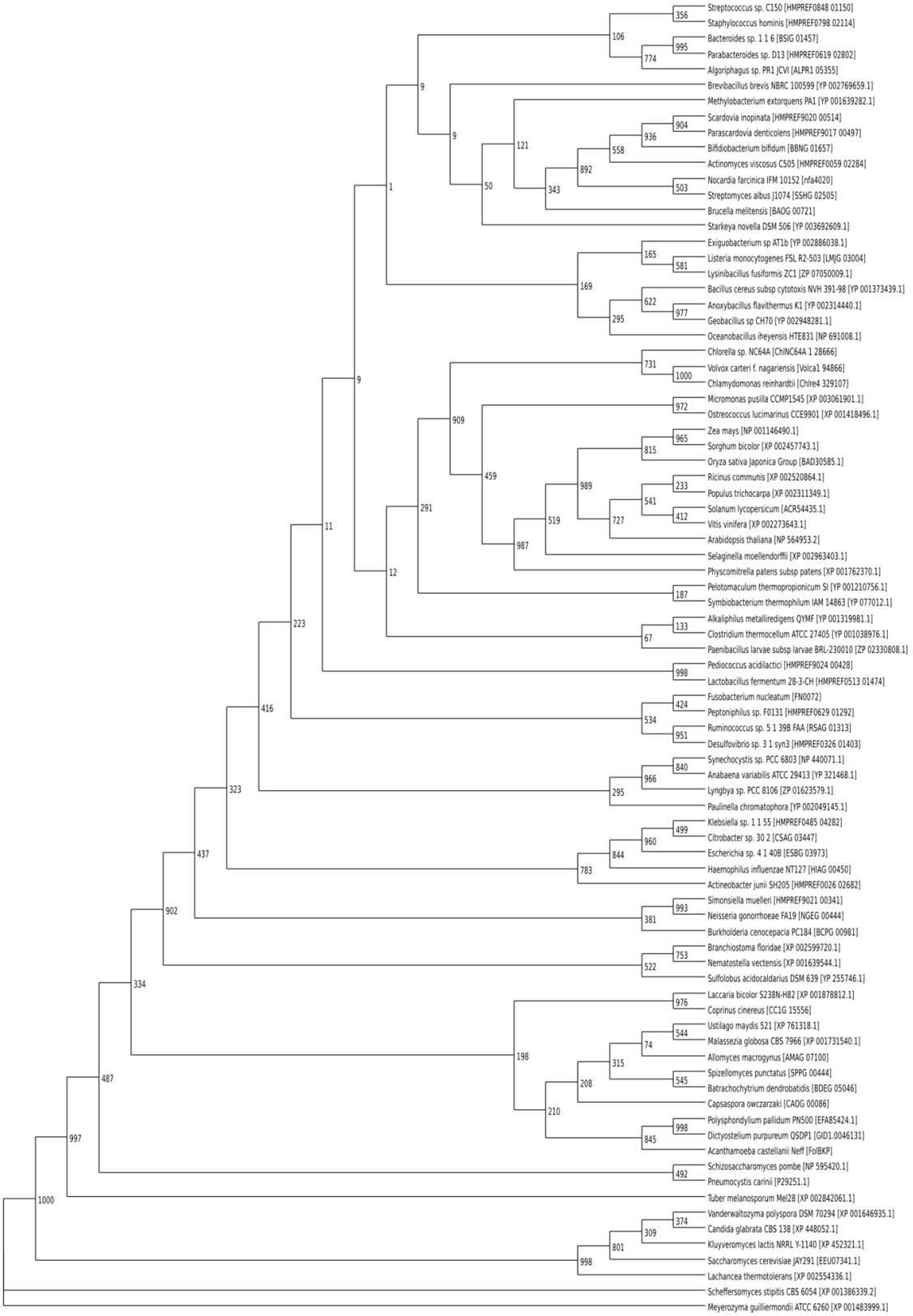

FIGURE 28 – DHNA-HPPK BOOTSTRAP SUPPORT VALUES

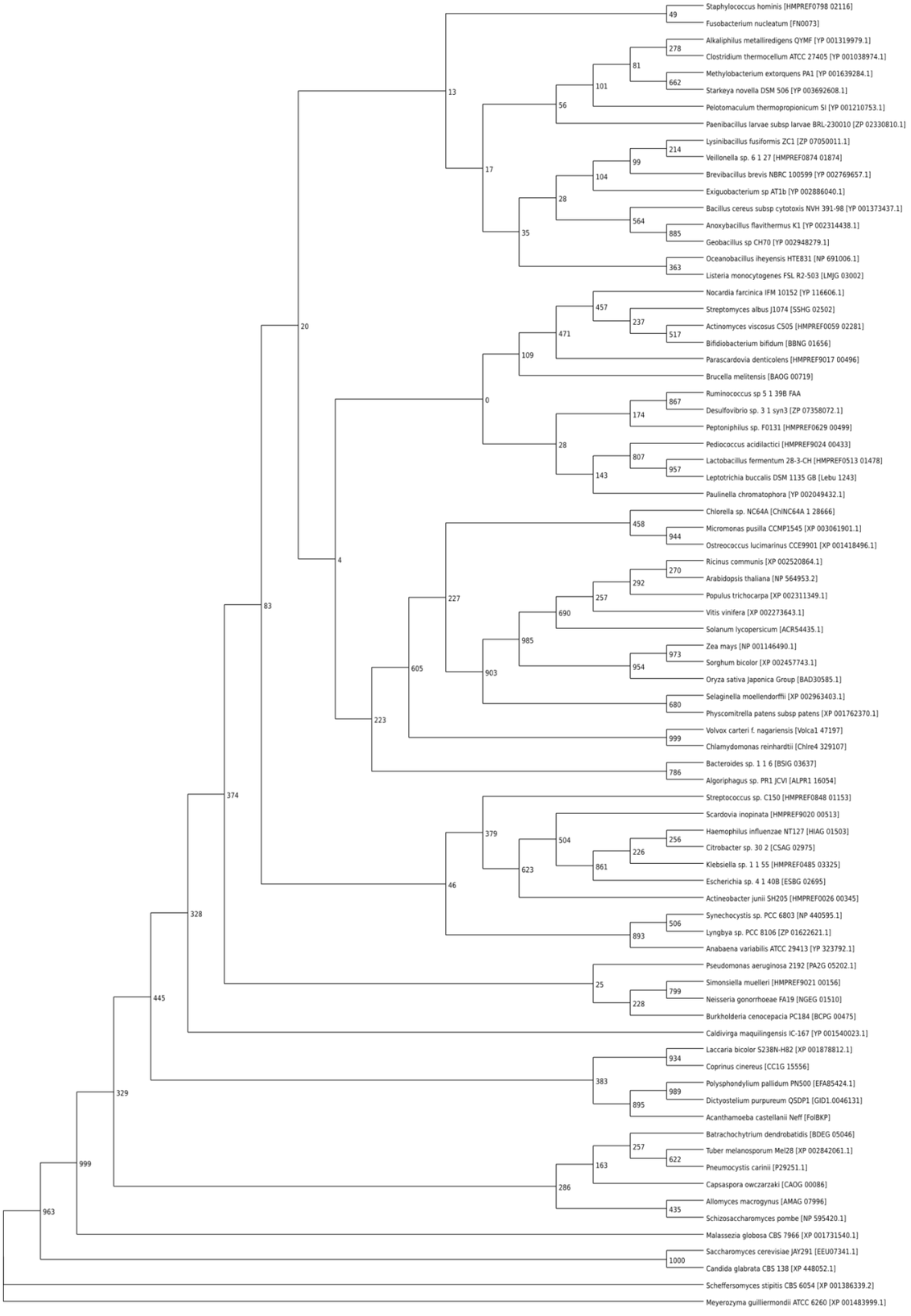

FIGURE 29 – DHNA-DHPS BOOTSTRAP SUPPORT VALUES

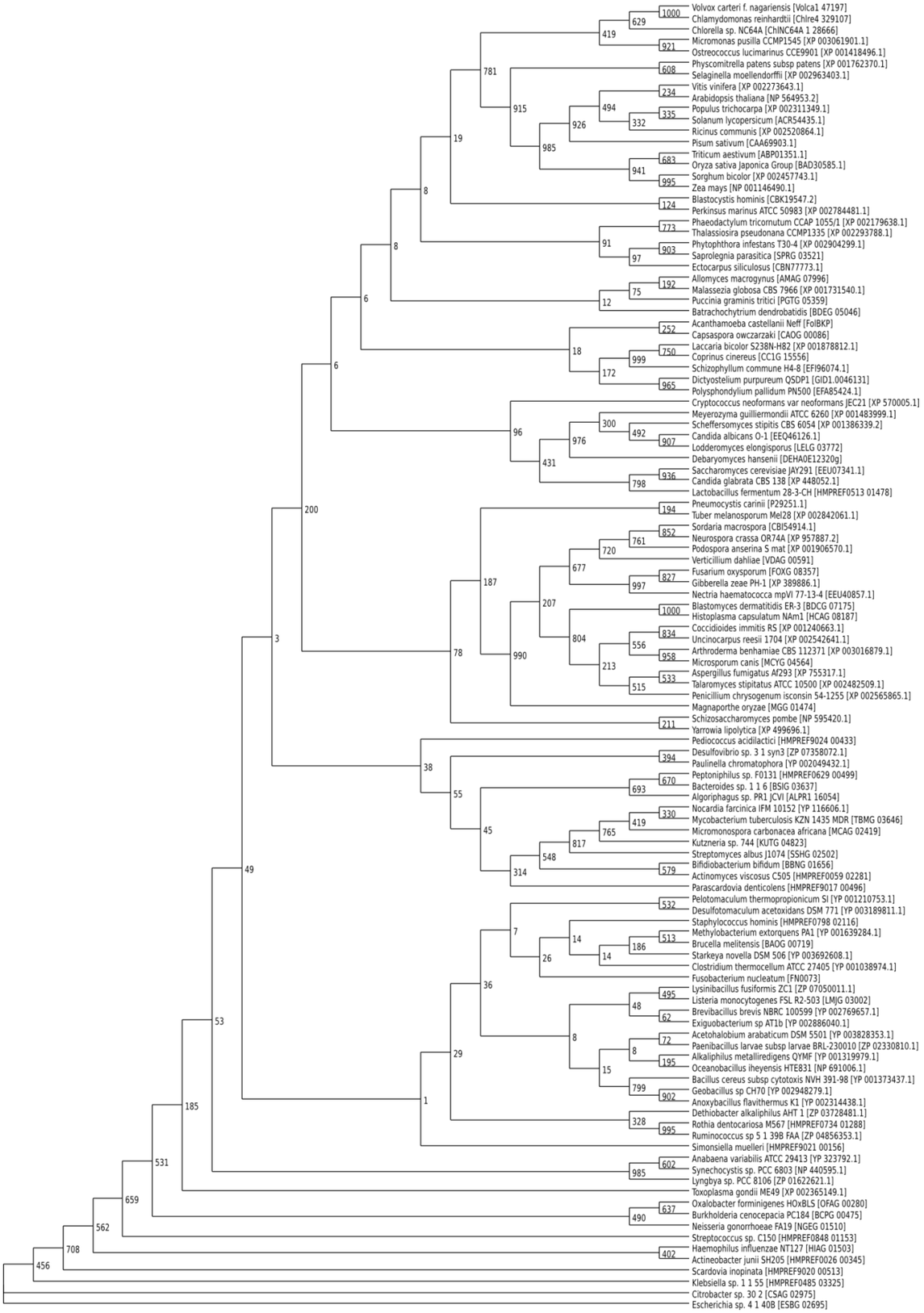

FIGURE 30 – HPPK-DHPS BOOTSTRAP SUPPORT VALUES

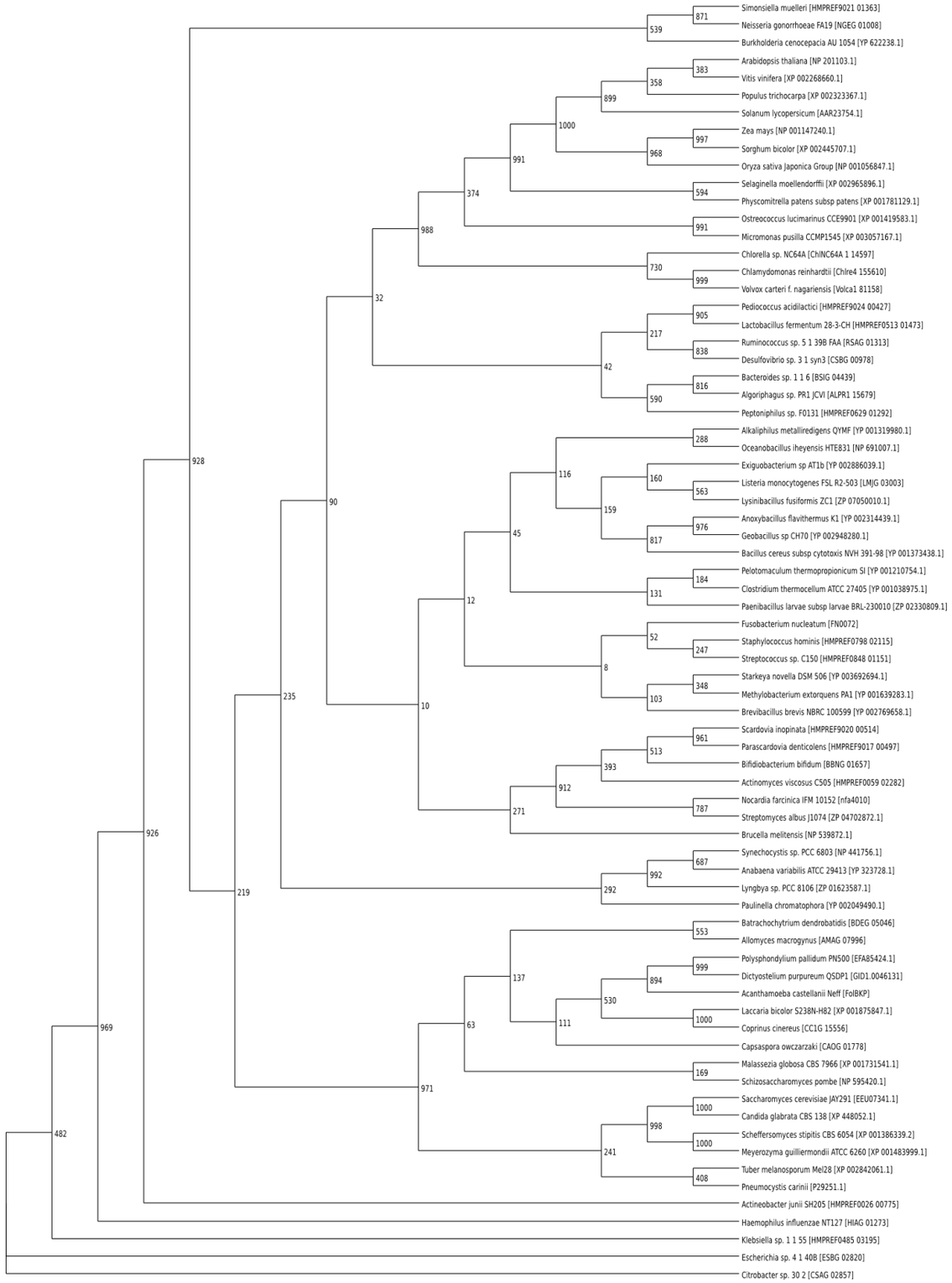

FIGURE 31 – DHNA-HPPK-DHPS BOOTSTRAP SUPPORT VALUES