



**HAL**  
open science

# Evolutionary Analysis of the Mammalian Tuftelin Sequence Reveals Features of Functional Importance

S. Delgado, D. Deutsch, J. Y. Sire

► **To cite this version:**

S. Delgado, D. Deutsch, J. Y. Sire. Evolutionary Analysis of the Mammalian Tuftelin Sequence Reveals Features of Functional Importance. *Journal of Molecular Evolution*, 2017, pp.1-11. 10.1007/s00239-017-9789-5 . hal-01516886

**HAL Id: hal-01516886**

**<https://hal.sorbonne-universite.fr/hal-01516886v1>**

Submitted on 2 May 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Evolutionary analysis of the mammalian tuftelin sequence reveals features of functional importance

S. Delgado<sup>1\*</sup>, D. Deutsch<sup>2</sup> and J.Y. Sire<sup>1</sup>

<sup>1</sup>Evolution et Développement du Squelette, UMR7138- Evolution Paris-Seine, Institut de Biologie (IBPS), Université Pierre et Marie Curie, Paris, France.

<sup>2</sup> Dental Research Laboratory, Faculty of Dental Medicine, Institute of Dental Sciences, The Hebrew University of Jerusalem-Hadassah, Jerusalem, Israel.

\* corresponding author: [sidney.delgado@upmc.fr](mailto:sidney.delgado@upmc.fr)

Running title: Evolutionary analysis of TUFT1

## Keywords

Tuftelin, TUFT1, MYZAP, Evolution, Mineralization, Mammals

## Acknowledgements

This collaborative study was initiated at the Tooth Morphogenesis and Differentiation conference in Lalonde-les-Maures (France) in Spring 2013. We thank the Université Pierre et Marie Curie, CNRS and ANR (Jaws project, 12-BSV7-0020) for their financial support.

We thank Dr. Kurt Liittschwager (USA) for his English correction.

**Abstract**

Tuftelin (TUFT1) is an acidic, phosphorylated glycoprotein, initially discovered in developing enamel matrix. TUFT1 is expressed in many mineralized and non-mineralized tissues. We performed an evolutionary analysis of 82 mammalian TUFT1 sequences to identify residues and motifs that were conserved during 220 million years (Ma) of evolution. We showed that 168 residues (out of the 390 residues composing the human TUFT1 sequence) are under purifying selection. Our analyses identified several, new, putatively functional domains, and confirmed previously described functional domains, such as the TIP39 interaction domain, which correlates with nuclear localization of the TUFT1 protein, that was demonstrated in several tissues. We also identified several sites under positive selection, which could indicate evolutionary changes possibly related to the functional diversification of TUFT1 during evolution in some lineages. We discovered that TUFT1 and MYZAP (myocardial zonula adherens protein) share a common ancestor that was duplicated circa 500 million years ago. Taken together, these findings expand our knowledge of TUFT1 evolution and provide new information that will be useful for further investigation of TUFT1 functions.

## INTRODUCTION

1  
2 When initially discovered in developing and mature bovine enamel, the acidic,  
3  
4 phosphorylated glycoprotein tuftelin (TUFT1) was thought to play a major role in the  
5  
6 structural organization and mineralization of enamel (Deutsch 1989; Deutsch et al. 1991).  
7  
8 Subsequent studies focused on its putative function during enamel formation in various  
9  
10 mammalian species (Deutsch et al. 1991, 1998; Bashir et al. 1998; McDougall et al. 1998).  
11  
12 The protein was originally identified in ameloblasts and in the extracellular matrix at the  
13  
14 dentine-enamel junction (DEJ), and its pattern and timing of expression during amelogenesis  
15  
16 suggested that TUFT1 could be involved in the initial stages of enamel mineralization  
17  
18 (Deutsch et al. 1991, 1998). The lack of a signal peptide at its N-terminus was, however,  
19  
20 intriguing (Paine et al. 2000). In fact, TUFT1 is principally located in the Tomes' processes of  
21  
22 secretory ameloblasts, that are shed during amelogenesis, and this may explain the presence  
23  
24 of TUFT1 in the enamel matrix. In ameloblasts, TUFT1 was shown to form a complex with  
25  
26 the Tuftelin-Interacting-Protein 39 (TIP39) (Paine et al. 2000). Several alternatively spliced  
27  
28 tuftelin transcripts have been detected in humans and mice (Mao et al. 2001; Deutsch et al.  
29  
30 2002). Most transcripts lack a single exon but some may be missing several exons. The  
31  
32 function of the resulting isoforms is still unknown. An analysis of SNPs in the *TUFT1* gene in  
33  
34 a Turkish population suggested that some mutations were associated with a high frequency of  
35  
36 caries, supporting the role of TUFT1 in enamel mineralization (Patir et al., 2008). Moreover,  
37  
38 further studies indicated that *TUFT1* may be involved in individual predisposition to tooth  
39  
40 hypomineralization (Jeremias et al. 2013).  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50

51 Ten years after its discovery, it was demonstrated that TUFT1 expression was  
52  
53 ubiquitous; TUFT1 is expressed in many mineralized and non-mineralized tissues, and in  
54  
55 cancer cells (Mao et al. 2001; Deutsch et al. 2002; Leiser et al. 2007). These findings suggest  
56  
57 that TUFT1 could have various fundamental roles that are not restricted to tooth  
58  
59  
60  
61  
62  
63  
64  
65

1 amelogenesis. Indeed, more recently, it was demonstrated that the expression level of TUFT1  
2 mRNA was significantly higher in tissues in which oxygen levels hover closely to hypoxia  
3  
4 under normal conditions (Leiser et al. 2007). In mouse brain, in a mouse mesenchymal  
5  
6 C3H10T1/2 stem cell model, and in a neuronal PC12 cell model, TUFT1 expression was  
7  
8 found to be induced by hypoxia via HIF1a (Leiser et al. 2010; Deutsch et al. 2011). During  
9  
10 NGF-mediated PC12 differentiation, TUFT1 expression was significantly induced in  
11  
12 correlation with neurite outgrowth, and partially blocked by K252a, a selective antagonist of  
13  
14 the NGF receptor TrkA (Leiser et al. 2010), revealing additional potential physiological  
15  
16 role(s) of TUFT1. Taken together these studies suggest that TUFT1 could have various  
17  
18 functions. However, although various putative functional sites were predicted and the protein  
19  
20 sequence was well conserved in the few mammalian species studied (Mao et al. 2001;  
21  
22 Deutsch et al. 2002), a connection between TUFT1 structure and its functional sites has not  
23  
24 been demonstrated. However, the previous analysis of the TUFT1 sequence was performed  
25  
26 more than a decade ago (Deutsch et al. 2002) and it is possible that new functional sites  
27  
28 and/or domains could be identified.

29  
30  
31  
32  
33  
34  
35  
36 The aim of the present study was to identify putative functional positions and domains  
37  
38 in the TUFT1 sequence by means of evolutionary analysis of this protein in mammals, i.e.  
39  
40 covering circa 200 million years of evolution in this lineage. Such analyses, which can reveal  
41  
42 unchanged positions over long geological times and hence strong functional constraints, have  
43  
44 proven to be efficient in highlighting important positions. We also investigated putative  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77  
78  
79  
80  
81  
82  
83  
84  
85  
86  
87  
88  
89  
90  
91  
92  
93  
94  
95  
96  
97  
98  
99  
100  
101  
102  
103  
104  
105  
106  
107  
108  
109  
110  
111  
112  
113  
114  
115  
116  
117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000

## MATERIALS AND METHODS

### Tuftelin sequences and alignment

A total of 82 coding sequences of mammalian *TUFT1* were obtained from *Ensembl* [<http://www.ensembl.org>] and *NCBI* [<http://www.ncbi.nlm.nih.gov>] databases. Species names and sequence references are listed in supplementary material 1 (SM1). The dataset was built with: six published full-length sequences (human, orangutan, baboon, cow, mouse and rat); 71 computer-predicted sequences, i.e. available from the automatic analysis of sequenced, or currently being sequenced, mammalian genomes; and five sequences obtained using BLAST from the whole genome shotgun (WGS) repository sequences. *TUFT1* sequences were individually checked through alignment to published cDNA sequences, with particular attention paid to the intron/exon boundaries. When necessary, the sequences were corrected and/or completed using a BLAST search against the WGS sequences. We performed codon alignments by projecting the results of the amino acid alignment onto the nucleotide sequences using Clustal X 2.0 (Higgins et al. 1996). We chose Clustal X 2.0 because the sequences are conserved, relatively short and contain few gaps. Putative problems of alignment generated by Clustal X 2.0 analyses were also tested using MUSCLE (Edgar, 2004). The same alignment was obtained using both alignment tools. The 82 *TUFT1* sequences are provided in the supplementary material 2 (SM2). A single transcript was identified in each species. Our final alignment is available in the supplementary material 3 (SM3).

In our final alignment, only 1,000 residues out of 32,902 were missing, representing 3% of the data. The positions with missing data were included in our analyses and treated as "unknown states". Gaps were removed in our evolutionary computations.

## Sliding window and non-synonymous substitution rate (dN) analyses

To identify strong functional constraints, a sliding window analysis and a non-synonymous substitution rate analysis of nucleotide sequence variability were conducted on our alignment using HYPHY (Kosakovsky Pond et al. 2005; Kosakovsky Pond and Muse 2005; Kosakovsky Pond and Frost 2005). HYPHY utilizes Ln likelihood to measure the selective pressure. For every variable site, four quantities were computed: normalized expected numbers (ES and EN) and observed numbers (NS and NN) of synonymous and non-synonymous substitutions, respectively. HYPHY estimates  $dN = NN/EN$  and  $dS = NS/ES$ .

The p-value derived from a two-tailed extended binomial distribution is used to assess significance. The test assumes that under neutrality, a random substitution will be synonymous with probability  $P = ES/(ES+EN)$ , and computes how likely it is that P, NS out of NN+NS substitutions are synonymous. At each position, the probability for the observed data is calculated by the likelihood algorithm, taking into account phylogenetic relationships. In the two analyses, we worked with the “local model”, in which all model parameters are estimated independently for each branch, and it was computed using tree topology found in a recent mammalian phylogeny (Meredith et al. 2011).

*The non-synonymous substitution rate (dN)* was calculated using the maximum likelihood method based on the HKY 85 model (Hasegawa et al. 1985).

*The Sliding Window* (available in “standard analyses” section and “Miscellaneous” subsection in HYPHY software) is a method to calculate the mean substitution rate along a protein sequence. This graphical representation is used to visualize selective pressures along the protein. Indeed, nucleotide diversity reflects selective constraints: the lower the variability, the higher the selection, and vice versa.

We chose the following parameters:

1 - the mean substitution rate was calculated using the maximum likelihood (ML) method based  
2 on the HKY 85 model (Hasegawa et al. 1985);  
3

4 - probabilities are calculated for a window of 15 bp with an overlap of 5 bp between each pair  
5 of windows. Other investigators have chosen larger windows for their analyses (e.g. Endo et  
6 al. 1996; Tsunoyama and Gojobori 1998; Schmid and Yang 2008) but, when applying the  
7 HYPHY method, it is not necessary to use large sliding windows and it is even recommended  
8 to avoid large windows, which produce a “smoothing” effect that could result in loss of  
9 evolutionary information.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20

### 21 **Distance tree**

22 The alignment was treated by MEGA 5.2.2 (Tamura et al. 2011) software  
23  
24 (<http://www.megasoftware.net>). The phylogenetic reconstructions used Neighbor-joining and  
25  
26 Maximum Likelihood methods, with Dayhoff model and a rate of substitution Gamma  
27 distributed (in both NJ and ML).  
28  
29  
30  
31  
32  
33  
34  
35

### 36 **Purifying selection analysis**

37 The search for site-specific purifying selection (i.e. biologically significant amino acids) in  
38 TUFT1 was carried out using the Consurf Server 2.4 (<http://consurf.tau.ac.il/>) (Ashkenazy et  
39 al. 2010; Celniker et al. 2013). The analysis was performed by comparing a null model, i.e. no  
40 purifying selection, and a model allowing purifying selections (HKY 85 model: Hasegawa et  
41 al. 1985). The results were then displayed on the human sequence. Different levels of  
42 purifying selection were indicated by a set of colours.  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55

### 56 **Positive selection analysis**

57  
58  
59  
60  
61  
62  
63  
64  
65



1 The search for site-specific positive selection in TUFT1 was carried out using the Selecton  
2 Server (<http://selecton.tau.ac.il/>) (Stern et al. 2007). The Selecton analysis was computed  
3  
4 using tree topology in a recent mammalian phylogeny (Meredith et al. 2011). The analysis  
5  
6 was performed using the M8 model (Yang et al. 2000). A proportion  $p_0$  of the sites was  
7  
8 drawn from a beta distribution (which is defined in the interval  $[0,1]$ ), and a proportion  
9  
10  $p_1 (= 1 - p_0)$  of the sites was drawn from an additional category  $\omega_s$  (which is constrained to be  
11  
12  $\geq 1$ ). Thus, sites drawn from the beta distribution were sites experiencing purifying selection,  
13  
14 whereas sites drawn from the  $\omega_s$  category are sites experiencing either neutral or positive  
15  
16 selection. Both  $p_0$  and  $\omega_s$  are estimated using ML. The results were then displayed on the  
17  
18 human sequence. Different levels of positive selection were indicated by a set of two colours.  
19  
20  
21  
22  
23

24 Model = Positive selection enabled (M8, beta +  $w \geq 1$ ); number of categories = 8.  
25  
26  
27  
28

### 29 **Putative functional sites**

30  
31 Search for post-translationally modified sites in the human TUFT1 sequence was performed  
32  
33 using the Prosite database [Sigrist et al. 2010; <http://prosite.expasy.org/>] to identify putative  
34  
35 N-glycosylation and phosphorylation sites.  
36  
37  
38  
39  
40

### 41 **Date calibration**

42  
43 Date calibration of the mammalian lineage was obtained from the “Fossil Calibration  
44  
45 Database” [<http://fossilcalibrations.org>]. The mammalian node is dated from 164.9 Ma to  
46  
47 201.5 Ma.  
48  
49  
50  
51

### 52 **Relationships**

53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 To improve our understanding of TUFT1 origins and possible functions, its putative  
2 relationships with other proteins were searched for by using PSI-BLAST (Position-specific  
3 iterative Blast) tool in NCBI site.  
4  
5  
6  
7  
8  
9

## 10 RESULTS

### 11 Alignment and sequence comparisons

12 The 82 TUFT1 sequences studied here represent 59 families distributed in 14 orders  
13 (including 1 Monotremata, 3 Marsupialia and 78 Placentalia species) and they are  
14 representative of the current mammalian phylogenetic diversity (SM1). The length of the  
15 TUFT1 sequences ranged from 383 residues in *Choloepus hoffmanni* (sloth) to 399 in  
16 *Echinops telfairi* (tenrec) with 390 amino acids in humans (SM2). Alignment of these  
17 sequences against the human sequence resulted in a 419 amino acid sequence that included 29  
18 gaps (SM3). Our alignment indicated that TUFT1 is composed of a succession of conserved  
19 and variable regions and revealed only a few insertions of amino acids, the largest being eight  
20 residues inserted in the region encoded by the 3' end of exon 1 in golden mole and five  
21 residues by the 3' end of exon 6 in tenrec (two Afrotheria species). Insertions and deletions of  
22 one to three residues were found in a few sequences and no sequence repeats were identified.  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43

### 44 Purifying selection

#### 45 *Analysis of non-synonymous substitution rate (Fig. 1A) and Sliding Window analysis (Fig.* 46 *1B)*

47 The dN/dS ratio analysis revealed that TUFT1 sequence is characterized by regions of weak  
48 selective pressures alternating with regions of strong selective pressures (Fig. 1A, B). The  
49 regions subjected to strong functional constraints display low -Ln likelihood and low dN  
50 values. The lower these values are, the more important the selective constraint was (for this  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 analysis the limits were determined arbitrarily; for example  $-\ln$  Likelihood  $< 250-200$  and  $dN$   
2  $< 0.6$ ). The strongest constraints were found in the regions/residues encoded by exon 3, exon  
3  
4 4, the 5' end of exon 5, and exons 10 to 12. In contrast, the regions encoded by exons 6 to 9  
5  
6 were found to be less conserved, meaning they were less subjected to functional constraints.  
7

### 8 ***Distance tree***

9  
10 The phylogenetic reconstruction was performed to (i) determine the rate of TUFT1 evolution  
11  
12 and (ii) check anomalies (long branches) occurring during evolution in some mammalian  
13  
14 lineages to avoid using a sequence that is no longer under selective pressure (i.e. a  
15  
16 pseudogene). Indeed, if TUFT1 sequences were conserved during mammalian evolution,  
17  
18 phylogenetic information would still be present and we would expect to obtain a phylogenetic  
19  
20 tree similar to that currently considered for mammalian relationships; otherwise, we would  
21  
22 expect to obtain long branches along with incorrect topology. We present an NJ tree in figure  
23  
24 1, but the ML tree is available in SM9.  
25  
26  
27  
28  
29  
30

31 The distance tree indicates a good equilibrium between conserved and variable regions  
32  
33 of TUFT1 sequences (Fig. 2).  
34  
35

36 A few *TUFT1* sequences display long branches, a feature which means a more rapid  
37  
38 accumulation of substitutions without changing the phylogenetic relations. This is the case for  
39  
40 manatee (*Trichechus manatus*) and tenrec (*Echinops telfairi*) in Afrotheria, bushbaby  
41  
42 (*Otolemur garnettii*) in Primates, guinea pig (*Cavia porcellus*) in Rodentia, hedgehog  
43  
44 (*Erinaceus europaeus*) in Laurasiatheria, and opossum (*Monodelphis domestica*) in  
45  
46 Marsupiala. Very few sequences are not located at the correct phylogenetic position such as  
47  
48 the shrew (*Sorex araneus*) and the kangaroo rat *Dipodomys ordii*. However, the *TUFT1*  
49  
50 sequence of *Dipodomys ordii* was not complete enough to have a good signal (SM2).  
51  
52  
53  
54  
55

### 56 ***Consurf analysis***

57  
58 The statistical tests supporting this analysis are available in SM6.  
59  
60  
61  
62  
63  
64  
65

Purifying selection detected by Selecton analysis at the amino acid level identified TUFT1 residues that were well conserved during mammalian evolution, i.e. that are predicted to play an important functional or structural role, or both (Fig. 3, SM4). A total of 121 conserved (= unchanged) and 47 conservative (= that could be only replaced with a residue possessing similar properties) positions were identified. These important positions represent more than 43% of the 390 residues composing the human TUFT1 sequence. These residues are either isolated or regrouped, forming motifs of various lengths:

(1) At the N-terminus, the motif <sup>1</sup>MNGT contains a putative N-glycosylation site (<sup>2</sup>N).

(2) The N-terminal region includes a large motif encoded by the 3' region of exon 3 to the 5' region of exon 5. It is composed of 57 residues, starting at <sup>64</sup>S and ending at <sup>121</sup>S. This motif contains 33 unchanged or conservative positions. The function of this motif is unknown to date, with the exception of the last residue <sup>121</sup>S, which is putatively phosphorylated (protein kinase C site).

(3) The C-terminal region contains another large motif, encoded by exons 10 to 12, consisting of 74 residues, from <sup>278</sup>E to <sup>351</sup>Q. This motif possesses 55 unchanged or conservative positions. Most of this motif is known to belong to the TIP-39 binding site.

(4) At the C-terminal extremity encoded by exon 13, the <sup>381</sup>PmPvIRVVET motif of unknown function is well conserved.

In addition to these motifs, our evolutionary analysis highlighted many conserved or conservative residues, hence those having a putative function. Among them, some were predicted as functional by *Prosites* analysis in the human sequence (see below, Fig. 4).

### Positive selection

Using the Selecton server, 13 positions were detected as positively selected during mammalian evolution, i.e. sites drawn from the  $\omega$ s category (Stern et al., 2007). Shades of

1 yellow indicate  $\omega > 1$ , with dark-yellow representing sites where reliable positive selection was  
 2 inferred, and light-yellow representing positive selection that was not statistically significant  
 3 (SM4, SM8). These 13 positions were indicated on the human sequence (Fig. 3). One is  
 4 isolated, at <sup>255</sup>V, and 12 are regrouped, forming three motifs under positive selection:  
 5  
 6  
 7  
 8  
 9  
 10 <sup>131</sup>SLHR; <sup>153</sup>AIYssPP; <sup>163</sup>TCI. One of them contained a “casein kinase II” phosphorylation  
 11 site predicted by *Prosite* on the human sequence (see below).  
 12  
 13  
 14  
 15  
 16

### 17 **Prediction of post translation modifications**

18  
 19 *Prosite* could not predict any functional sites, however, it predicted 14 sites that undergo post-  
 20 translation modifications, and hence these positions may be involved in TUFT1 function  
 21 (SM5). Most of these sites were also predicted in a previous analysis (Deutsch et al. 2002).  
 22  
 23  
 24 These were: two N-glycosylation sites (positions 2 and 356), six casein kinase II (CK2)  
 25 phosphorylation sites (phosphoserines 122, 157 and phosphothreonines 9, 35, 175, 322), and  
 26  
 27 four protein kinase (PKC) phosphorylation sites (phosphoserines 121, 171, 370, 378). One  
 28 amidation site (46) and one N-myristoylation site (355), close to the N-gly 356, were also  
 29 predicted. All the *Prosite*-predicted sites were reported on the 3D TUFT1 sequence published  
 30 by Deutsch et al. (2002) including sites that were considered as functional in our evolutionary  
 31 analysis (Fig. 4). Cysteines were also indicated in this figure.  
 32  
 33  
 34  
 35  
 36  
 37  
 38  
 39  
 40  
 41  
 42  
 43  
 44  
 45

### 46 **Relationships of TUFT1**

47  
 48 Using PSI-BLAST with the human TUFT1, we found a single related protein, MYZAP  
 49 (myocardial zonula adherens protein), that is known in many vertebrate taxa (including  
 50  
 51  
 52  
 53  
 54  
 55  
 56  
 57  
 58  
 59  
 60  
 61  
 62  
 63  
 64  
 65  
 66  
 67  
 68  
 69  
 70  
 71  
 72  
 73  
 74  
 75  
 76  
 77  
 78  
 79  
 80  
 81  
 82  
 83  
 84  
 85  
 86  
 87  
 88  
 89  
 90  
 91  
 92  
 93  
 94  
 95  
 96  
 97  
 98  
 99  
 100  
 101  
 102  
 103  
 104  
 105  
 106  
 107  
 108  
 109  
 110  
 111  
 112  
 113  
 114  
 115  
 116  
 117  
 118  
 119  
 120  
 121  
 122  
 123  
 124  
 125  
 126  
 127  
 128  
 129  
 130  
 131  
 132  
 133  
 134  
 135  
 136  
 137  
 138  
 139  
 140  
 141  
 142  
 143  
 144  
 145  
 146  
 147  
 148  
 149  
 150  
 151  
 152  
 153  
 154  
 155  
 156  
 157  
 158  
 159  
 160  
 161  
 162  
 163  
 164  
 165  
 166  
 167  
 168  
 169  
 170  
 171  
 172  
 173  
 174  
 175  
 176  
 177  
 178  
 179  
 180  
 181  
 182  
 183  
 184  
 185  
 186  
 187  
 188  
 189  
 190  
 191  
 192  
 193  
 194  
 195  
 196  
 197  
 198  
 199  
 200  
 201  
 202  
 203  
 204  
 205  
 206  
 207  
 208  
 209  
 210  
 211  
 212  
 213  
 214  
 215  
 216  
 217  
 218  
 219  
 220  
 221  
 222  
 223  
 224  
 225  
 226  
 227  
 228  
 229  
 230  
 231  
 232  
 233  
 234  
 235  
 236  
 237  
 238  
 239  
 240  
 241  
 242  
 243  
 244  
 245  
 246  
 247  
 248  
 249  
 250  
 251  
 252  
 253  
 254  
 255  
 256  
 257  
 258  
 259  
 260  
 261  
 262  
 263  
 264  
 265  
 266  
 267  
 268  
 269  
 270  
 271  
 272  
 273  
 274  
 275  
 276  
 277  
 278  
 279  
 280  
 281  
 282  
 283  
 284  
 285  
 286  
 287  
 288  
 289  
 290  
 291  
 292  
 293  
 294  
 295  
 296  
 297  
 298  
 299  
 300  
 301  
 302  
 303  
 304  
 305  
 306  
 307  
 308  
 309  
 310  
 311  
 312  
 313  
 314  
 315  
 316  
 317  
 318  
 319  
 320  
 321  
 322  
 323  
 324  
 325  
 326  
 327  
 328  
 329  
 330  
 331  
 332  
 333  
 334  
 335  
 336  
 337  
 338  
 339  
 340  
 341  
 342  
 343  
 344  
 345  
 346  
 347  
 348  
 349  
 350  
 351  
 352  
 353  
 354  
 355  
 356  
 357  
 358  
 359  
 360  
 361  
 362  
 363  
 364  
 365  
 366  
 367  
 368  
 369  
 370  
 371  
 372  
 373  
 374  
 375  
 376  
 377  
 378  
 379  
 380  
 381  
 382  
 383  
 384  
 385  
 386  
 387  
 388  
 389  
 390  
 391  
 392  
 393  
 394  
 395  
 396  
 397  
 398  
 399  
 400  
 401  
 402  
 403  
 404  
 405  
 406  
 407  
 408  
 409  
 410  
 411  
 412  
 413  
 414  
 415  
 416  
 417  
 418  
 419  
 420  
 421  
 422  
 423  
 424  
 425  
 426  
 427  
 428  
 429  
 430  
 431  
 432  
 433  
 434  
 435  
 436  
 437  
 438  
 439  
 440  
 441  
 442  
 443  
 444  
 445  
 446  
 447  
 448  
 449  
 450  
 451  
 452  
 453  
 454  
 455  
 456  
 457  
 458  
 459  
 460  
 461  
 462  
 463  
 464  
 465  
 466  
 467  
 468  
 469  
 470  
 471  
 472  
 473  
 474  
 475  
 476  
 477  
 478  
 479  
 480  
 481  
 482  
 483  
 484  
 485  
 486  
 487  
 488  
 489  
 490  
 491  
 492  
 493  
 494  
 495  
 496  
 497  
 498  
 499  
 500  
 501  
 502  
 503  
 504  
 505  
 506  
 507  
 508  
 509  
 510  
 511  
 512  
 513  
 514  
 515  
 516  
 517  
 518  
 519  
 520  
 521  
 522  
 523  
 524  
 525  
 526  
 527  
 528  
 529  
 530  
 531  
 532  
 533  
 534  
 535  
 536  
 537  
 538  
 539  
 540  
 541  
 542  
 543  
 544  
 545  
 546  
 547  
 548  
 549  
 550  
 551  
 552  
 553  
 554  
 555  
 556  
 557  
 558  
 559  
 560  
 561  
 562  
 563  
 564  
 565  
 566  
 567  
 568  
 569  
 570  
 571  
 572  
 573  
 574  
 575  
 576  
 577  
 578  
 579  
 580  
 581  
 582  
 583  
 584  
 585  
 586  
 587  
 588  
 589  
 590  
 591  
 592  
 593  
 594  
 595  
 596  
 597  
 598  
 599  
 600  
 601  
 602  
 603  
 604  
 605  
 606  
 607  
 608  
 609  
 610  
 611  
 612  
 613  
 614  
 615  
 616  
 617  
 618  
 619  
 620  
 621  
 622  
 623  
 624  
 625  
 626  
 627  
 628  
 629  
 630  
 631  
 632  
 633  
 634  
 635  
 636  
 637  
 638  
 639  
 640  
 641  
 642  
 643  
 644  
 645  
 646  
 647  
 648  
 649  
 650  
 651  
 652  
 653  
 654  
 655  
 656  
 657  
 658  
 659  
 660  
 661  
 662  
 663  
 664  
 665  
 666  
 667  
 668  
 669  
 670  
 671  
 672  
 673  
 674  
 675  
 676  
 677  
 678  
 679  
 680  
 681  
 682  
 683  
 684  
 685  
 686  
 687  
 688  
 689  
 690  
 691  
 692  
 693  
 694  
 695  
 696  
 697  
 698  
 699  
 700  
 701  
 702  
 703  
 704  
 705  
 706  
 707  
 708  
 709  
 710  
 711  
 712  
 713  
 714  
 715  
 716  
 717  
 718  
 719  
 720  
 721  
 722  
 723  
 724  
 725  
 726  
 727  
 728  
 729  
 730  
 731  
 732  
 733  
 734  
 735  
 736  
 737  
 738  
 739  
 740  
 741  
 742  
 743  
 744  
 745  
 746  
 747  
 748  
 749  
 750  
 751  
 752  
 753  
 754  
 755  
 756  
 757  
 758  
 759  
 760  
 761  
 762  
 763  
 764  
 765  
 766  
 767  
 768  
 769  
 770  
 771  
 772  
 773  
 774  
 775  
 776  
 777  
 778  
 779  
 780  
 781  
 782  
 783  
 784  
 785  
 786  
 787  
 788  
 789  
 790  
 791  
 792  
 793  
 794  
 795  
 796  
 797  
 798  
 799  
 800  
 801  
 802  
 803  
 804  
 805  
 806  
 807  
 808  
 809  
 810  
 811  
 812  
 813  
 814  
 815  
 816  
 817  
 818  
 819  
 820  
 821  
 822  
 823  
 824  
 825  
 826  
 827  
 828  
 829  
 830  
 831  
 832  
 833  
 834  
 835  
 836  
 837  
 838  
 839  
 840  
 841  
 842  
 843  
 844  
 845  
 846  
 847  
 848  
 849  
 850  
 851  
 852  
 853  
 854  
 855  
 856  
 857  
 858  
 859  
 860  
 861  
 862  
 863  
 864  
 865  
 866  
 867  
 868  
 869  
 870  
 871  
 872  
 873  
 874  
 875  
 876  
 877  
 878  
 879  
 880  
 881  
 882  
 883  
 884  
 885  
 886  
 887  
 888  
 889  
 890  
 891  
 892  
 893  
 894  
 895  
 896  
 897  
 898  
 899  
 900  
 901  
 902  
 903  
 904  
 905  
 906  
 907  
 908  
 909  
 910  
 911  
 912  
 913  
 914  
 915  
 916  
 917  
 918  
 919  
 920  
 921  
 922  
 923  
 924  
 925  
 926  
 927  
 928  
 929  
 930  
 931  
 932  
 933  
 934  
 935  
 936  
 937  
 938  
 939  
 940  
 941  
 942  
 943  
 944  
 945  
 946  
 947  
 948  
 949  
 950  
 951  
 952  
 953  
 954  
 955  
 956  
 957  
 958  
 959  
 960  
 961  
 962  
 963  
 964  
 965  
 966  
 967  
 968  
 969  
 970  
 971  
 972  
 973  
 974  
 975  
 976  
 977  
 978  
 979  
 980  
 981  
 982  
 983  
 984  
 985  
 986  
 987  
 988  
 989  
 990  
 991  
 992  
 993  
 994  
 995  
 996  
 997  
 998  
 999  
 1000

1 protein that does not possess a signal peptide, like TUFT1. MYZAP is expressed in adherens  
 2 junctions of myocardial and vascular endothelial cells, and in junctions of various epithelia  
 3  
 4 (Rickelt et al. 2011; Pieperhoff et al. 2012). The alignment of TUFT1 and MYZAP points to  
 5  
 6 only a few similarities.  
 7  
 8  
 9

## 10 11 **DISCUSSION**

### 12 **Purifying selection**

13  
 14 During millions of years of evolution, purifying selection does not retain mutations that lead to  
 15  
 16 deleterious effects. Such selection results in the preservation of all sensitive positions, i.e.  
 17  
 18 those possessing amino acids that play either functionally or structurally important roles.  
 19  
 20 Residues are either conserved or conservative. In the TUFT1 sequences, our evolutionary  
 21  
 22 analysis identified a total of 168 conserved and conservative positions during 164.9 Ma of  
 23  
 24 mammalian evolution indicating that the important functions supported by these positions  
 25  
 26 were present earlier in TUFT1 evolution, in non-mammalian vertebrates. We compare the  
 27  
 28 putative functional sites predicted by *Prosite* with the crucial positions revealed by our  
 29  
 30 evolutionary analysis, then we discuss the presence of various conserved domains (Figs 3, 4,  
 31  
 32 SM5).  
 33  
 34  
 35  
 36  
 37  
 38  
 39  
 40

### 41 ***Several predicted functional sites are confirmed by evolutionary analysis***

42  
 43 *N-glycosylated residues.* - The two predicted N-linked glycosylation sites (N-gly 2, <sup>2</sup>NGTR  
 44  
 45 and N-gly 356, <sup>356</sup>NFST) are under purifying selection, which strongly suggests that these  
 46  
 47 post-translational modifications (attachment of a glycan, N-acetylglucosamine, to asparagine)  
 48  
 49 of TUFT1 are functionally and/or structurally important.  
 50  
 51

52  
 53 *N-myristoylation site.* - The predicted N-myristoylation site (N-Myr 355, <sup>355</sup>GNfsTQ), located  
 54  
 55 in the C-terminal region, near the second N-Gly conserved site, was not validated by our  
 56  
 57  
 58  
 59  
 60  
 61  
 62  
 63  
 64  
 65

analysis. Indeed, the glycine (G), which is the crucial, post-translationally modified residue of such a site in the N-terminal region, is located at a variable position.

*Phosphorylated sites.* - Our analysis indicated that only three out of the ten putative phosphorylated residues predicted by Prosite are under purifying selection. Only phosphorylation of serines by protein kinase C (PKC) were validated by our analysis, through conservation of the position housing a serine: Ser-P 121, <sup>121</sup>SsK; Ser-P 171, <sup>171</sup>SIR; Ser-P 378, <sup>378</sup>SpK. Previous post-translational modification analyses using MS/MS sequencing of the recombinant human TUFT1 protein produced in a eukaryotic system, also revealed one phosphorylated site at Serine 378 (Shay et al., 2009).

None of the casein kinase II phosphorylation sites were confirmed. In a previous study, Deutsch et al. (2002) suggested that phosphorylation sites might play an important role in providing potential sites for specific chelation of calcium ions, which could explain why TUFT1 is presumed to play an important role in enamel mineralization. Our findings could support such a role, although several, previously predicted phosphorylation sites were not conserved during mammalian evolution.

*Amidation site.* - The amidation site (amidation 46, <sup>46</sup>aGRK) predicted by Prosite is not under purifying selection since all residues of this motif, including the glycine residue that provides the amine group, are located at variable positions according to our analysis.

#### ***Evaluation of previously identified functional domains by evolutionary analysis***

*TUFT1-TIP39 interaction domain.* - More than 15 years ago, Paine et al. (2000) showed that TUFT1 interacts with TIP39 (Tuftelin Interacting Protein – 39 kDa; also known as TFIP11 - Tuftelin Interacting Protein 11). These authors showed that the region of interaction with TIP39 is located near the TUFT1 C-terminus, between amino acids 294 and 348 (Paine et al. 2000). This part of the protein was also suggested to be responsible for TUFT1 self-assembly (Paine et al. 2000). Our analysis confirmed the importance of this TUFT1 region because 42

1 out of the 54 amino acids putatively implicated in this interaction were conserved. Only 12  
2 amino acids are at a variable position. In contrast, most of the conserved and conservative  
3 residues are clustered in long stretches within this region, a finding that indicates a crucial  
4 role. In addition, we also found that several conserved residues are located at both extremities  
5 of this domain (residues 285-293 and 349-360), which indicates that the functional region is  
6 probably wider than the region previously identified by Paine et al. (2000). TIP39 (TFIP11) is  
7 a nuclear speckle-localized protein that may play a role in spliceosome disassembly in Cajal  
8 bodies (Wen et al. 2005; Stanek et al. 2008). In adult tissues, TUFT1 is expressed mainly in  
9 the cytoplasm. However, nuclear localization of the protein was demonstrated in several  
10 tissues (Deutsch et al. 2002; Leiser et al. 2007). During early mouse embryonic development,  
11 TUFT1 was detected mainly in the cytoplasm, while at later embryonic stages and post-  
12 natally, its expression in neuronal cells is concentrated in the perinuclear/nuclear region  
13 (Deutsch unpublished; Shilo et al. unpublished), possibly indicating interaction with TIP39.  
14  
15 *Calcium-binding domain.* - An EF-hand, calcium-binding domain was previously suggested  
16 in human, bovine, and murine TUFT1, from residue 125 to 137 (Mao et al. 2001; Deutsch et  
17 al. 2002). However, no experimental results supported this motif. Our analysis shows that  
18 only one of the residues in this region is conserved in mammals, a finding which indicates  
19 that this region lacks an important function (but see below, *Positive selection*).

### 43 ***Domains of unknown function***

44 Our evolutionary analysis predicts that four additional motifs could play an important role not  
45 previously identified, either through protein analysis software or in other previous studies.  
46 Indeed, these positions are subjected to strong purifying selection. Three are located between  
47 residues 64 and 121, whereas the fourth constitutes the C-terminus of TUFT1. Each of them  
48 possesses one or two glutamic acid (E) and one lysine (K) or arginine (R). Remarkable  
49 positions highlighted by our evolutionary analyses are summarized in SM7.  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



## Positive selection

Positive selection means that one allele was selected because it improves fitness. Therefore, positive selection increases the prevalence of adaptive traits. The results of purifying selection are generally easy to interpret. However, this is not the case for the results of positive selection, because the residues detected as being positively selected are often not related to a motif, and hence interpretation of a putative role is difficult. Another problem of positive selection is the possibility of obtaining false positive results. However, our analysis of TUFT1 indicates that several positively-selected positions are located in close proximity to one another, and belong to two putative motifs, which were not under purifying selection.

Four positions under positive selection (<sup>131</sup>SLHR) are located in a TUFT1 region that was previously described as a “calcium-binding domain”, but predicted to be variable in our evolutionary analysis. We do not know the role of this motif, however, its unknown function was acquired recently in mammalian history, most probably in placental mammals, which could explain why that motif was not detected as being important in our analysis. Therefore, the potential role of TUFT1 in mineralisation (Deutsch et al. 2002) could be a recent feature, on the geological scale.

The second motif revealed as being positively selected is the <sup>157</sup>sPPE encoded by the 3' region of exon 6. This short sequence was predicted by Prosite as a casein kinase II phosphorylation site (SppE) but not validated by our evolutionary analysis of conserved positions. As discussed above, the positive selection of the two prolines probably occurred recently in the mammalian lineage and could mean that this motif is functional in placental mammals for example, strengthening the phosphorylation pattern of TUFT1 (Deutsch et al. 2002). A mutated enamelin phosphorylation site causes amelogenesis imperfecta (Chan et al. 2010), which emphasizes the importance of conservation of phosphorylation pattern during evolution of secretory calcium-binding phosphoproteins (Al-Hashimi et al. 2009; Silvent et al.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

2013). Our results confirmed the importance of mineralization among the different functions of these ubiquitous proteins by reinforcement of phosphorylation.

In addition, the presence of a positively selected valine at position 255, close to the well-conserved <sup>256</sup>ALEE motif, could increase the fitness and the role of this highly conserved motif, although its role is not yet known.

Finally, the putative role of the <sup>163</sup>TCI motif encoded by exon 7 and composed of three positively selected residues is unknown. The presence of a cysteine could be important if it were involved in the formation of a disulfide bond that could promote a new tertiary structure for TUFT1. However, none of the other cysteines in TUFT1 are conserved, which does not support such a model.

### **Alternative splicing**

Alternatively spliced TUFT1 mRNA transcripts have been detected in different tissues (Mao et al. 2001; Deutsch et al. 2002). An isoform lacking exon 2 was identified in the mouse kidney, whereas another lacking exon 2, part of exon 4, exon 5-11 and part of exon 12 was found in mouse liver. Various TUFT1 isoforms were also identified in tooth buds (lacking exon 2, or 3 or 6). Both the lack of exons 5-9 in some transcripts and the fact that they are variable could indicate their recent recruitment, i.e. in tetrapod ancestors for example. This hypothesis could also explain the presence of amino acids under positive selection in exons 6 and 7 (new constraints resulting from new selective pressures). It should be noted that these two alternatively spliced exons each have a two-amino acid motif under strong positive selection.

### **Relationships of TUFT1**

1 The single protein we found related to TUFT1 is MYZAP. Both proteins share some amino  
2 acid sequence similarities and lack a signal peptide, and their genes display similar exonic  
3 structures. These findings suggest that one might have derived from the other after  
4 duplication of an ancestral gene. Their probable presence in the genome of ancestral  
5 osteichthyans and their location on different chromosomes could indicate that their origin  
6 dates back to the genome duplication that occurred at the onset of vertebrate diversification,  
7 approximately 500 million years ago. The long history of these two proteins could explain  
8 why they do not share many similarities.  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21

### 22 **Additional remarks**

23  
24 Recent functional studies pointed to the involvement of TUFT1 in adaptation to hypoxia and  
25 in differentiation of neurons (Leiser et al. 2010, Deutsch et al. 2011). However, no interacting  
26 proteins or sequences within TUFT1 were indicated as being involved in these functions. Two  
27 putative HIF consensus DNA binding sites, compatible with hypoxia responsive elements  
28 (HRE, 5'-RCGTG-3') in the *TUFT1* promoter region, at positions -1296 and -27 of human  
29 *TUFT1* (upstream to exon 1; GenBank: AH009496.1), and at positions -130 and -8 of the  
30 mouse *TUFT1* promoter (GenBank: NC\_000069.5), were identified (Leiser et al. 2010).  
31  
32 These sequences, which were not included in the current analyses, point to the regulatory role  
33 of HIF1a on TUFT1 expression.  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45

46 TUFT1 is an acidic protein and although predictions of its secondary structure reveals  
47 two long coiled-coil regions, consurf and other prediction algorithms indicate that most  
48 residues are exposed, whereas a few buried residues could contribute to 3D structure. Hence,  
49 analysis of linear motifs was performed to reveal motifs predicting function. Unlike  
50 amelogenin, a search for linear motifs using the eukaryotic linear motif (ELM) server (Dinkel  
51 et al. 2016) revealed no motifs. The major contribution of our consurf analysis was to  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1 highlight four unknown regions under strong purifying selection and indicate functional and  
 2 buried residues, while previous studies can only point out (i) motifs indicated by yeast two-  
 3 hybrid systems (TIP39 and self-assembly), (ii) results of previous *in silico* analysis (Ca  
 4 binding domain), or (iii) post-translational modifications that might point to a function.  
 5  
 6  
 7  
 8  
 9

## 10 11 12 **References**

- 13  
 14 Al-Hashimi N, Sire JY, Delgado S (2009) Evolutionary Analysis of Mammalian Enamelin,  
 15 the Largest Enamel Protein, Supports a Crucial Role for the 32 kDa Peptide and Reveals  
 16 Selective Adaptation in Rodents and Primates. *J Mol Evol* 69(6):635-656.  
 17  
 18 Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating  
 19 evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic*  
 20 *Acids Res* 38 (Web Server issue):W529-533.  
 21  
 22 Bashir MM, Abrams WR, Tucker T, Sellinger B, Budarf M, Emanuel B, Rosenbloom J  
 23 (1998) Molecular cloning and characterization of the bovine and human tuftelin genes.  
 24 *Connect. Tissue Res* 39:13-24.  
 25  
 26 Celniker G, Nimrod G, Ashkenazy H, Glaser F, Martz E, Mayrose I, Pupko T, Ben-Tal N  
 27 (2013) ConSurf: Using evolutionary data to raise testable hypotheses about protein  
 28 function. *Isr J Chem* 53:199-206.  
 29  
 30 Chan HC, Mai L, Oikonomopoulou A, Chan HL, Richardson AS, Wang SK, Simmer JP, Hu  
 31 JC (2010) Altered enamel phosphorylation site causes amelogenesis imperfecta. *J Dent*  
 32 *Res* 89:695-699.  
 33  
 34 Deutsch D (1989) Structure and function of enamel gene products. *Anat Rec* 224:189-210.  
 35  
 36 Deutsch D, Palmon A, Fisher LW, Kolodny N, Termine JD, Young MF (1991) Sequencing of  
 37 bovine enamel (‘‘tuftelin’’) a novel acidic enamel protein. *J Biol Chem* 266:16021-16028.  
 38  
 39 Deutsch D, Palmon A, Young MF, Selig S, Kearns WG, Fisher LW (1994) Mapping of the  
 40 human tuftelin (TUFT1) gene to chromosome 1 by fluorescence in situ hybridization.  
 41 *Mamm. Genome* 5:461-462.  
 42  
 43 Deutsch D, Palmon A, Dafni L, Catalano-Sherman J, Young MF, Fisher LW (1995) The  
 44 enamel (tuftelin) gene. *Int Dev Biol* 39:135-143.  
 45  
 46 Deutsch D, Palmon A, Dafni L, Mao Z, Leytin V, Young M, Fisher LW (1998) Tuftelin –  
 47 aspects of protein and gene structure. *Eur J Oral Sci* 106 (Suppl. 1): 315–323.  
 48  
 49 Deutsch D, Shay B, Rosenfeld E, Leiser Y, Fermon E, Taylor A, Charuvi K, Cohen Y, Haze  
 50 A, Fuks A, Dafni L, Mao Z (2002) The human tuftelin gene and the expression of tuftelin  
 51 in mineralizing and nonmineralizing tissues. *Connect Tissue Res* 43:425-434.  
 52  
 53 Deutsch D, Silverstein N, Shilo D, Lecht S, Lazarovici P, Blumenfeld A (2011) Biphasic  
 54 influence of hypoxia on tuftelin expression in mouse mesenchymal C3H10T1/2 stem cells.  
 55 *Eur J Oral Sci* 119 (suppl.1):55-61.  
 56  
 57 Dinkel H, Van Roey K, Michael S, Kumar M, Uyar B, Altenberg B, Milchevskaya V,  
 58 Schneider M, Kühn H, Behrendt A, Dahl SL, Damerell V, Diebel S, Kalman S, Klein S,  
 59  
 60  
 61  
 62  
 63  
 64  
 65

1 Knudsen AC, Mäder C, Merrill S, Staudt A, Thiel V, Welti L, Davey NE, Diella F, Gibson  
2 TJ (2016) ELM 2016-data update and new functionality of the eukaryotic linear motif  
3 resource. *Nucleic Acids Res* 44(D1):D294-300.

4 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high  
5 throughput. *Nucleic Acids Res* 32(5):1792-1797.

6  
7 Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection  
8 may operate. *Mol Biol Evol* 13:685-690.

9  
10 Jeremias FL, Koruyucu M, Kuchler EC, Bayram M, Tuna EB, Deeley K, Pierri RA, Souza JF,  
11 Fragelli CM, Paschoal MA, Gencay K, Seymen F, Caminaga RM, dos Santos-Pinto L,  
12 Vieira AR (2013) Genes expressed in dental enamel development are associated with  
13 molar-incisor hypomineralization. *Arch Oral Biol* 58:1434-1442.

14  
15 Kosakovsky Pond SL, Frost SDW (2005) A genetic algorithm approach to detecting lineage-  
16 specific variation in selection pressure. *Mol Biol Evol* 22:478-485.

17  
18 Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using  
19 phylogenies. *Bioinformatics* 21:676-679.

20  
21 Leiser Y, Blumenfeld A, Haze A, Dafni L, Taylor AL, Rosenfeld E, Fermon E, Gruenbaum-  
22 Cohen Y, Shay B, Deutsch D (2007) Localization, quantification, and characterization of  
23 tuftelin in soft tissues. *Anat Rec* 290:449-454.

24  
25 Leiser Y, Silverstein NC, Blumenfeld A, Shilo D, Haze A, Rosenfeld E, Shay B, Tabakman  
26 R, Lecht S, Lazarovici P, Deutsch D (2010) The induction of tuftelin expression in PC12  
27 cell line during hypoxia and NGF induced differentiation. *J. Cell Physiol* 226:165-172.

28  
29 MacDougall M, Simmons D, Dodds A, Knight C, Luan X, Zeichner-David M, Zhang C, Ryu  
30 OH, Qian Q, Simmer JP, Hu C-C (1998) Cloning, characterization, and tissue expression  
31 pattern of mouse tuftelin cDNA. *J Dent Res* 77:1970-1978.

32  
33 Mao Z, Shay B, Hekmati M, Fermon E, Taylor A, Dafni L, Heikenheimo K, Lustmann J,  
34 Fisher LW, Young MF, Deutsch D (2001) The human tuftelin gene: cloning and  
35 characterization. *Gene* 279:181-196.

36  
37 Meredith RW1, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik  
38 E, Simão TL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C,  
39 Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS,  
40 Murphy WJ (2011) Impacts of the Cretaceous Terrestrial Revolution and KPg extinction  
41 on mammal diversification. *Science* 334(6055):521-524.

42  
43 Paine CT, Paine ML, Luo W, Okamoto CT, Lyngstadaas SP, Snead ML (2000) A tuftelin-  
44 interacting protein (TIP39) localizes to the apical secretory pole of mouse ameloblasts. *J*  
45 *Biol Chem* 275:22284-22292.

46  
47 Patir A, Seymen F, Yildirim M, Deeley K, Cooper ME, Marazita ML, Vieira AR (2008)  
48 Enamel formation genes are associated with high caries experience in Turkish children.  
49 *Caries Res* 42:394-400.

50  
51 Pieperhoff S, Rickelt S, Heid H, Claycomb WC, Zimbelmann R, Kuhn C, Winter-  
52 Simanowski S, Kuhn C, Frey N, Franke WW (2012) The plaque protein myozap identified  
53 as a novel major component of adhering junctions in endothelia of the blood and the lymph  
54 vascular systems. *J Cell Mol Med* 16:1709-1719.

55  
56 Rickelt S, Kuhn C, Winter-Simanowski S, Zimbelmann R, Frey N, Franke WW (2011)  
57 Protein myozap--a late addition to the molecular ensembles of various kinds of adherens  
58  
59  
60  
61  
62  
63  
64  
65

junctions. *Cell Tissue Res* 346:347-359.

- 1  
2 Schmid K, Yang Z (2008) The trouble with sliding windows and the selective pressure in  
3 BRCA1. *PLoS ONE* 3:e3746
- 4  
5 Sergei L. Kosakovsky Pond SL, Spencer V, Muse SV (2005) HyPhy: Hypothesis testing  
6 using Phylogenies. In *Statistical Methods for Molecular Evolution. Statistics for Biology*  
7 *and Health, Part II*, pp. 125-181.
- 8  
9 Shay B, Gruenbaum-Cohen Y, Tucker AS, Taylor AL, Rosenfeld E, Haze A, Dafni L, Leiser  
10 Y, Fermon E, Danieli T, Blumenfeld A, Deutsch D (2009) High yield expression of  
11 biologically active recombinant full length human tuftelin protein in baculovirus-infected  
12 insect cells. *Protein Expr Purif* 68:90-98.
- 13  
14 Silvent J, Sire JY, Delgado S (2013) The Dentin Matrix Acidic Phosphoprotein 1 (DMP1) in  
15 the light of mammalian evolution. *J Mol Evol* 76(1-2):59-70.
- 16  
17 Stanek, D, Pridalova-Hnilicova J, Novotny I, Huranova M, Blazikova M, Wen X, Sapra AK,  
18 Neugebauer KM (2008) Spliceosomal small nuclear ribonucleoprotein particles repeatedly  
19 cycle through Cajal bodies. *Mol Biol Cell* 19:2534-2543.
- 20  
21 Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T (2007) Selecton 2007:  
22 advanced models for detecting positive and purifying selection using a Bayesian inference  
23 approach. *Nucleic Acids Res* 35(Web Server issue):W506-511.
- 24  
25 Sigrist CJA, Cerutti L, de Castro E, Langendijk-GenevauxPS, Bulliard V, Bairoch A, Hulo N  
26 (2010) PROSITE, a protein domain database for functional characterization and  
27 annotation. *Nucleic Acids Res* 38(Database issue):D161-D166.
- 28  
29 Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5:  
30 Molecular evolutionary genetics analysis using maximum likelihood, evolutionary  
31 distance, and maximum parsimony methods. *Mol Biol Evol* 28:2731-2739.
- 32  
33 Tsunoyama K, Gojobori T (1998) Evolution of nicotinic acetylcholine receptor subunits. *Mol*  
34 *Biol Evol* 15:518-527.
- 35  
36 Wen X, Lei Y-P, Zhou YL, Okamoto CT, Snead ML, Paine ML (2005) Structural  
37 organization and cellular localization of tuftelin-interacting protein 11 (TFIP11). *Cell Mol*  
38 *Life Sci* 62:1038-1046.
- 39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

## Figure Captions

**Figure 1.** Evolutionary analysis of the mammalian TUFT1 sequences using SLAC (A) and the Sliding Window (B) analysis.

A. Non-synonymous substitution rate (dN) along TUFT1 sequences.

B. Logarithm of substitution rate per site along the TUFT1 sequences estimated for a window of 15 bp with an overlap of 5 bp between each pair of windows.

The regions with the lowest rate of non-synonymous substitution (A) and the lower Ln likelihood (B) indicate strong constraints, which reflect high selective pressures.

**Figure 2.** Distance tree obtained from alignment of the 82 mammalian *TUFT1* sequences.

The longer the branches are, the higher the evolutionary rate of the taxa. Scale bar = number of substitutions per site.

**Figure 3.** Evolutionary chart of TUFT1 calculated from the 82 mammalian sequences. This

chart is a simplification of Consurf analysis (SM4), and was deduced from the results

obtained when dN/dS was calculated at each codon of TUFT1 by Conseq (purifying

selection) and by Selecton (positive selection). The human sequence was used as reference.

Positions subjected to purifying selection are marked as black (conserved positions) and grey

(conservative positions) background. Positively selected positions are indicated with an

asterisk. The two frames represent putative functional domains predicted in previous studies

(see references in text).

**Figure 4.** Schematic representation of the human TUFT1 sequence, on which are indicated

the domains and remarkable positions identified in previous studies and present work (after

Deutsch et al., 2002). The regions and amino acid positions that were conserved during 200

1 million years of mammalian evolution are indicated in green. They are predicted to play an  
2 important role either for the function or the structure of the protein. In contrast, the regions  
3 and positions that are considered variable are indicated in red. It is worth noting that all  
4 cysteines and most positions predicted by *Prosite* as being functional were not validated by  
5 our analyses.  
6  
7  
8  
9  
10

11  
12  
13  
14 **Supplementary material 1 (SM1).** Scientific names, common names, families, orders and  
15 references in GenBank for the 82 *TUFT1* sequences used in our study. Published sequences in  
16 bold. Alphabetical order of Latin names. Amino acid sequences are available in SM2. XM\_  
17 and NM\_ sequences were obtained from *NCBI* database. ENS sequences from *Ensembl*  
18 release 80 and 81. Other sequences were obtained in GenBank using blast on genomes being  
19 sequenced [<https://www.ncbi.nlm.nih.gov/Traces/wgs/>].  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30

31 **Supplementary material 2 (SM2).** Amino acid sequences of the 82 mammalian TUFT1  
32 proteins included in the study. Number of amino acids between brackets. ? = unknown  
33 residues.  
34  
35  
36  
37  
38  
39  
40

41 **Supplementary material 3 (SM3).** The 82 mammalian TUFT1 sequences were aligned  
42 against the human sequence and were ordered following mammalian relationships. (|= exon  
43 limits; (.)= residue identical to human TUFT1 residue; (-)= indel; (?)= unknown amino acid;  
44 (\*)= stop codon. See SM2 for amino acid sequences.  
45  
46  
47  
48  
49  
50  
51  
52

53 **Supplementary material 4 (SM4).** Results of the ConSurf analysis of the TUFT1 sequences  
54 from 82 mammalian species, revealing a panorama of the various selections acting on  
55 TUFT1.  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65



1  
2 **Supplementary material 5 (SM5).** Prediction of post translational modifications of the  
3  
4 human TUFT1 sequence using *Prosite* database.  
5  
6

7  
8  
9 **Supplementary material 6 (SM6).** Statistical tests supporting Consurf analysis.  
10  
11

12  
13  
14 **Supplementary material 7 (SM7).** Non-exhaustive list of remarkable amino acid  
15  
16 positions of TUFT1 highlighted in our study.  
17  
18

19  
20  
21 **Supplementary material 8 (SM8).** Results of the Conseq analysis of TUFT1 sequences from  
22  
23 82 mammalian species.  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

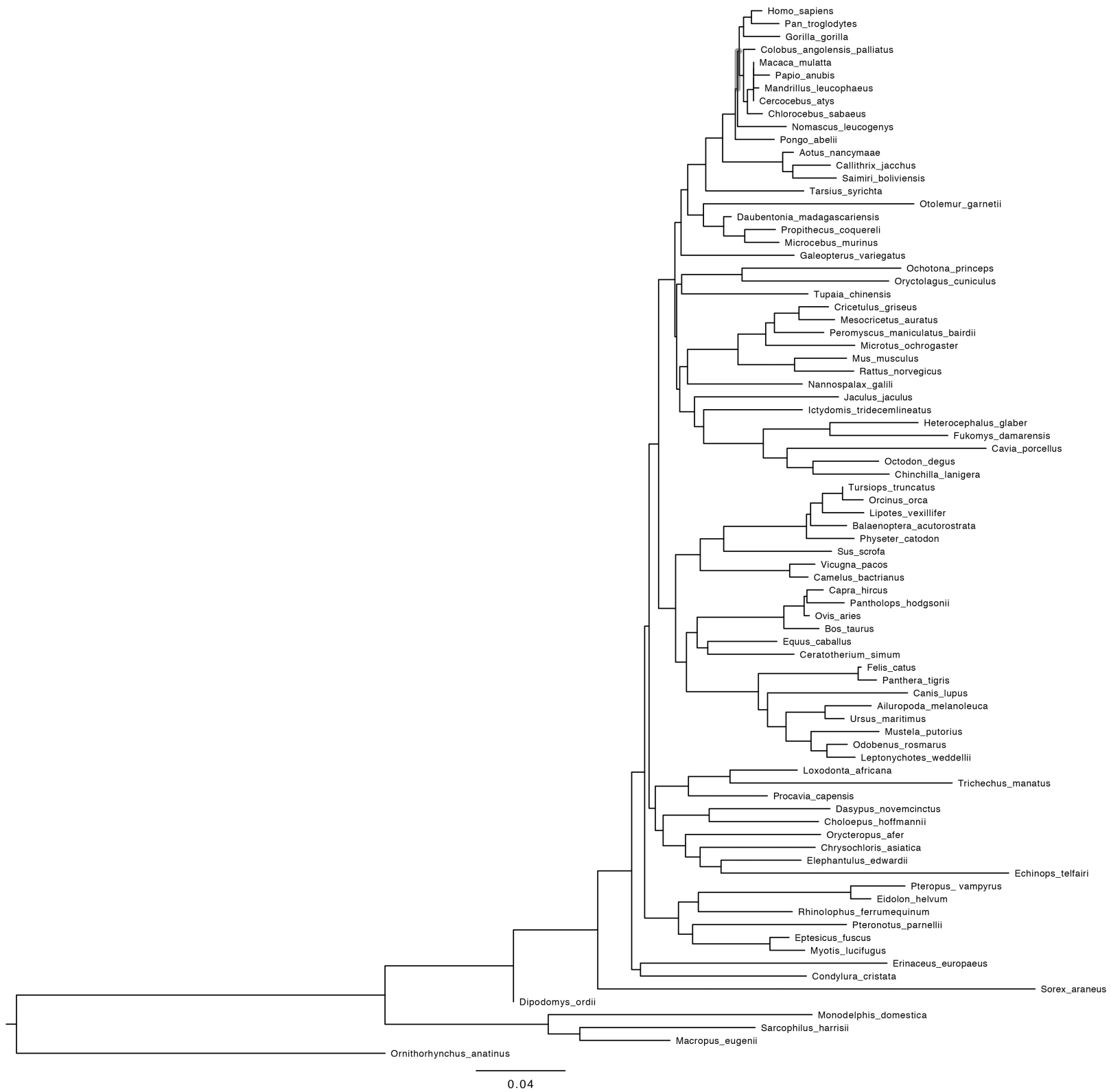


Figure 2. Delgado et al.



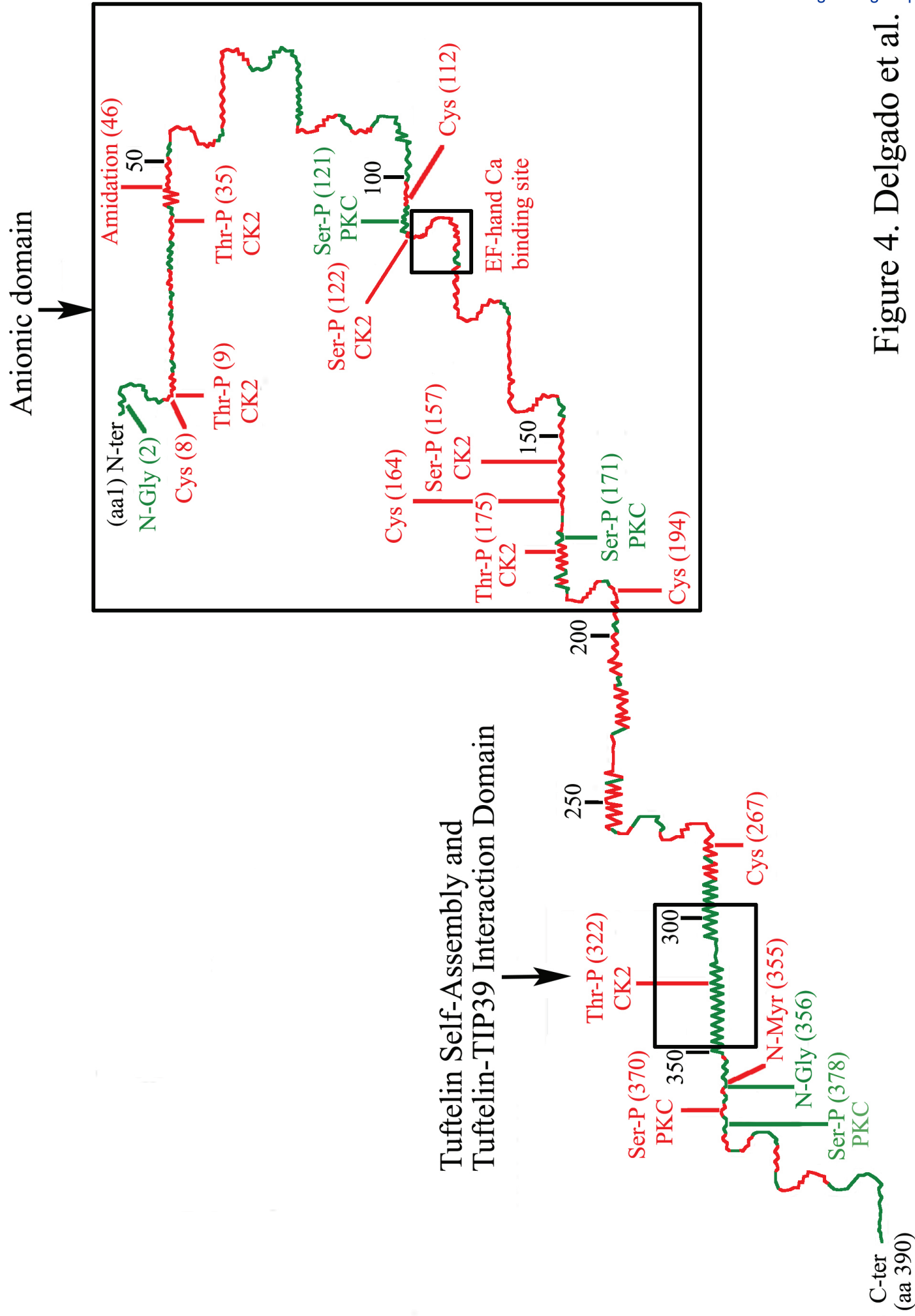


Figure 4. Delgado et al.

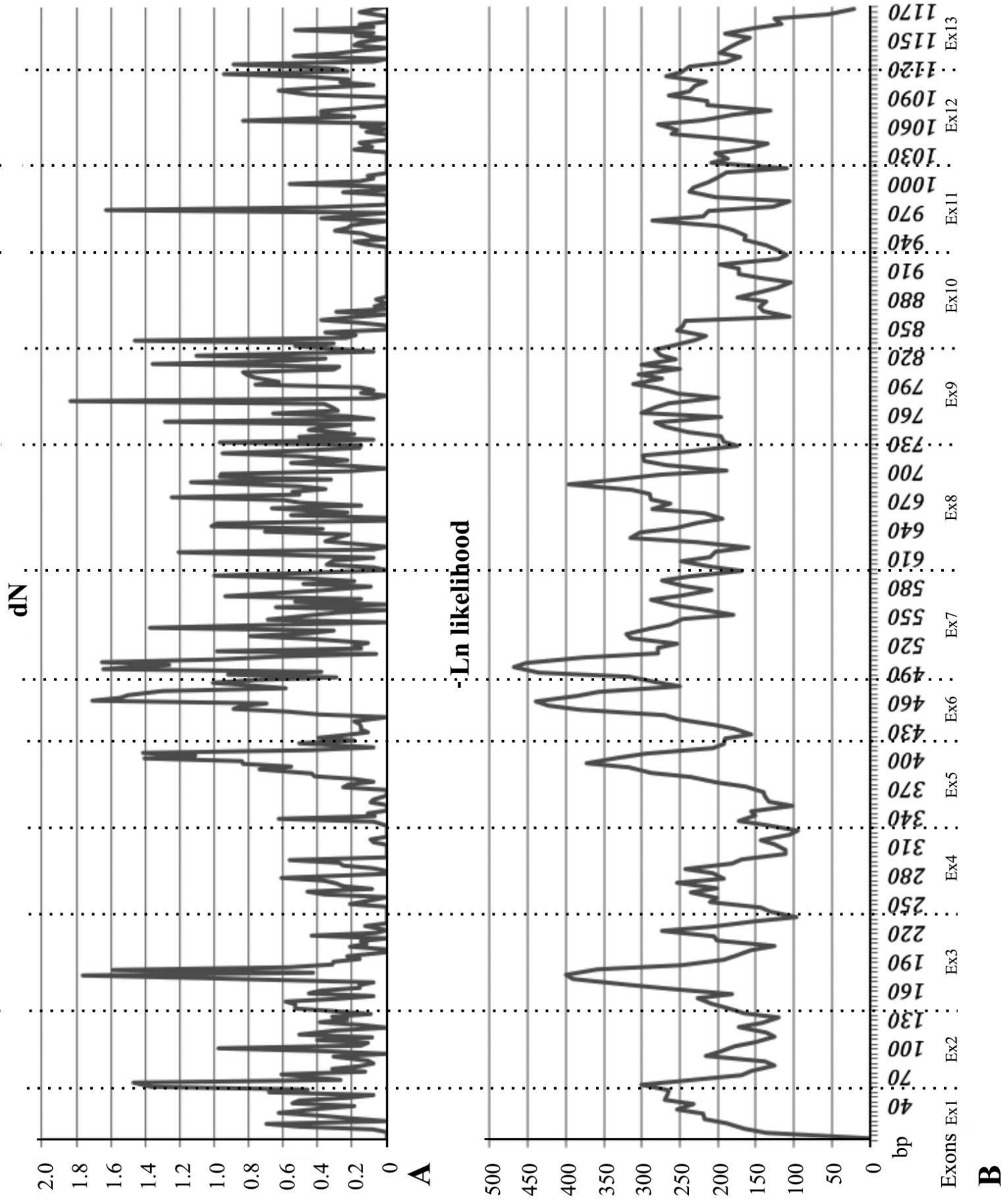


Figure 1. Delgado et al.

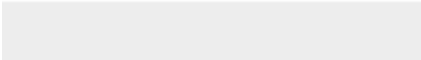




Click here to access/download  
**Supplementary Material**  
SM9.pdf



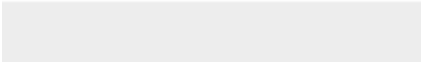


Click here to access/download  
**Supplementary Material**  
SM2.pdf





Click here to access/download  
**Supplementary Material**  
SM3.pdf







Click here to access/download  
**Supplementary Material**  
SM4.jpg






Click here to access/download  
**Supplementary Material**  
SM5.pdf





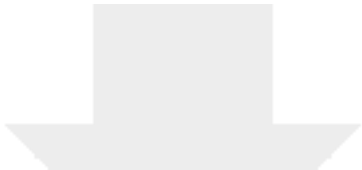
Click here to access/download  
**Supplementary Material**  
SM1.pdf





Click here to access/download  
**Supplementary Material**  
SM6.pdf





Click here to access/download  
**Supplementary Material**  
SM7.pdf





Click here to access/download  
**Supplementary Material**  
SM8.pdf

