Itch Response: Phylogenetic Analysis of the Evolution of Prdm8

Kaitlin Hamilton

03-727 Phylogenetics

December 6, 2013

Introduction

Prdm8 is a protein that serves as a key mediator in development by acting as a transcriptional regulator with an intrinsic methyltransferase activity (1). As a result of its activity, this protein is a member of the histone methyltransferase superfamily and contains the domain characteristic to this family, which is a SET domain (1). Prdm8 actually contains a PR/SET domain, but the PR domain has been shown to be very similar to the SET domain (2). Histone methyltransferases are responsible for modifying histones by transferring up to three methyl groups to either lysine or arginine residues depending upon the specific type of enzyme involved (2). Prdm8 transfers methyl groups to lysine residues. This activity is carried out by the SET domain and is done by the domain binding to the lysine it will modify plus an S-Adenosyl methionine cofactor (3). A specific tyrosine residue in the binding pocket deprotonates the ε-amino group of the lysine, allowing it to attack the methyl group on the S-Adenosyl methionine and have the methyl group transferred on to it (3). The Prdm8 protein can also feature a zinc finger binding domain (1). This domain recognizes and binds to particular sequences of DNA (4).

Ross, et al. in a 2012 study discovered that Prdm8 interacts with another protein, Bhlhb5, in mice (1). Bhlhb5 is a transcription factor found in excitatory neurons, and Ross, et al. suggested that Bhlhb5 may bind to its targets and then recruit Prdm8 in order to form a complex that would repress the expression of specific genes in order to allow proper neural circuit formation (1). They came to this conclusion based upon the fact that mice in which either of these two proteins was lacking exhibited the same phenotypes of improperly formed nervous systems and incessant itching that led to the development of skin lesions (1). Ross, et al. also discovered in a blastn search that Prdm8 matched most similarly not to another SET domain protein but to Zfp488, a zinc finger domain-containing protein (1). An alignment between Prdm8 and Zfp488 revealed that the C terminus, which is where the zinc finger domain of Prdm8 is located, of each protein was 82% similar (Appendix A1).

What is significant about this finding by Ross, et al. is that Zfp488 has been documented to interact with Olig2, and Olig2 is a close relative of Bhlhb5 (1). This suggests that the Prdm8 and Zfp488 pair and the Bhlbh5 and Olig2 pair are each similar proteins that resulted from a duplication event in each pair's respective protein family tree (1). A hypothesis as to what Ross, et al. thought such trees would look like can be found in the appendix (Appendix A2). Thus, the purpose of the study described in this paper was to focus specifically on the Prdm8 protein and determine when the duplication event may have occurred that links Prmd8 and Zfp488. A BLAST search was carried out using the *Mus musculus* Prdm8 sequence used by Ross, et al. in their study with the goal of obtaining similar sequences containing SET domains and zinc finger domains that would allow for a reconstruction of the evolutionary history of these domains and their associated Prdm8 and Zfp488 proteins.

Methods

Acquiring Sequences

The sequences for this study were obtained by performing a BLAST search in UniProt using the *Mus musculus* Prdm8 sequence that was utilized in the Ross, et al. study (Appendix A3). The search was carried out on the UniProt Eukaryotic database at a threshold of 10. In order to reduce the number of results received, only those sequences with an E value equal to or lower than 8.0×10^{-18} and a query coverage greater than 36% were originally selected. This resulted in the initial acquisition of 92 sequences. The number of sequences acquired was further reduced in Jalview by removing incomplete sequences and sequences containing sequencing errors. This resulted in a final collection of 52 sequences representing 33 different species (Appendix A4). These sequences were then further edited by being renamed with their five letter species code and UniProt ID number.

Multiple Sequence Alignment

Multiple sequence alignments were carried on the sequences in Jalview using MAFFT, MUSCLE, ProbCons, and TCoffee, all of which were run with default settings. The alignments were ranked based upon how well they aligned the PR/SET and zinc finger domains as identified in the *Mus musculus* Prdm8 query sequence on its UniProt information page. The PR/SET domain occurred between amino acid residues 16 to 131 and the C_2H_2-type zinc finger domains occurred between amino acid residues 154 to 182 and 624 to 647 in this sequence. The TCoffee

program aligned these regions best, and so it was chosen for the next step of the study. Prior to being used, however, the alignment was trimmed manually by removing any columns consisting of more than 50% gaps. This was done in stages in order to determine the optimal amount of trimming since there were many gaps present in areas outside of the conserved PR/SET and zinc finger domains. The first stage removed all columns consisting of more than 80% gaps, the second stage removed all columns consisting of more than 70% gaps, the third stage removed all columns consisting of more than 60% percent gaps, and the fourth stage removed all columns consisting of more than 50% gaps. Each stage was evaluated by generating a Neighbor Joining tree with bootstraps in Seaview and comparing the tree to those of the other stages. The tree in which all columns consisting of more than 60% gaps were removed turned out to have the best resolution and bootstrap scores, so this version of the trimmed alignment was used for further analysis (Appendix A5).

Model Selection

The trimmed TCoffee alignment was next uploaded into TOPALi in order to determine the best model for tree construction of the sequences. The JTT+G model had the lowest AIC and BIC scores, so it was chosen as the construction model (Appendix A6). It had an AIC1 score of 29,626.41, an AIC2 score of 29627.04, and a BIC score of 48,000.78. It also had an α equal to 0.932.

Phylogenetic Tree Construction

PhyML software in Seaview was utilized to build a Maximum Likelihood tree using the JTT+G model with an α equal to 0.932, no invariable sites, and "Tree searching operations" set to "Best of NNI and SPR" (Appendix A7). The remaining parameters were left at their default values. A Neighbor Joining tree with 1,000 replicates and using observed distances was also created in Seaview (Appendix A8).

Species Tree Construction

A species tree for all of the species represented in the sequences was generated using the NCBI Taxonomy Common Tree. Uncertainties in branching were resolved in this tree by editing the Newick format of this tree in Notepad++ in accordance to species trees found in literature (5, 6, 7, 8). The revised species tree was then uploaded into Notung 2.7 and color annotated according to taxonomic groups (Appendix A9).

Gene and Species Tree Reconciliation

The Maximum Likelihood gene tree was uploaded into Notung 2.7 and color annotated in the same manner as the species tree (Appendix A10). This tree was then reconciled with the species tree using default settings. Since the original gene tree was unrooted, a rooting analysis was performed on the reconciled tree in Notung in order to determine the optimal branch for the root that would minimize the number of duplications and losses. In an attempt to lower the event score for the reconciled tree, the tree was also rearranged with an edge weight threshold of 0.75.

Results and Discussion

The multiple sequence alignments performed on the 52 sequences analyzed in this study demonstrated that the PR/SET and zinc finger domains were highly conserved amongst the sequences that contained these domains. This indicates the importance of the particular sequences of these domains for carrying out their specific functions across species.

The Maximum Likelihood gene tree generated in Seaview features many strongly supported branches since many of the bootstrap values were higher than 0.80 (Appendix A8). However, this tree also featured a large polytomy towards its top in which the branching of many of the primate sequences was uncertain (Appendix A8). This most likely occurred due to the high similarity between the sequences obtained for the primate species. This polytomy was ultimately resolved when the Maximum Likelihood gene tree was reconciled with the species tree in Notung (Appendix A11). The initial reconciled gene tree had a score of 107 with 26 duplications and 68 losses. As mentioned in the Methods section, this tree was rearranged with an edge weight threshold of 0.75, and this lowered the score to 72.5 with 21 duplications and 41 losses. This second tree was the one used for further analyses.

The results of the UniProt BLAST search returned mostly sequences that matched the PR/SET domain and not the zinc finger domain of the original *Mus musculus* query sequence. There was only one sequence that matched only the zinc finger domain of the query, and this was a zinc finger protein with the UniProt ID F6UEU6 from a duckbill platypus. Each gene in the reconciled gene tree was highlighted to indicate which portion of the query sequence it covered, and this can be seen in Appendix A12. There is clearly an order to the way in which the genes are organized in regards to which domains they contain since there are, for instance, blocks of red sequences that containing both the PR/SET and zinc finger domains and containing highly

similar sequences in between these domains and groups of blue and green sequences that covered and contained only the PR/SET and zinc finger domains. What is of importance to note is the block of mainly green sequences at the top of the tree that also contains the one yellow labeled sequence that matched and contained just the zinc finger domain. The green label indicates sequences that matched and covered only the PR/SET and zinc finger domains. The fact that the yellow sequence occurs within this group could possibly indicate that somewhere along the branch containing F6UEU6, the SET domain was lost and only the zinc finger domain was retained. It is possible then that the duplication event suggested by Ross, et al. that led to the creation of Zfp488 and Prdm8 may have occurred somewhere in this section of the tree. However, there is not enough data to strongly support such a conclusion because there was only one zinc finger protein obtained in the sequences used in this study.

The fact that only one zinc finger protein appeared in the BLAST search for highly similar sequences and that this zinc finger protein was not very similar to the query sequence, as indicated by its branching far away from Q8BZ97 in the reconciled gene tree, is interesting because Ross, et al. had found in their blastn search that the sequence Q8BZ97 was more closely related to Zfp488 than any PR/SET domain containing proteins. The results of this study demonstrated the exact opposite. In order to investigate this a bit further, only the C-terminal zinc finger domain of the query sequence was used to perform a BLAST search using the RefSeq database of NCBI at a threshold of 10. A total of 82 sequences were obtained ranging from Prdm8 sequences to several Zfp488 and Zfp488-like sequences. A similar procedure to that outlined previously in the Methods section was performed except the JTT+I+G model was used to generate a Maximum Likelihood tree, which can be seen in Appendix A12. The Zfp488 sequences were highlighted in yellow, and as can be seen in the tree, they segregated far from the query sequence. Ross, et al. did not do any phylogenetic analysis to support their claim that the *Mus musculus* Prdm8 sequence is more closely related to Zfp488 than to other PR/SET domain containing proteins, so the findings of this study suggest that the suggested relationship by Ross, et al. may be incorrect.

In the future, the relationship between Zfp488 and Prdm8 can be further analyzed by expanding the amount of data in the reconciled gene tree found in Appendix A10 in order to better determine where the duplication event linking the two may have occurred.

References

1) Bhlb5 and Prdm8 Form a Repressor Complex Involved in Neuronal Circuit Assembly. Ross *et al.* Neuron. 2012 Jan 26; 73(2):292-303.

2) Posttranslational Modifications of Histones by Methylation. Wood, Adam and Ali Shilatifard. Proteins in Eukaryotic Transcription, 2004; 67:201-222.

3) Structure and Catalytic Mechanism of a SET Domain Protein Methyltransferase. Trievel, Raymond, *et al.* Cell, 2002 Oct 4; 111(1):91-103.

4) Zinc finger domain. Genetics Home Reference. National Institute of Health, 2013 Dec 3. <http://ghr.nlm.nih.gov/glossary=zincfingerdomain>.

5) Complex Distribution of Avian Color Vision Systems Revealed by Sequencing the SWS1 Opsin from Total DNA. Odeen, Anders, and Olle Hastad. Mol. Biol. Evol, 2003; 20(6):855-861.

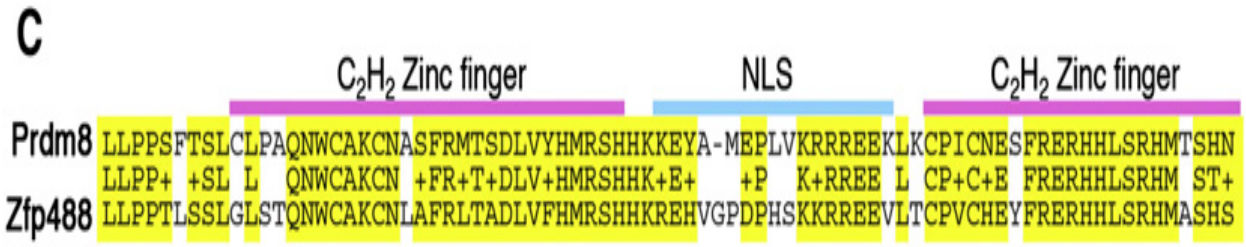
6) Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Nishihara, Hidenori, *et al.* Proceedings of the National Academy of the Sciences, 2006; 108(26):9929-9934.

7) A complete phylogeny of the whales, dolphins, and even-tode hoofed mammals (Cetartiodactyla). Price, Samantha, *et al.* Biol. Rev., 2005; 80:445-473.

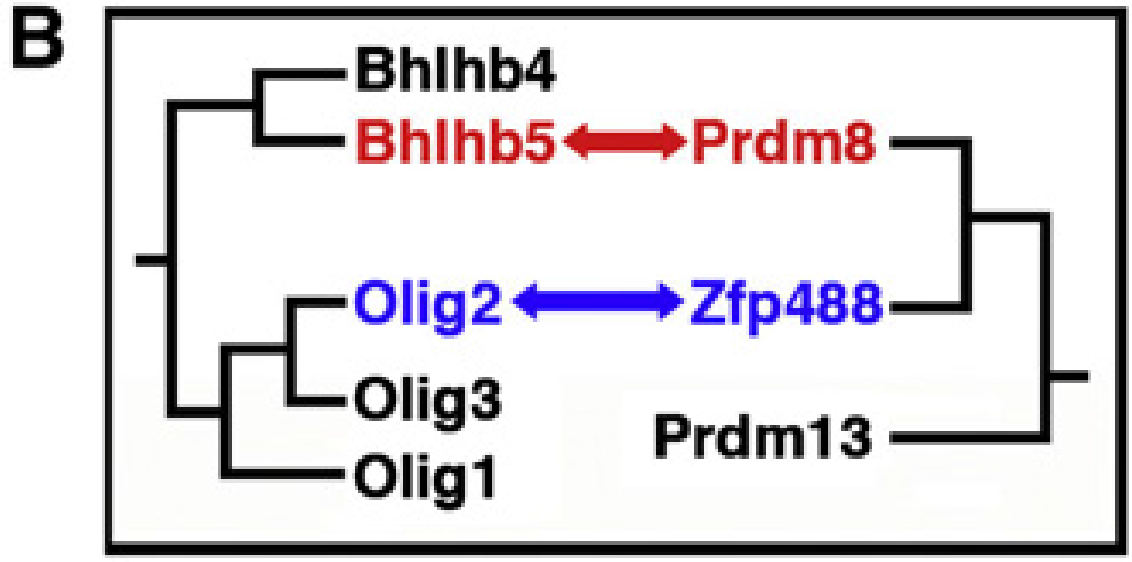
8) Novel evolutionary relationship among four fish model systems. Chen, Wei-Jen, *et al.* Trends in Genetics, 2004; 20(9):424-431.

Appendix

A1: Ross, et al. Alignment of Prdm8 and Zfp488



A2: Ross, et al. Suggested Relationship Between Zfp488/Prdm8 Pair and Bhlhb5/Olig2 Pair



A3: FASTA Sequence of *Mus musculus* Prdm8 Query

```
>sp|Q8BZ97|PRDM8_MOUSE PR domain zinc finger protein 8 OS=Mus musculus GN=Prdm8
PE=2 SV=1
MEDSGIQRGIWDGDAKAVQQCLTDIFTSVYTTCDIPENAIFGPCVLSHTSLYDSIAFVAL
KSTDKRTVPYIFRVDTSAANGSSEGLMWLRLVQSARDKEEQNLEAYIKNGQLFYRSLRRI
AKDEELLVWYGKELTELLLLCPSRAHKMNGSSPYTCLECSQRFQFEFPYVAHLRFRCPKR
LHSTDANPQDEQGGGLGTKDHGGGGGGKEQQQQQQQQQQEAPLIPGPKFCKAGPIHHYPA
SSPEASNPPGSAGASSAKPSTDFHNLARELENSRGNSSCVAAPGVGSGGSGHQEAELSPD
GVATGGCKGKRRFPEEAAAEGGGAGLAGGRARFSERPLATSKEELVCTPQQYRAAGSYFG
LEENGRLFAPPSPETGEAKRSAFVEVKKAGRAVGLQEEAAATDGAGGTAEDPDAGGGVAG
GGSNGSSTPAAGSPGAPEKLLAPRPGGSLPGRLEGGSPARGSAFTSVSQLGGGGGAGTAG
TAGGSGGGQTAASDERKSAFSQPARSFSQLSPLVLGQKLGALEPCHPGDGVGPTRLYPAA
ADPLAVKLQGAADLNGACGPLASGGGGGLPKQSPFLYATAFWPKSSAAAAAAAAAAAGPL
QLQLPSALTLLPPSFTSLCLPAQNWCAKCNASFRMTSDLVYHMRSHHKKEYAMEPLVKAA
AGGETQVPHLQRVLQGASPPVPAHDLA
```

A4: Species List - Five Letter Code (Species Name)

MELGA (Meleagris gallopavo)

CHICK (Gallus gallus)

CHEMY (Chelonia mydas)

ORNAN (Ornithorhynchus anatinus)

DANRE (Danio rerio)

ORYLA (Oryzias latipes)

XIPMA (Xiphophorus maculatus)

ORENI (Oreochromis niloticus)

TAKRU (Takifugu rubripes)

TETNG (Tetraodon nigroviridis)

XENTR (Xenopus tropicalis)

CAVPO (Cavia porcellus)

MYOBR (Myotis brandtii)

RABIT (Oryctolagus cuniculus)

RAT (Rattus norvegicus)

NOMLE (Nomascus leucogenys)

MOUSE (Mus musculus)

BOSMU (Bos grunniens mutus)

BOVIN (Bos Taurus)

PIG (Sus scrofa)

HORSE (Equus caballus)

9CETA (Camelus ferus)

CANFA (Canis familiaris)

TAEGU (Taeniopygia guttata)

CALJA (Callithrix jacchus)

MACMU (Macaca mulatta)

PONAB (Pongo abelii)

HUMAN (Homo sapiens)

GORGO (Gorilla gorilla gorilla)

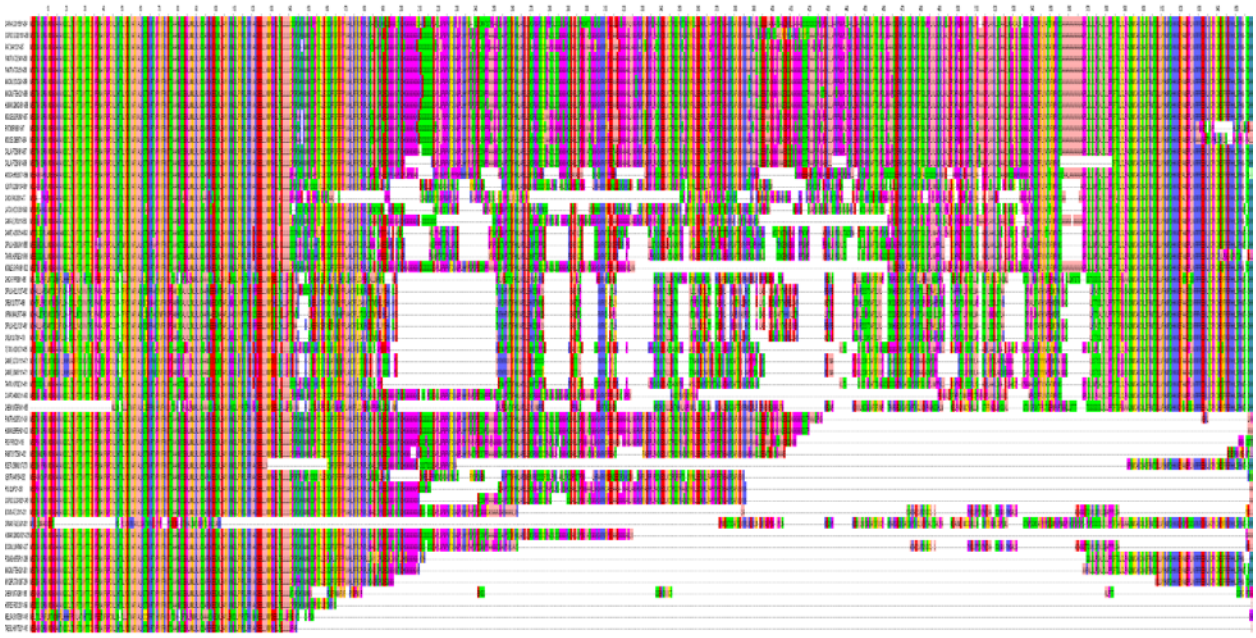
PANTR (Pan troglodytes)

SARHA (Sarcophilus harrisii)

ANOCA (Anolis carolinensis)

LATCH (Latimeria chalumnae)

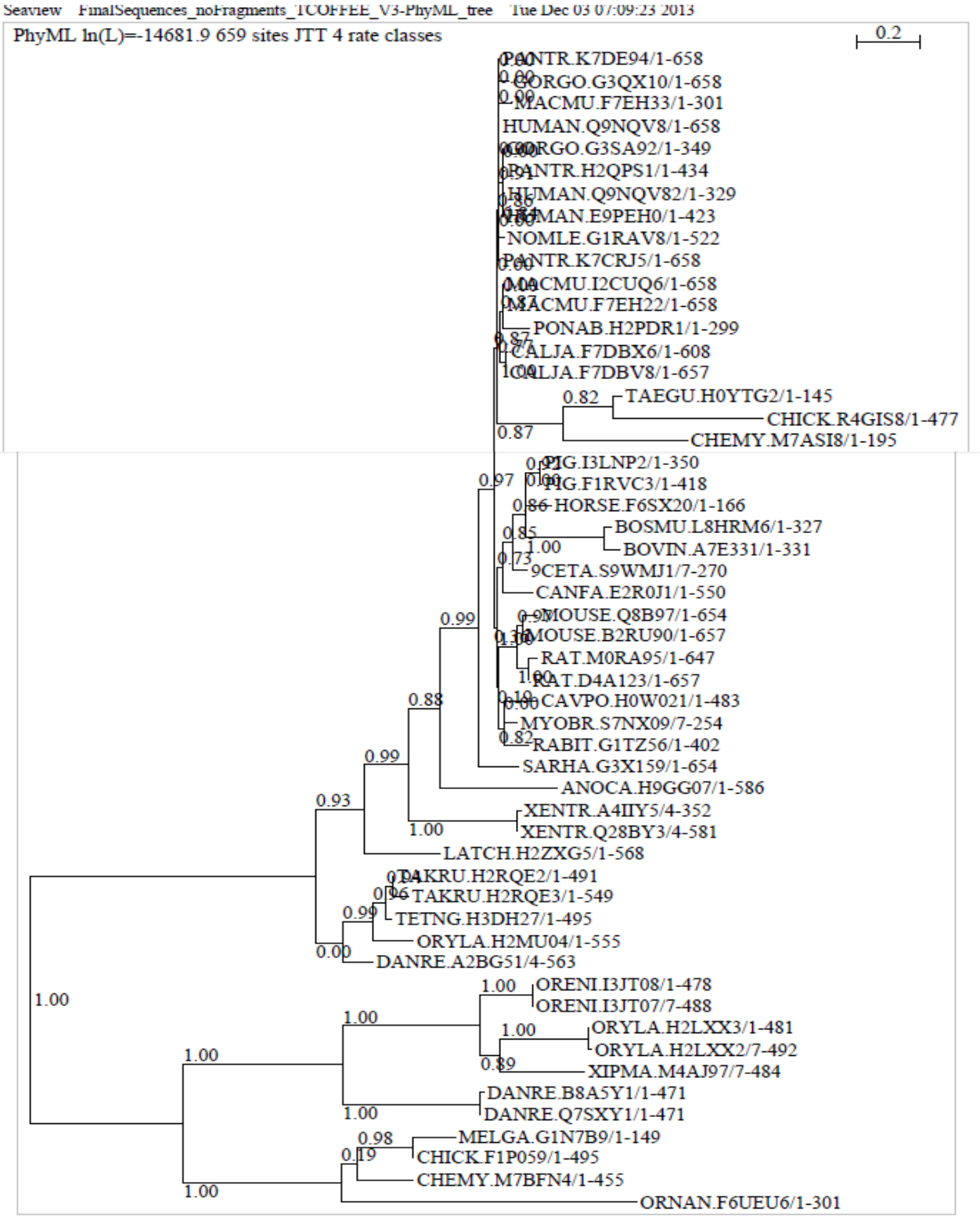
A5: Trimmed TCoffee Multiple Sequence Alignment



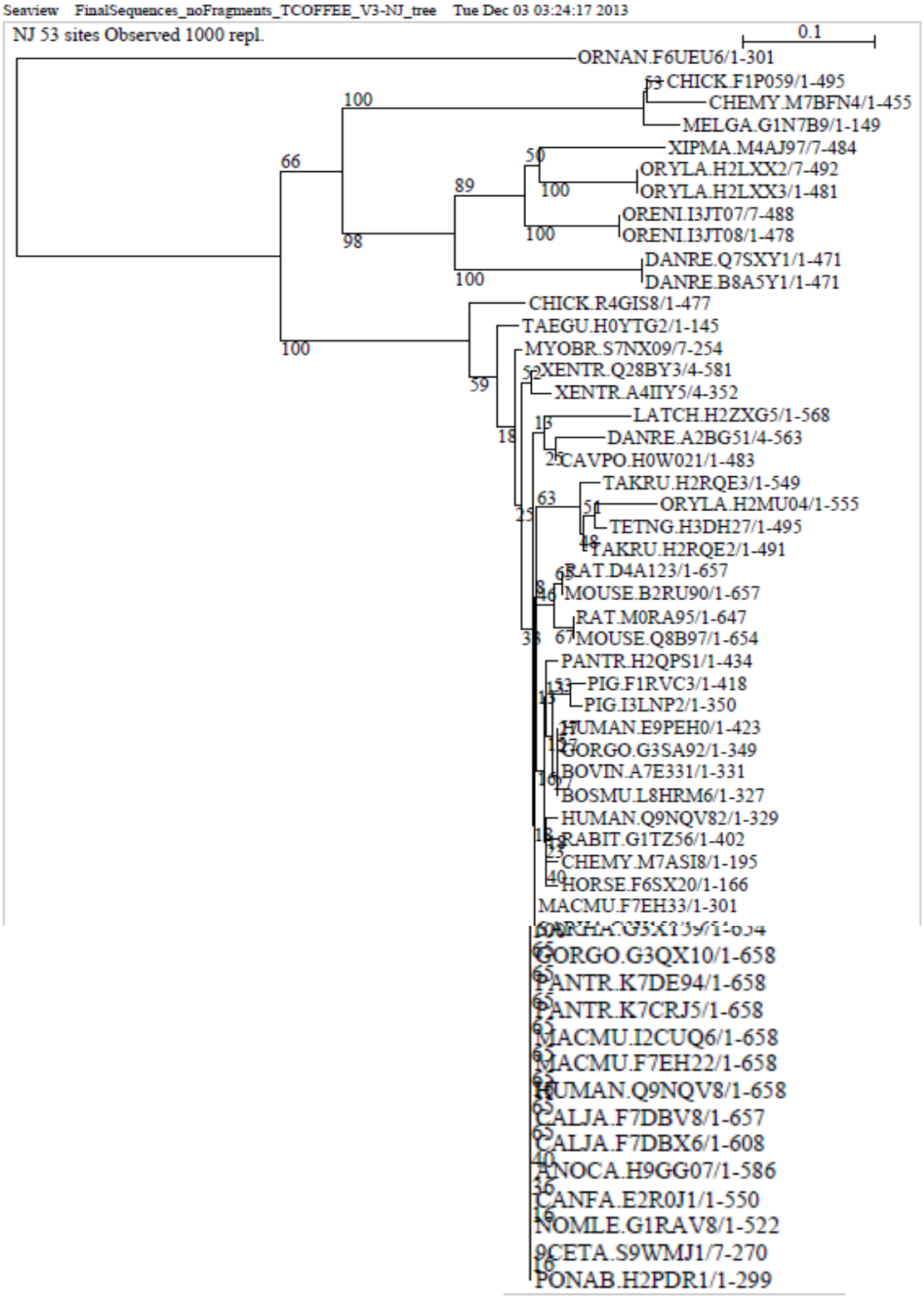
A6: Topali Model Generator Information

JTT+G Model	
AIC_1	29626.41
AIC_2	29627.04
BIC	30506.36
Alpha	0.932
pINV	N/A

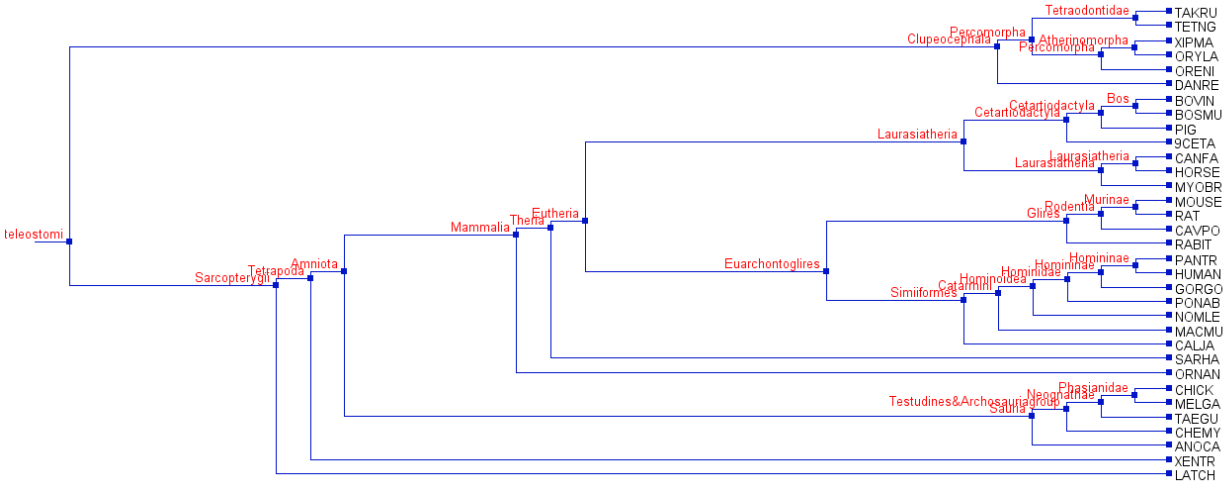
A7: Maximum Likelihood Tree Generated Using JTT+G Model

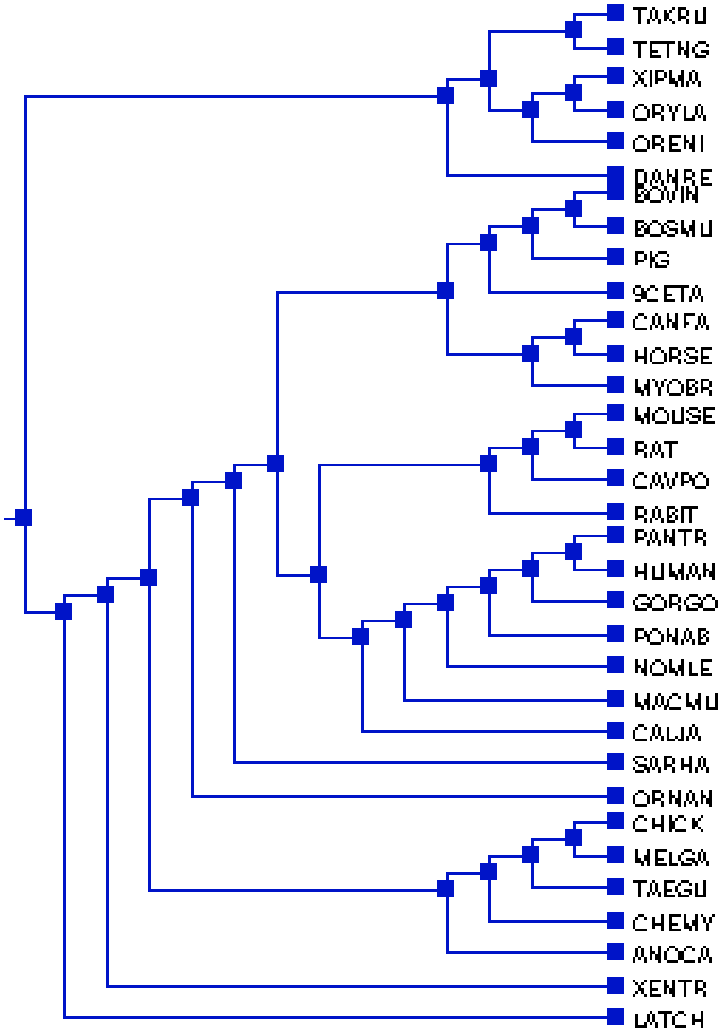


A8: Neighbor Joining Tree Generated Using 1000 Replicates

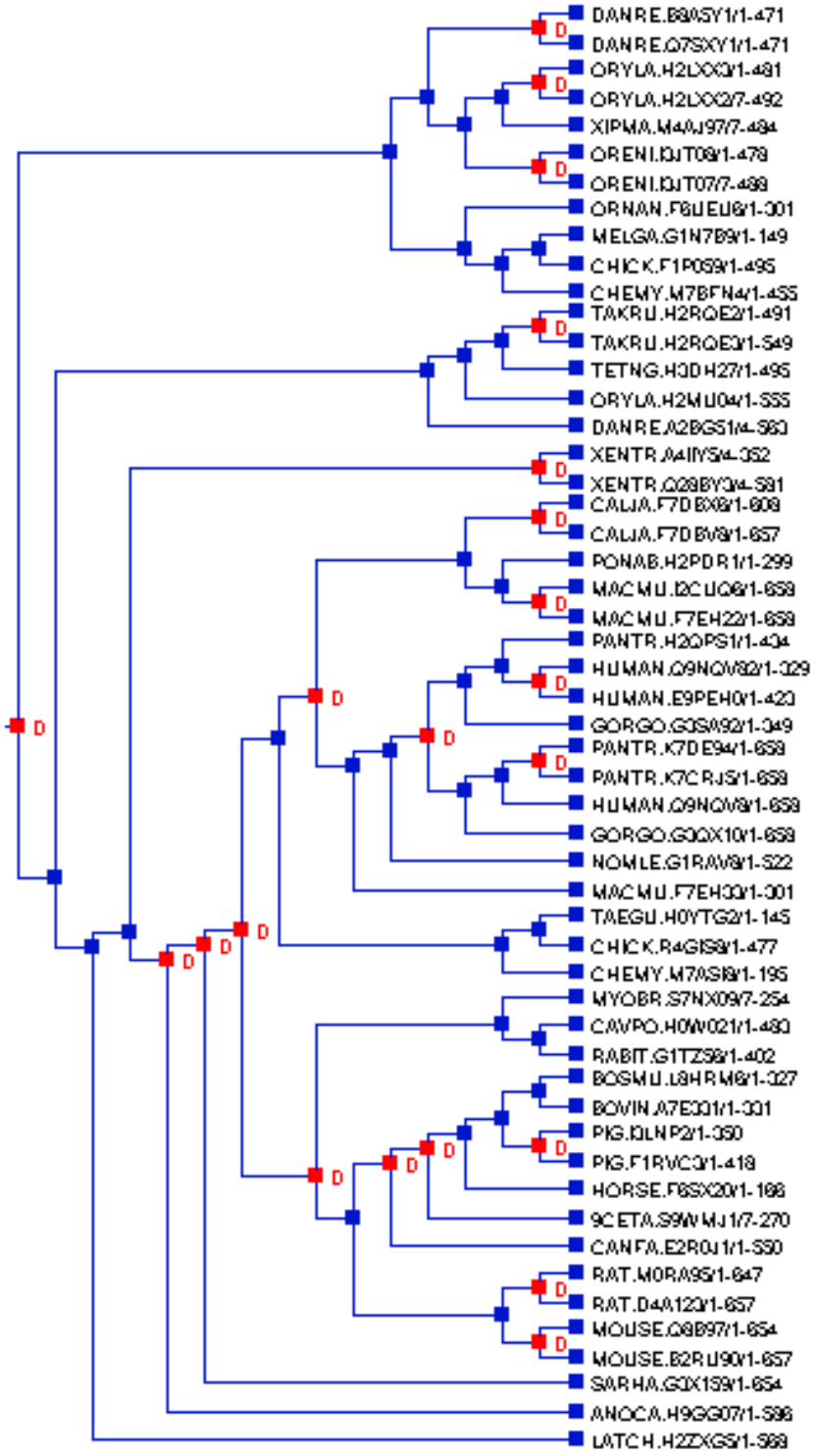


A9: Species Tree

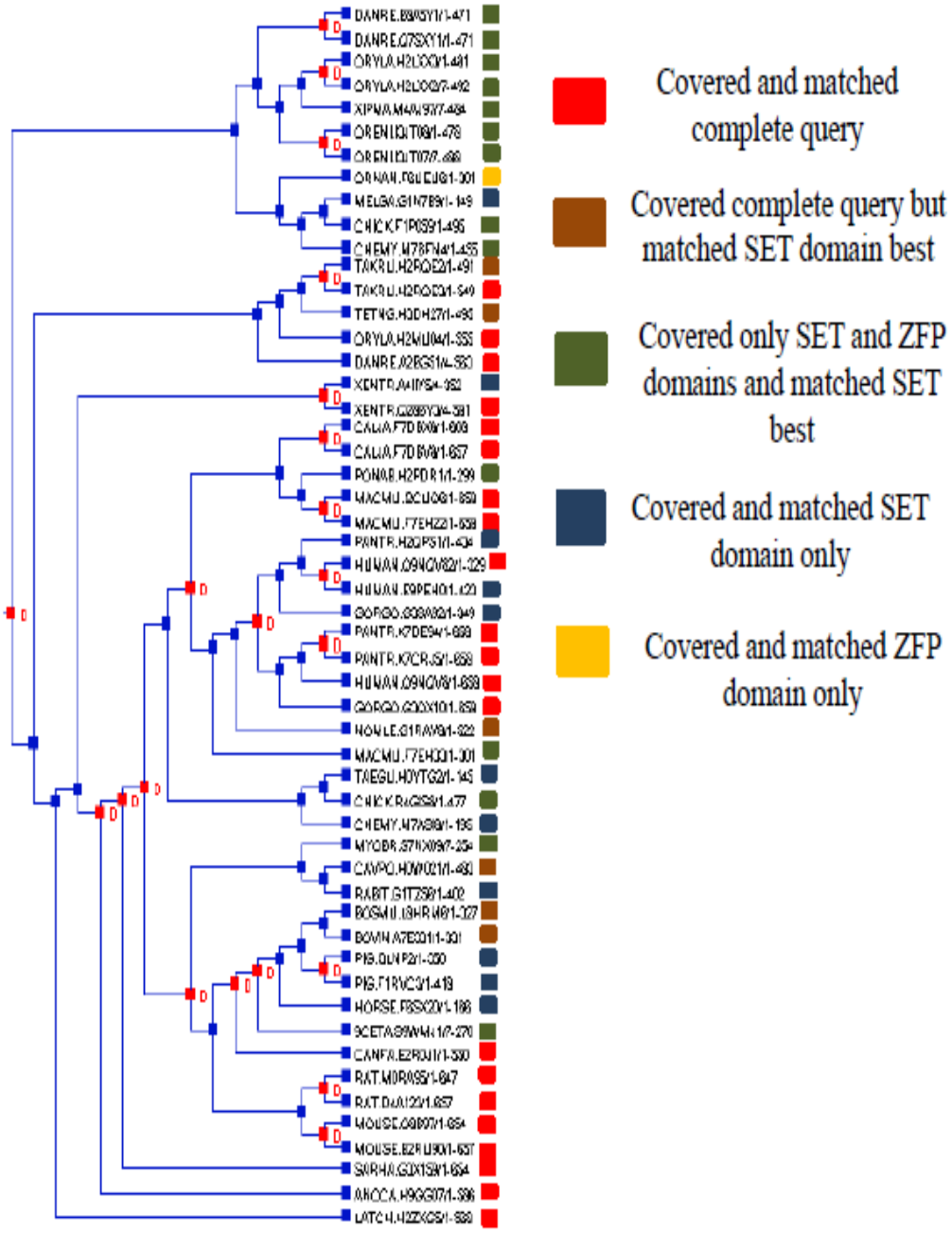




A10: Reconciled Gene Tree



A11: Reconciled Gene Tree with Coverage and Matching Labels



A12: Regenerated Maximum Likelihood Gene Tree Containing More Zfp Proteins