What's in a Likelihood? Simple models of protein evolution and the contribution of structurally viable reconstructions.

SUPPLEMENTARY MATERIALS

Clemens Lakner[1,2], Mark T. Holder[3], Nick Goldman[4], and Gavin J. P. Naylor[2]

[1] *Department of Biological Science, Section: Ecology and Evolution, Florida State University, Tallahassee, Florida 32306-4120, USA*
[2] *Department of Scientific Computing, Florida State University, Tallahassee, Florida 32306-4120, USA*
[3] *Department of Ecology and Evolution, 6031 Haworth, University of Kansas, 1200 Sunnyside Ave, Lawrence KS 66045*
[4] *European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK*

Contents

Reference Data Sets

Sequences used for the transition paths are indicated in boldface font.

1. HPPK (GenBank Accession.version)
   BAB96719.1, AAP15678.1, AAL19147.1, **AAO67923.1 (*Salmonella typhi*)**, CAE13168.1, AAS60560.1, **CAG76218.1 (*Pectobacterium atrosepticum*)**, BAC95526.1, **AAF93760.1 (*Vibrio cholerae*)**, CAG21480.1, **YP_001140682.1 (*Aeromonas salmonicida*)**;

2. Lysozyme (UniProtKB)
   LYZ1_MOUSE, **LYSC_HUMAN (*Homo sapiens*)**, LYSC_GORGO, LYSC_PONPY, LYSC_HYLLA, LYZ2_MOUSE, LYSC_NASLA, LYSCK_SHEEP, LYSC_SAISC, LYSC_COLAN, LYSC_RABIT, LYSC_TRAOB, LYSC_SEMEN, LYSC_PYGNE, LYSC1_RAT, LYSC_MACMU, LYSC_PAPAN, LYSC_TRISI, LYSC2_CANFA, LYSC_SAGOE, LYSC_ALLNI, LYSC_CALJA, LYSC_CERAE, LYSM_BOVIN, LYSC_MIOTA, LYSC2_RAT, LYSCN_BOVIN, LYSC3_PIG, LYSC_AMYCA, **LYSC_BUFAN (*Bufo andrewsi*)**, LYSC_PHOVI, LYSC_LEPWE, **LYSC_CAMDR (*Camelus dromedarius*)**, **LYSC_CHEMY (*Chelonia mydas*)**;

3. Myoglobin (UniProtKB)
   MYG_CARCR, MYG_CHEMY, MYG_GRAGE, MYG_VARVA, MYG_URILO, MYG_ANAPO, MYG_PHAFI, MYG_CHICK, MYG_CERMN, MYG_AETPY, **MYG_STRCA (*Struthio camelus*)**, **MYG_ALLMI (*Alligator mississippiensis*)**, MYG_PERPO, MYG_OCHPR, MYG_LEPMU, MYG_LUTLU, MYG_NYCCO, MYG_TUPGL, MYG_SAISC, MYG_CEBAP, MYG_PROGU, MYG_OCHCU, MYG_PIG, MYG_CALJA, MYG_AOTTR, MYG_ROUAE, MYG_GORBE, MYG_RABIT, MYG_CTEGU, MYG_LAGLA, MYG_LAGMA, MYG_MACFA, MYG_ERYPA, MYG_GALCR, MYG_DIDMA, MYG_HYLAG, MYG_HUMAN, MYG_ZALCA, MYG_PONPY, MYG_PHOVI, MYG_PANTR, **MYG_EQUBU (*Equus burchellii*)**, MYG_ORYAF, MYG_VULCH, MYG_ORNAN, MYG_CANFA, MYG_LYCPI, **MYG_RAT (*Rattus norvegicus*)**, MYG_SPAEH, MYG_ERIEU, MYG_TACAC, MYG_INIGE, MYG_MACRU, MYG_MOUSE, MYG_MELME, MYG_INDPC, MYG_ZIPCA, MYG_MEGNO, MYG_MESST, MYG_BALBO, MYG2_STEAT, MYG_BALAC, MYG_GLOME, MYG1_STEAT, MYG_BALPH, MYG_ORCOR, MYG_PENEL, MYG_SHEEP, MYG_PHOSI, MYG_BOVIN, MYG_PHODA, MYG_PHYCA, MYG_KOGBR, MYG_CEREL, MYG_BUBBU, MYG_LOXAF, MYG_ELEMA;

4. Parvalbumin (UniProtKB)
   PRVA_RAT (*Rattus norvegicus*), PRVA_MOUSE, PRVA_GERSP, PRVA_MACFU, PRVA_HUMAN, PRVA_BOVIN, PRVA_RABIT, PRVA_FELCA, PRVM_CHICK, PRVA_RANCA, PRVA_RANES, **PRVA_AMPME (*Amphiuma means*)**, **PRVA_LATCH (*Latimeria chalumnae*)**, **PRVA_TRISE (*Triakis semifasciata*)**;

Trees

Two taxa

HPPK

LG: (Salmonella:0.72945,Vibrio:0.0)

WAG: (Salmonella:0.66055,Vibrio:0.0);

Poisson: (Salmonella:0.60856,Vibrio:0.0);

Lysozyme

LG: (LYSC_HUMAN:0.54236,LYSC_CHEMY:0.0);

WAG: (LYSC_HUMAN:0.51281,LYSC_CHEMY:0.0);

Poisson: (LYSC_HUMAN:0.48017,LYSC_CHEMY:0.0);

Myoglobin

LG: (MYG_EQUBU:0.49951,MYG_ALLMI:0.0);

WAG: (MYG_EQUBU:0.46228,MYG_ALLMI:0.0);

Poisson: (MYG_EQUBU:0.46225,MYG_ALLMI:0.0);

Parvalbumin

LG: (PRVA_RAT:0.58710,PRVA_TRISE:0.0);

WAG: (PRVA_RAT:0.54033,PRVA_TRISE:0.0);

Poisson: (PRVA_RAT:0.54166,PRVA_TRISE:0.0);

Three taxa

HPPK

LG: (Aeromonas:0.367204,Salmonella:0.514048,Vibrio:0.228230);

WAG: (Aeromonas:0.337242,Salmonella:0.464291,Vibrio:0.218742);

Poisson: (Aeromonas:0.267048,Salmonella:0.412442,Vibrio:0.224312);

Lysozyme

LG: (LYSC_HUMAN:0.357404,LYS_BUFAN:0.340339,LYSC_CHEMY:0.203214);

WAG: (LYSC_HUMAN:0.345938,LYS_BUFAN:0.319364,LYSC_CHEMY:0.188813);

Poisson: (LYSC_HUMAN:0.313610,LYS_BUFAN:0.300220,LYSC_CHEMY:0.181252);

Myoglobin

LG: (MYG_EQUBU:0.229045,MYG_STRCA:0.100012,MYG_ALLMI:0.264132);

WAG: (MYG_EQUBU:0.219730,MYG_STRCA:0.093948,MYG_ALLMI:0.245053);

Poisson: (MYG_EQUBU:0.211233,MYG_STRCA:0.102493,MYG_ALLMI:0.248147);

Parvalbumin

LG: (PRVA_RAT:0.189325,PRVA_AMPME:0.316503,PRVA_TRISE:0.395588);

WAG: (PRVA_RAT:0.187683,PRVA_AMPME:0.286560,PRVA_TRISE:0.361415);

Poisson: (PRVA_RAT:0.229427,PRVA_AMPME:0.247373,PRVA_TRISE:0.330310);

Four taxa

HPPK

1. LG
   ((Vibrio:0.327906,Salmonella:0.400892):0.010265,Pectobacterium:0.302588,Aeromonas:0.445498);
   ((Vibrio:0.330949,Pectobacterium:0.307360):0.000882,Salmonella:0.403020,Aeromonas:0.448683);
   ((Vibrio:0.230261,Aeromonas:0.363793):0.220977,Salmonella:0.338829,Pectobacterium:0.219824);

2. WAG
   ((Vibrio:0.313270,Salmonella:0.368740):0.005007,Pectobacterium:0.282739,Aeromonas:0.407168);
   ((Vibrio:0.314001,Pectobacterium:0.283188):0.008442,Salmonella:0.364117,Aeromonas:0.403835);
   ((Vibrio:0.222122,Aeromonas:0.327729):0.206134,Salmonella:0.314046,Pectobacterium:0.204833);

3. Poisson
   ((Vibrio:0.284759,Salmonella:0.325622):0.024539,Pectobacterium:0.252609,Aeromonas:0.332781);
   ((Vibrio:0.293604,Pectobacterium:0.262784):0.003472,Salmonella:0.333768,Aeromonas:0.340022);
   ((Vibrio:0.216645,Aeromonas:0.270011):0.189444,Salmonella:0.273903,Pectobacterium:0.187624);

Lysozyme

1. LG
   ((LYSC_HUMAN:0.226501,LYS_BUFAN:0.421583):0.000004,LYSC_CAMDR:0.250204,LYSC_CHEMY:0.328016);
   ((LYSC_HUMAN:0.162835,LYSC_CAMDR:0.200649):0.219274,LYS_BUFAN:0.325866,LYSC_CHEMY:0.213459);
   ((LYSC_HUMAN:0.226500,LYSC_CHEMY:0.328016):0.000004,LYS_BUFAN:0.421583,LYSC_CAMDR:0.250204);

2. WAG
   ((LYSC_HUMAN:0.220983,LYS_BUFAN:0.395049):0.000004,LYSC_CAMDR:0.238081,LYSC_CHEMY:0.314577);
   ((LYSC_HUMAN:0.164085,LYSC_CAMDR:0.195008):0.208546,LYS_BUFAN:0.303897,LYSC_CHEMY:0.203699);
   ((LYSC_HUMAN:0.220982,LYSC_CHEMY:0.314578):0.000004,LYS_BUFAN:0.395049,LYSC_CAMDR:0.238080);

3. Poisson
   ((LYSC_HUMAN:0.196094,LYS_BUFAN:0.375805):0.000004,LYSC_CAMDR:0.228509,LYSC_CHEMY:0.293868);
   ((LYSC_HUMAN:0.147103,LYSC_CAMDR:0.188600):0.189729,LYS_BUFAN:0.289089,LYSC_CHEMY:0.194504);
   ((LYSC_HUMAN:0.196094,LYSC_CHEMY:0.293868):0.000004,LYS_BUFAN:0.375805,LYSC_CAMDR:0.228508);

Myoglobin

1. LG
   ((MYG_ALLMI:0.346734,MYG_EQUBU:0.137471):0.007576,MYG_STRCA:0.201163,MYG_RAT:0.154511);
   ((MYG_ALLMI:0.261681,MYG_STRCA:0.102564):0.161860,MYG_EQUBU:0.097480,MYG_RAT:0.123639);
   ((MYG_ALLMI:0.336742,MYG_RAT:0.141416):0.021991,MYG_EQUBU:0.131997,MYG_STRCA:0.201699);

2. WAG
((MYG_ALLMI:0.311230,MYG_EQUBU:0.142668):0.004823,MYG_STRCA:0.180253,MYG_RAT:0.157309);
((MYG_ALLMI:0.243448,MYG_STRCA:0.095948):0.152621,MYG_EQUBU:0.098243,MYG_RAT:0.113698);
((MYG_ALLMI:0.301248,MYG_RAT:0.142007):0.021703,MYG_EQUBU:0.135894,MYG_STRCA:0.180156);

3. Poisson
((MYG_ALLMI:0.336832,MYG_EQUBU:0.124078):0.007050,MYG_STRCA:0.198378,MYG_RAT:0.144539);
((MYG_ALLMI:0.250671,MYG_STRCA:0.097718):0.148530,MYG_EQUBU:0.091079,MYG_RAT:0.120739);
((MYG_ALLMI:0.320826,MYG_RAT:0.120287):0.033088,MYG_EQUBU:0.115644,MYG_STRCA:0.194298);

Parvalbumin

1. LG
((PRVA_RAT:0.179552,PRVA_AMPME:0.326468):0.176410,PRVA_LATCH:0.292100,PRVA_TRISE:0.267793);
((PRVA_RAT:0.231982,PRVA_LATCH:0.348190):0.032714,PRVA_AMPME:0.343972,PRVA_TRISE:0.317235);
((PRVA_RAT:0.244997,PRVA_TRISE:0.339128):0.000004,PRVA_LATCH:0.360526,PRVA_AMPME:0.364043);

2. WAG
((PRVA_RAT:0.179067,PRVA_AMPME:0.292863):0.153355,PRVA_LATCH:0.272352,PRVA_TRISE:0.250651);
((PRVA_RAT:0.222339,PRVA_LATCH:0.323013):0.027584,PRVA_AMPME:0.312089,PRVA_TRISE:0.299048);
((PRVA_RAT:0.232508,PRVA_TRISE:0.315793):0.000004,PRVA_LATCH:0.332321,PRVA_AMPME:0.326638);

3. Poisson
((PRVA_RAT:0.199973,PRVA_AMPME:0.266704):0.141846,PRVA_LATCH:0.268246,PRVA_TRISE:0.230462);
((PRVA_RAT:0.231054,PRVA_LATCH:0.311383):0.038143,PRVA_AMPME:0.287142,PRVA_TRISE:0.274608);
((PRVA_RAT:0.237759,PRVA_TRISE:0.284504):0.018877,PRVA_LATCH:0.313572,PRVA_AMPME:0.293818);

Likelihoods (lnL)

Two taxa

HPPK

LG: -735.662167
WAG: -741.261095
Poisson: -791.086604

Lysozyme

LG: -601.541213
WAG: -594.482931
Poisson: -619.852608

Myoglobin

LG: -665.949741
WAG: -668.857905
Poisson: -723.725963

Parvalbumin

LG: -482.837648
WAG: -478.894401
Poisson: -532.923162

Three taxa

HPPK

LG: -972.537795
WAG: -977.544549
Poisson: -1042.765084

Lysozyme

LG: -787.024360
WAG: -776.966605
Poisson: -826.511054

Myoglobin

LG: -794.556943
WAG: -796.066717
Poisson: -872.272315

Parvalbumin

LG: -629.383650
WAG: -625.072729
Poisson: -697.673948

Four taxa

HPPK

Tree #	LG	WAG	Poisson
1	-1197.547	-1201.159	-1285.722
2	-1197.597	-1201.081	-1286.574
3	-1180.046*	-1182.973*	-1262.192*

Lysozyme

Tree #	LG	WAG	Poisson
1	-955.261	-942.007	-1006.594
2	-930.941*	-919.064*	-986.027*
3	-955.260	-942.007	-1006.594

Myoglobin

Tree #	LG	WAG	Poisson
1	-937.633	-941.064	-1045.844
2	-911.131*	-913.278*	-1009.918*
3	-936.682	-939.841	-1041.132

Parvalbumin

Tree #	LG	WAG	Poisson
1	-775.110*	-768.997*	-866.085*
2	-783.536	-776.476	-874.209
3	-784.903	-777.338	-875.069

Plots

Two taxa

a) Z_s (see paper for PE_d and PE_s)

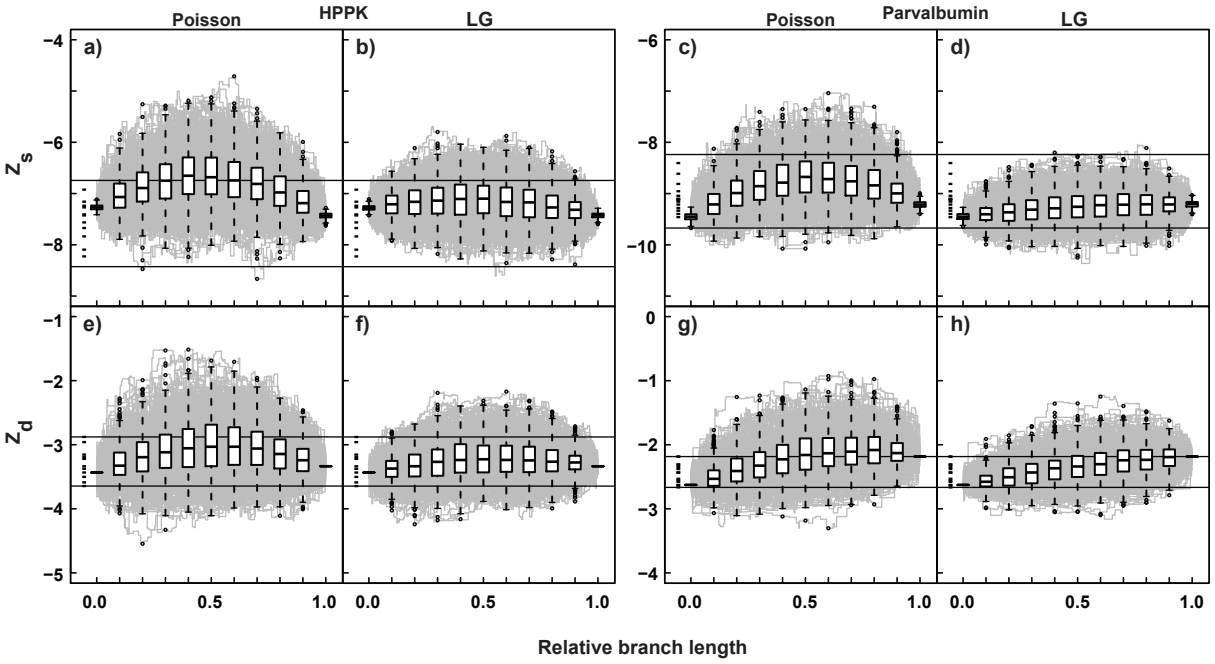

Figure 1. Constrained mappings (Z_s based on $10,000$ shuffles).

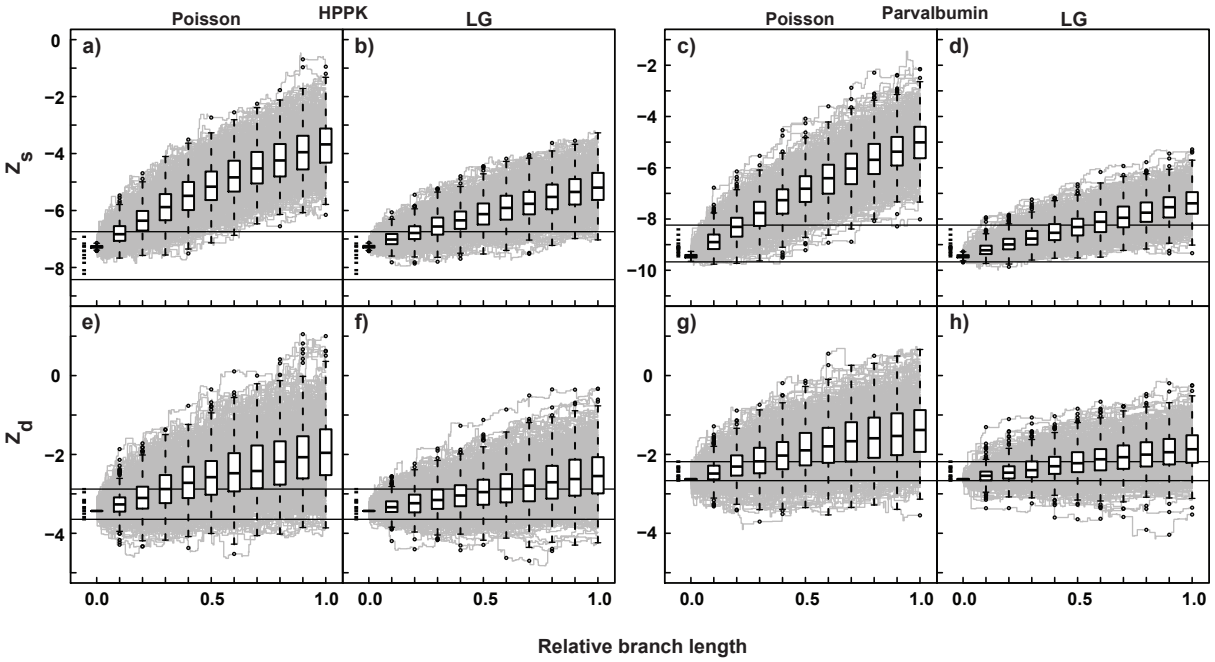

Figure 2. Unconstrained mappings (Z_s based on $10,000$ shuffles).

In the remaining plots the PE_s and Z_s scores are based on $1,000$ shuffles. Results presented in the paper are based on $10,000$ shuffles which greatly reduced the variance. The median PW_s and Z_s scores were very similar for $1,000$ and $10,000$ sequence shuffles.

Three taxa

a) PE_s

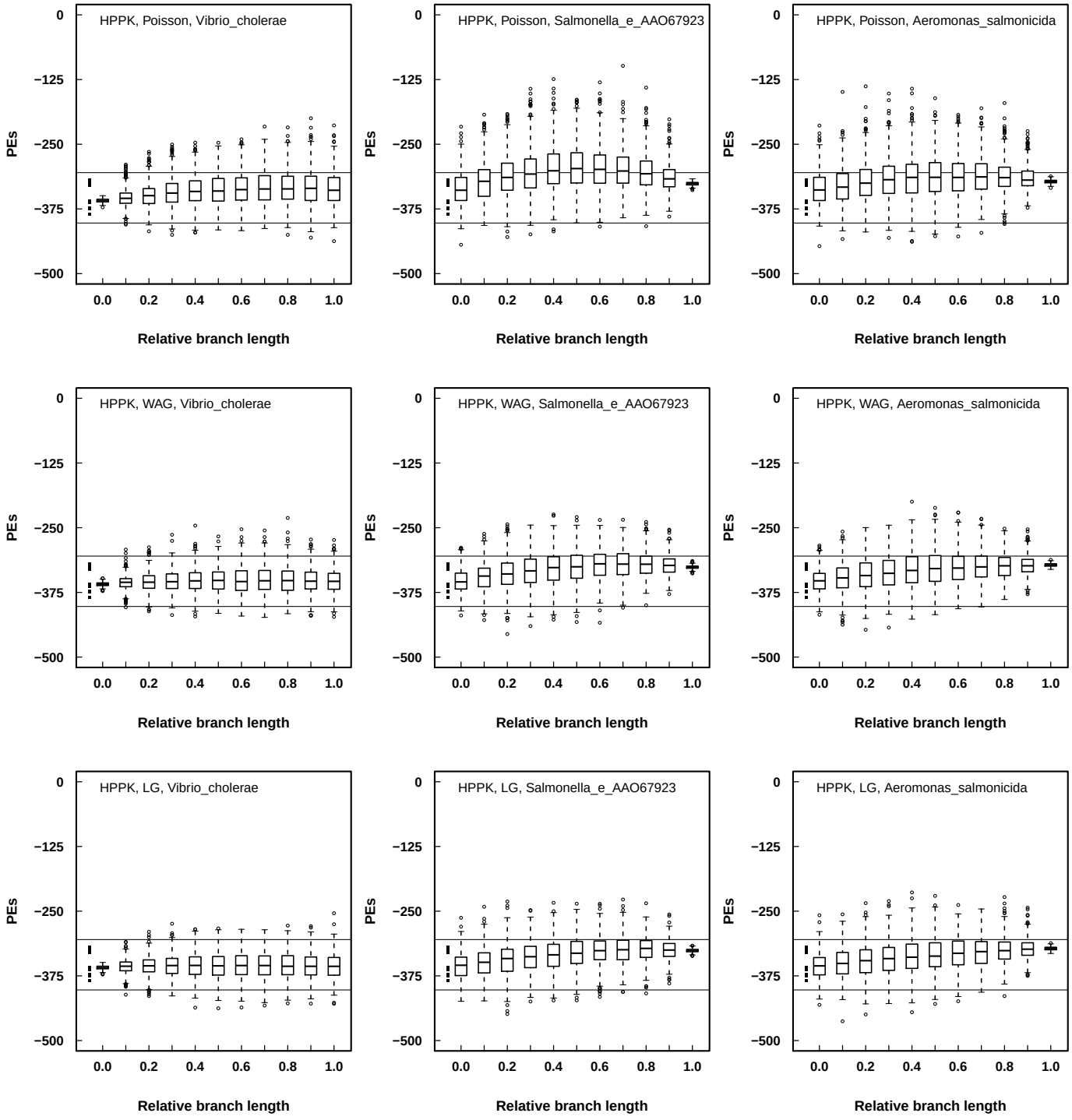

Figure 3. HPPK, PE_s

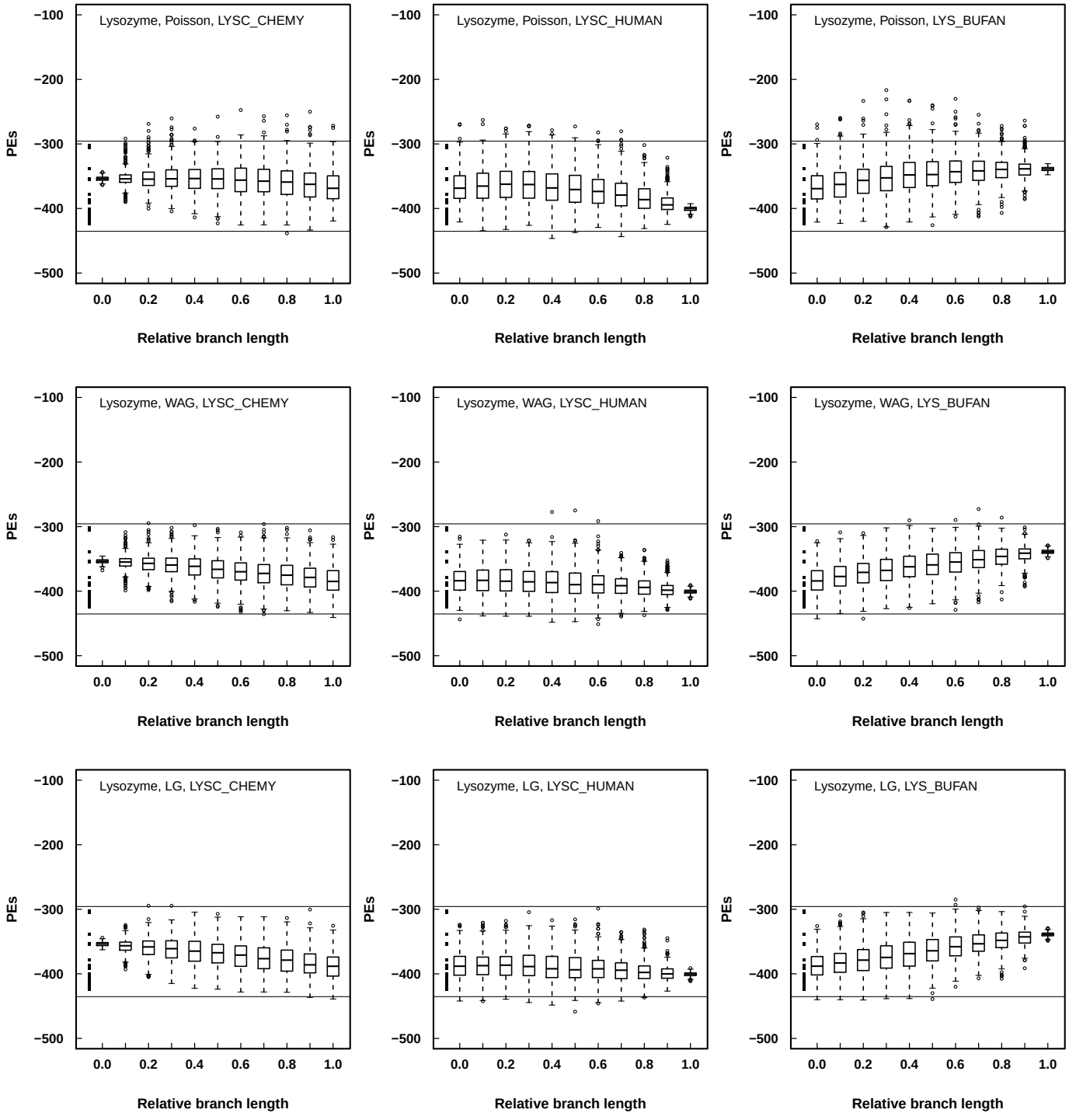

Figure 4. Lysozyme, PE_s

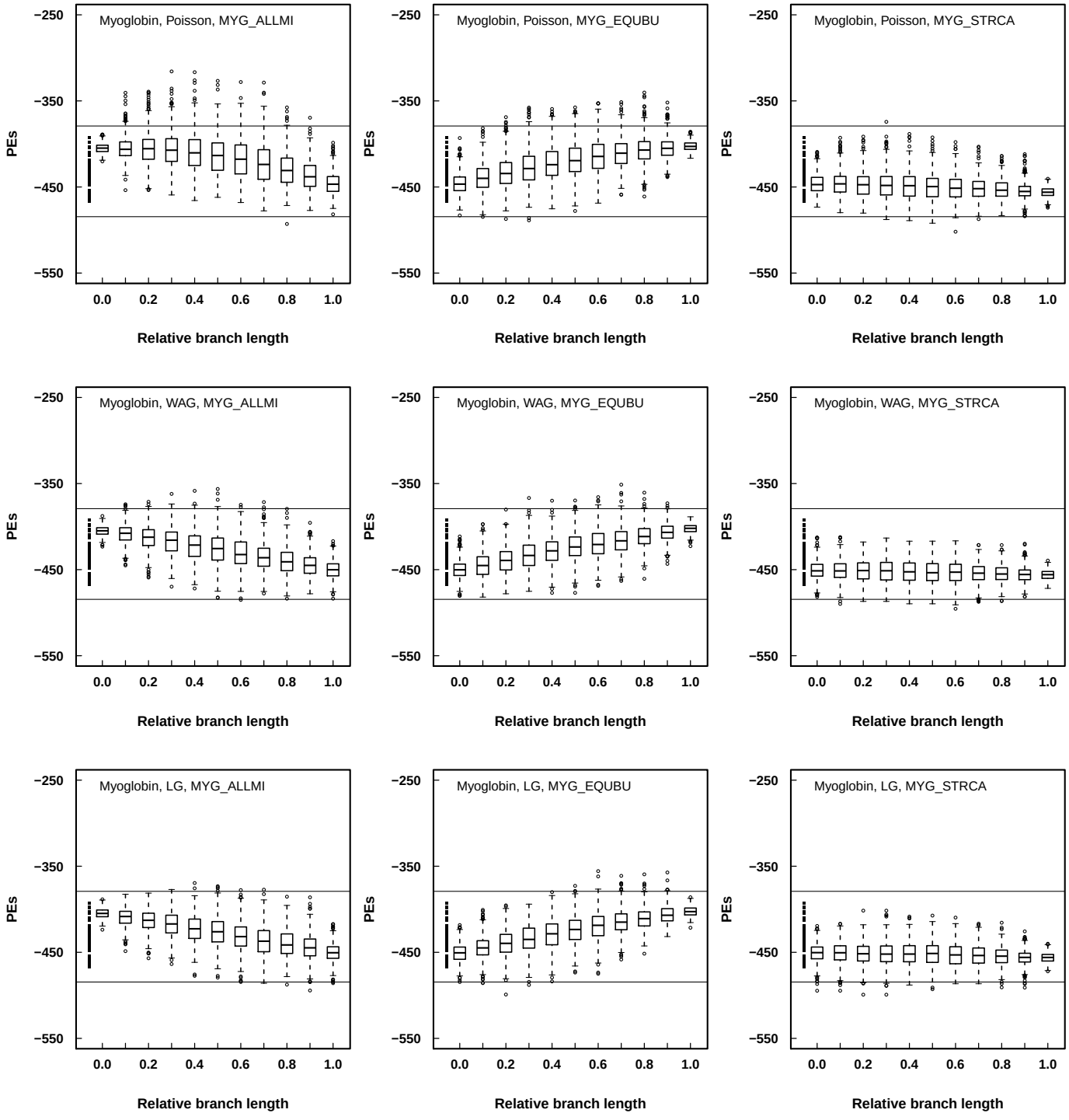

Figure 5. Myoglobin, PE_s

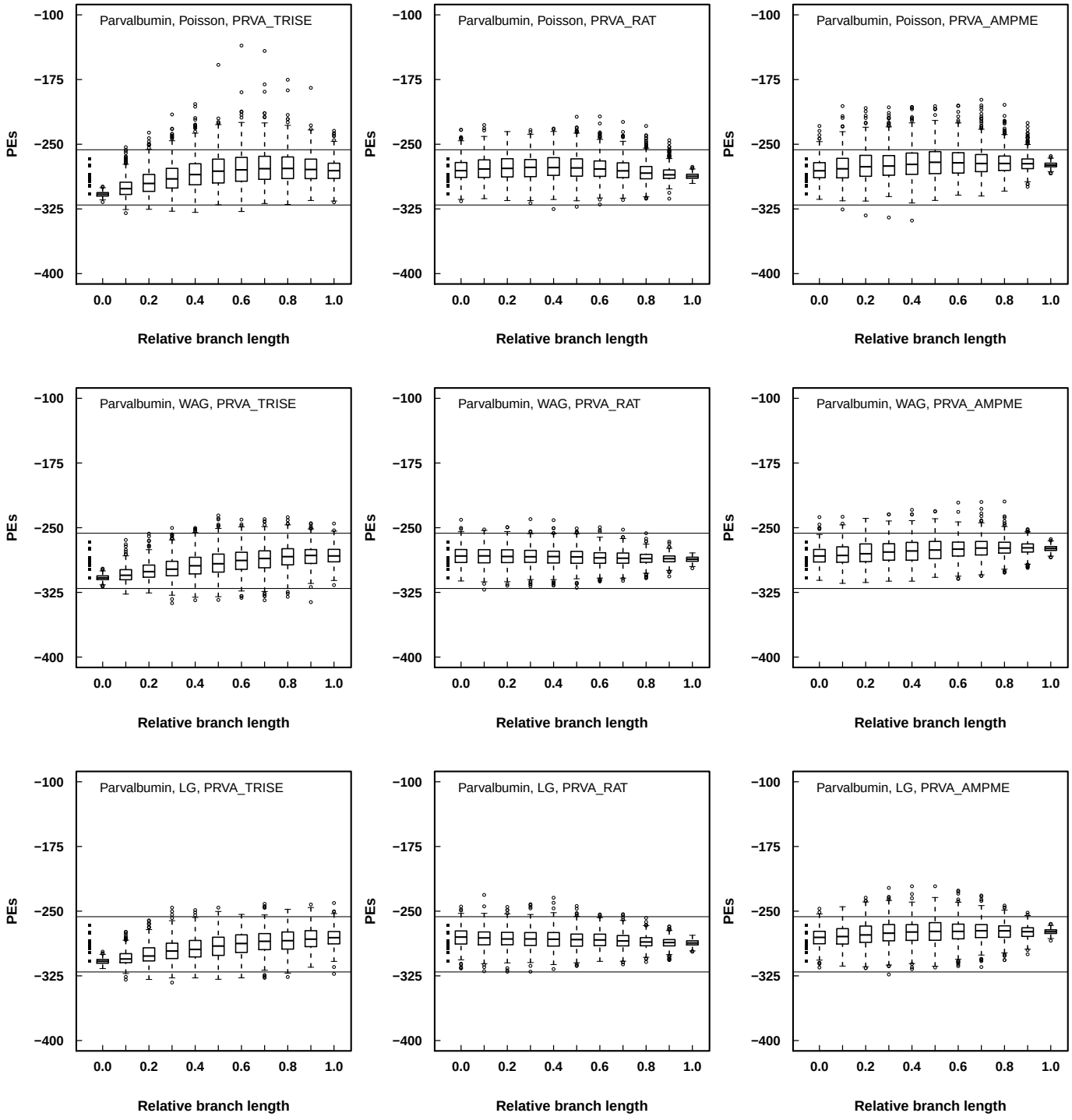

Figure 6. Parvalbumin, PE_s

b) Z_s

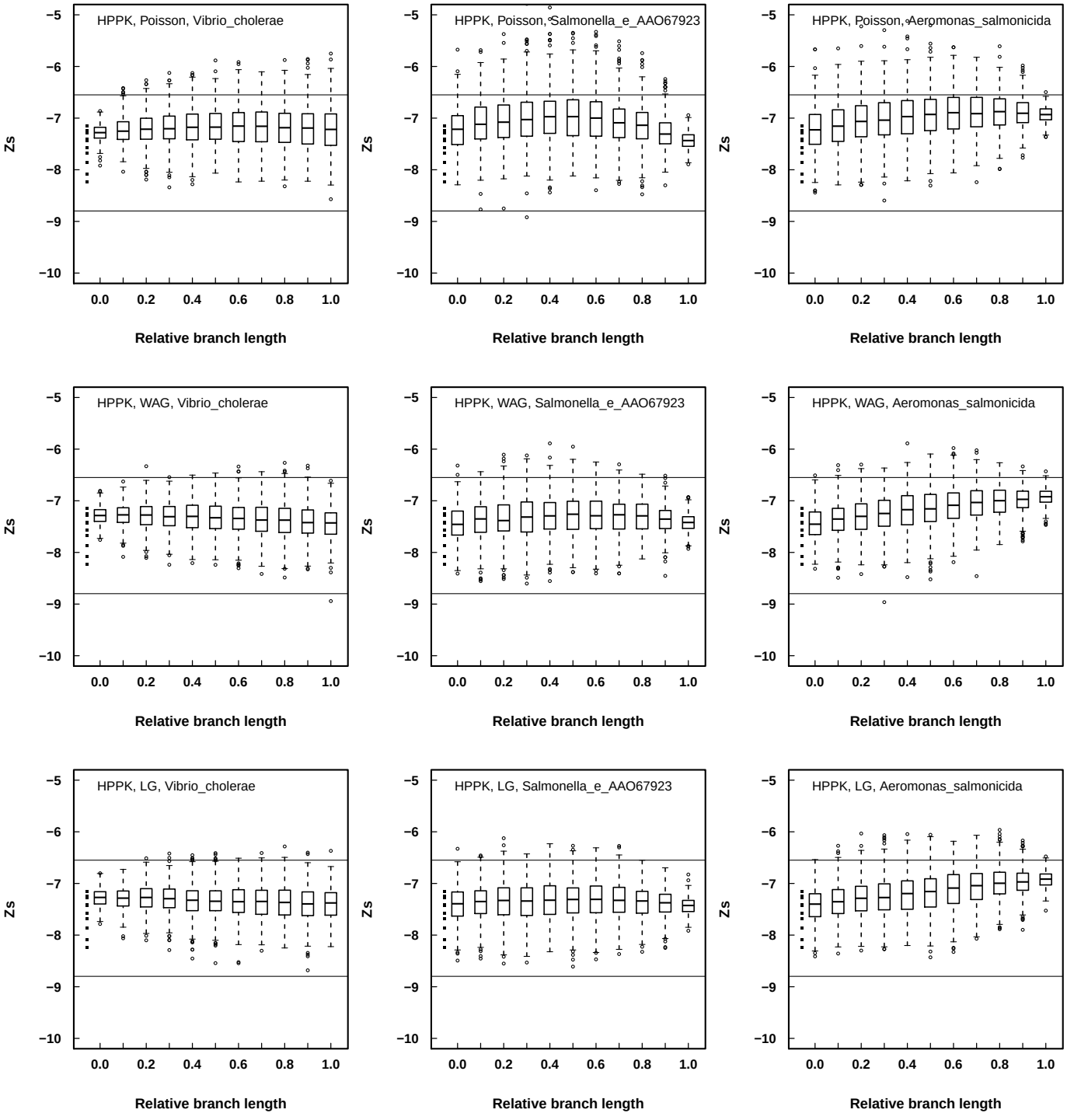

Figure 7. HPPK, Z_s

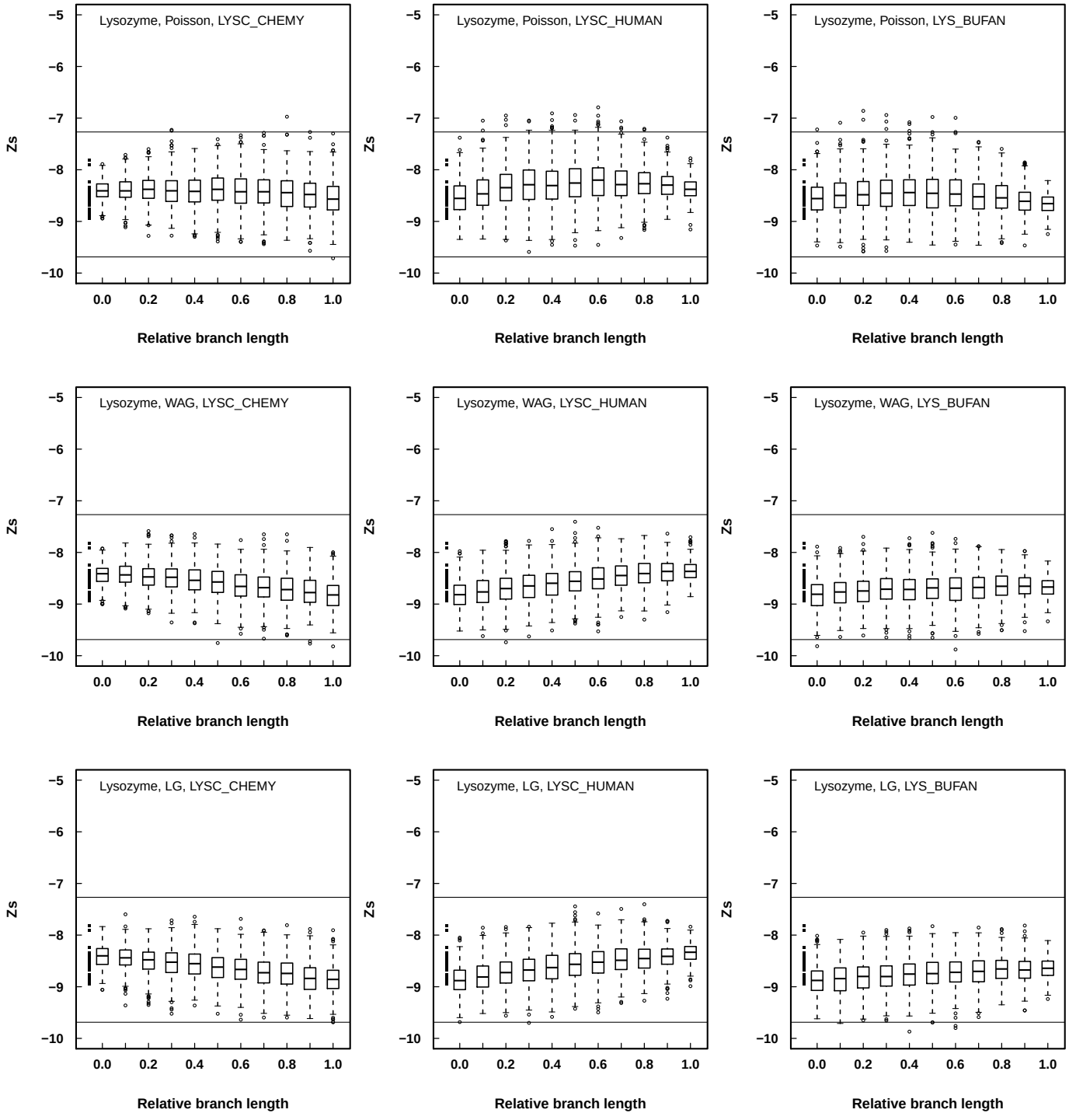

Figure 8. Lysozyme, Z_s

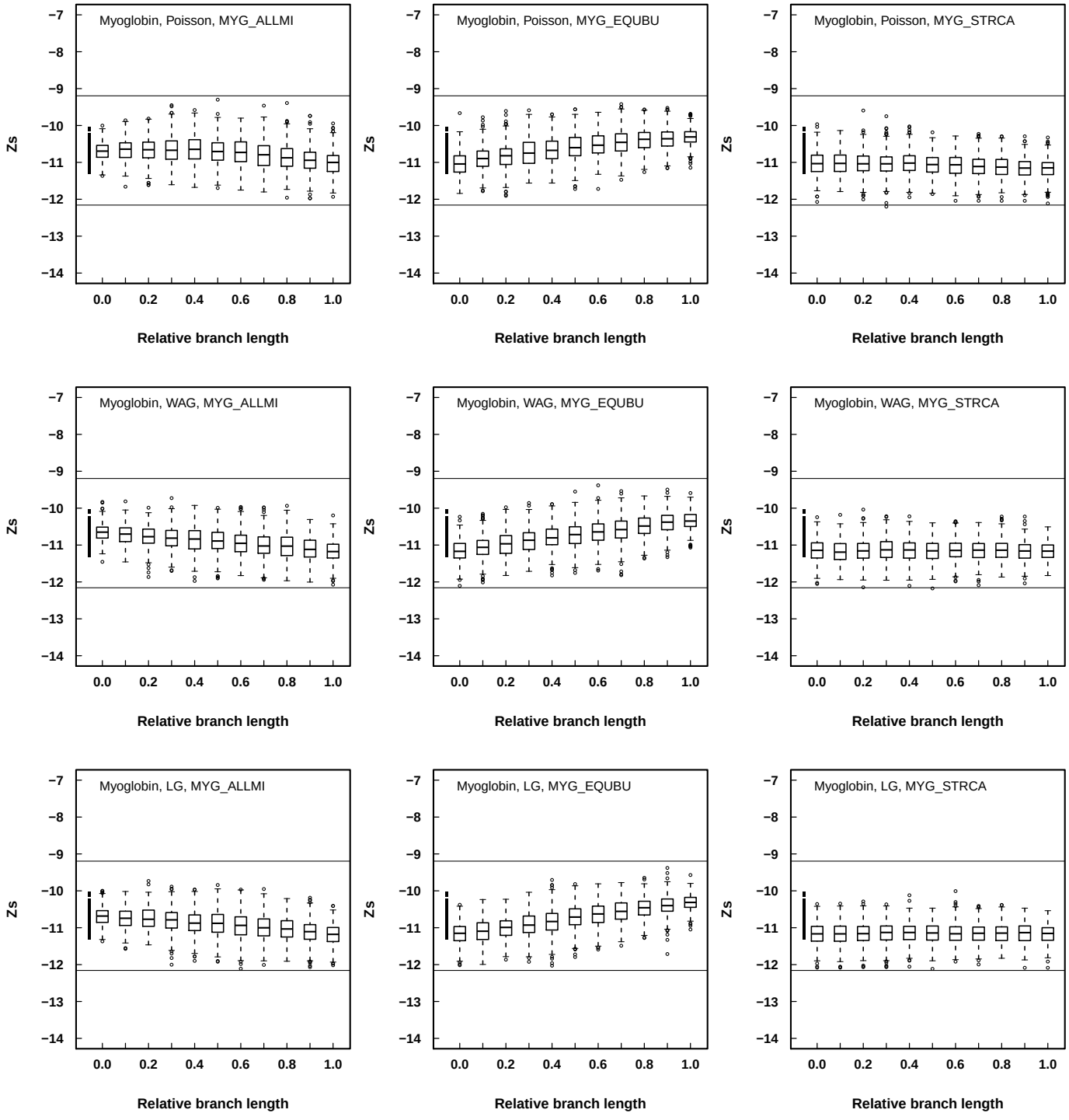

Figure 9. Myoglobin, Z_s

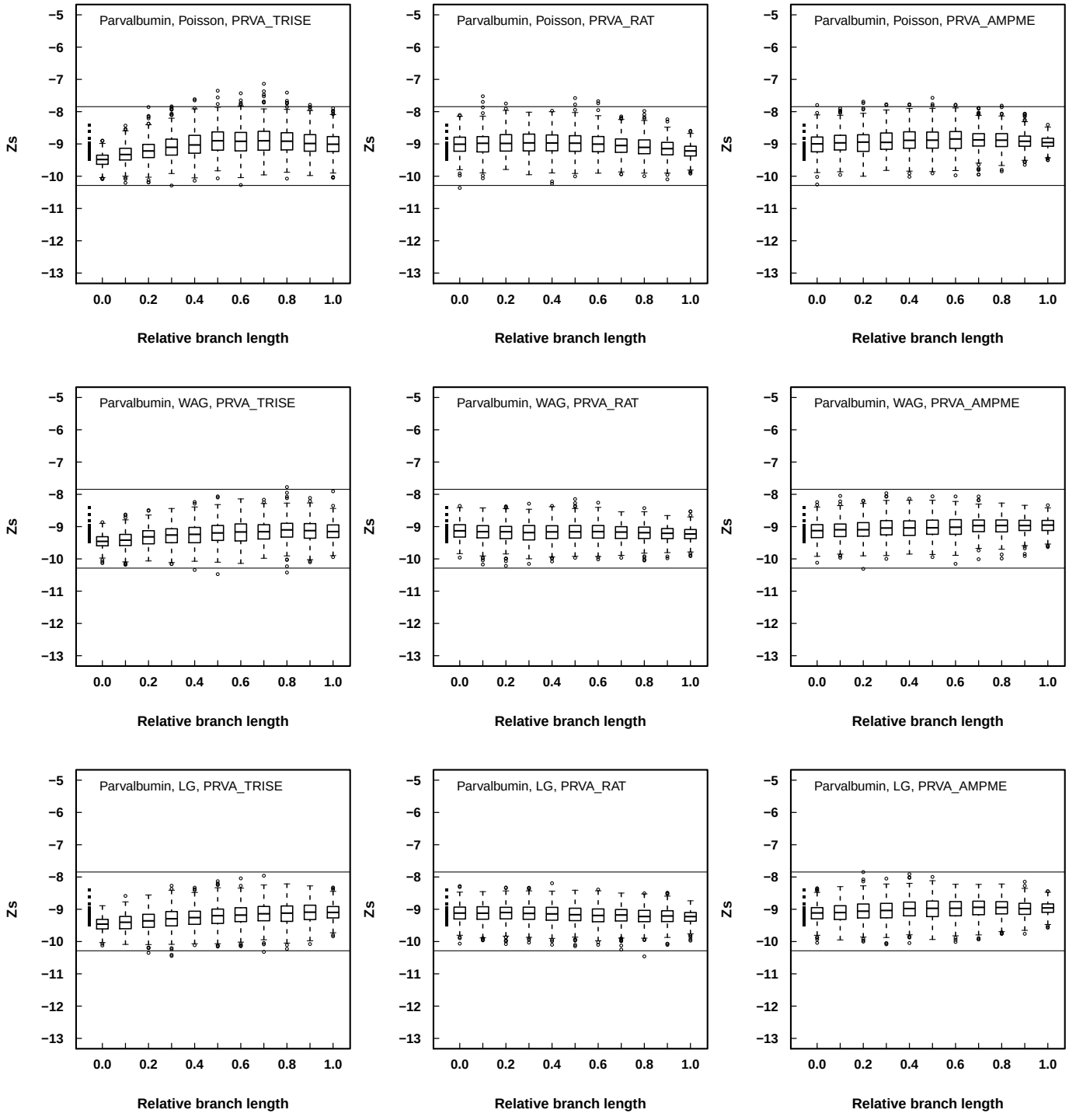

Figure 10. Parvalbumin, Z_s

Four taxa

a) PE_s

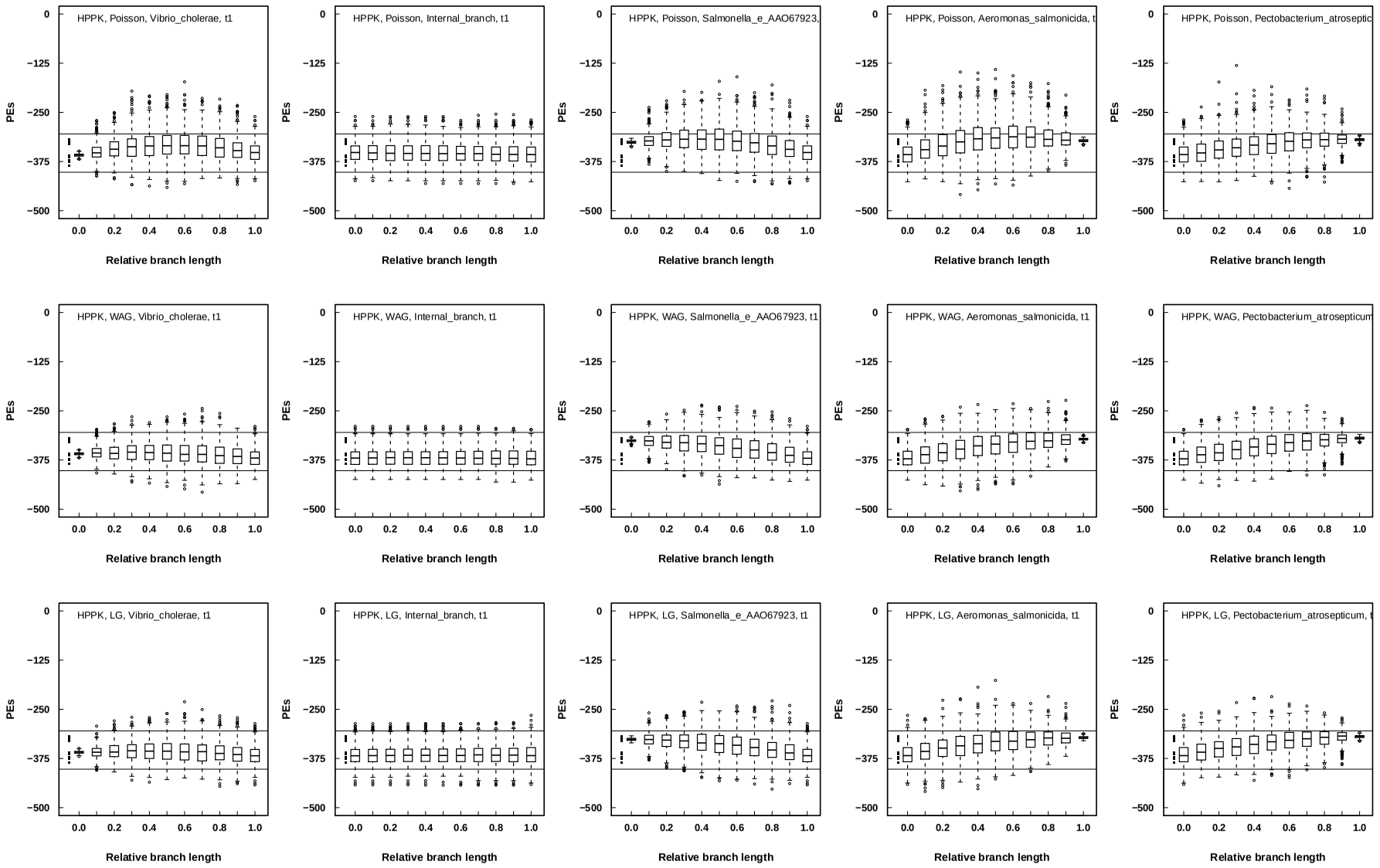

Figure 11. HPPK, PE_s, Tree 1

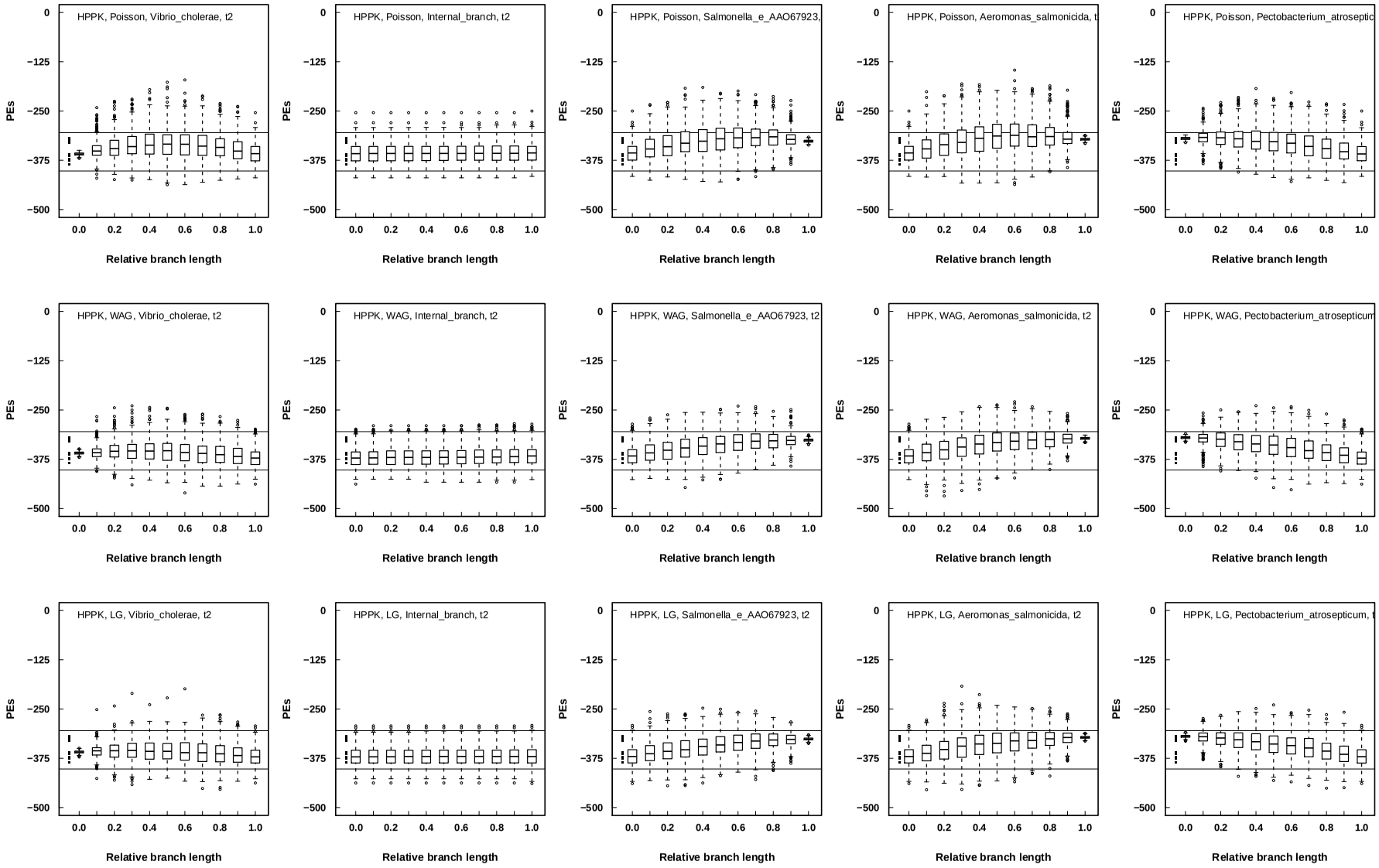

Figure 12. HPPK, PE_s, Tree 2

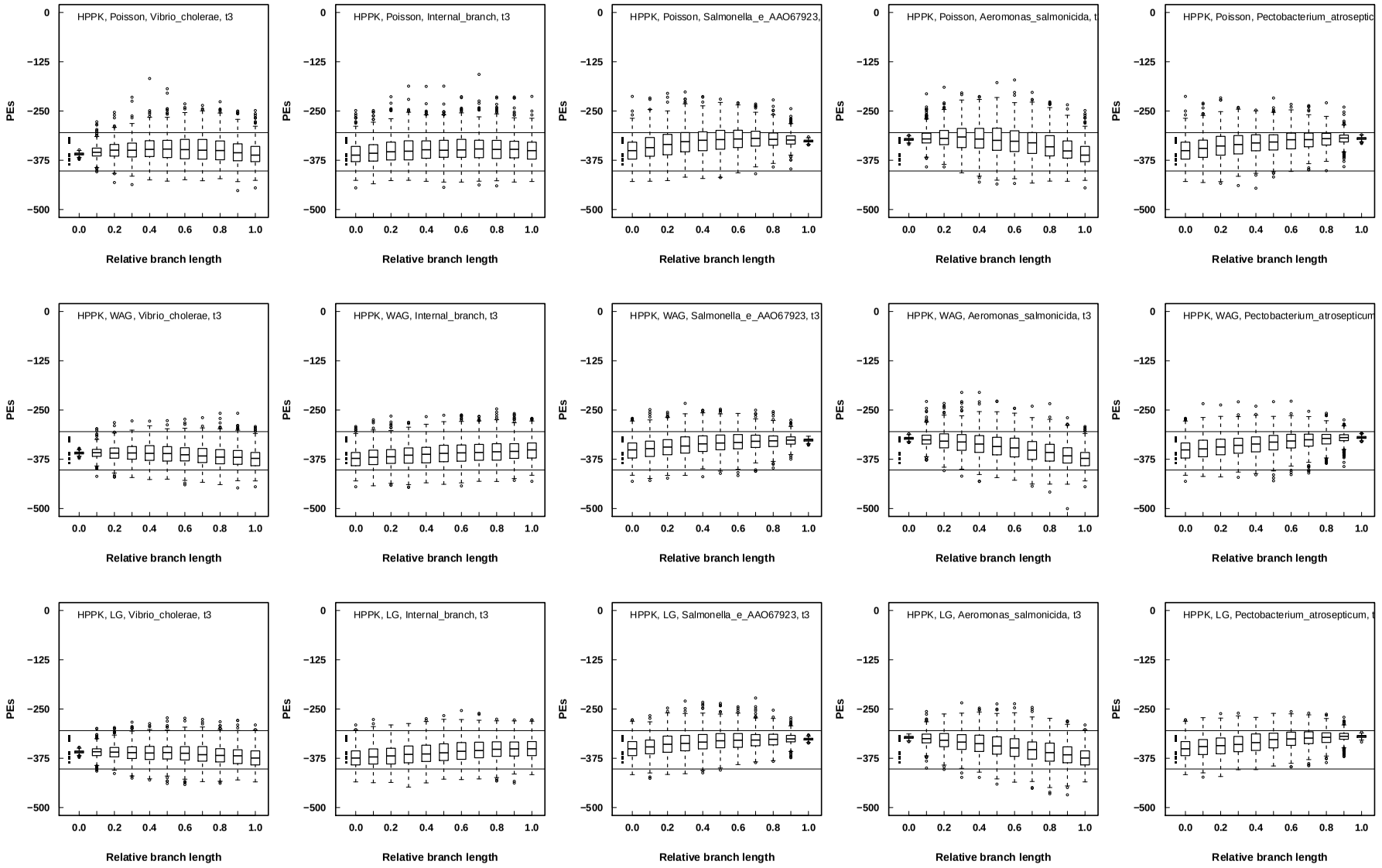

Figure 13. HPPK, PE_s, Tree 3

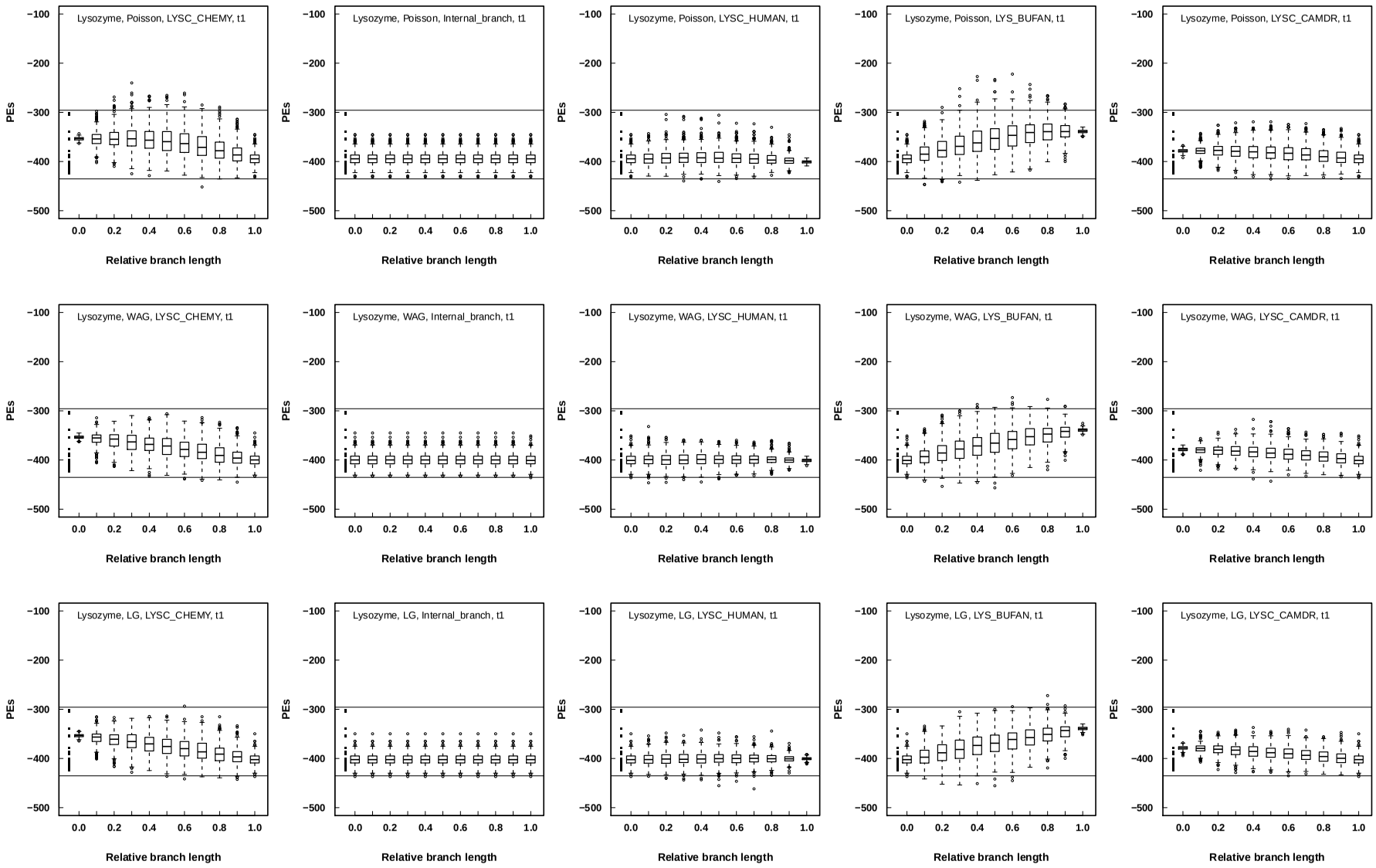

Figure 14. Lysozyme, PE_s, Tree 1

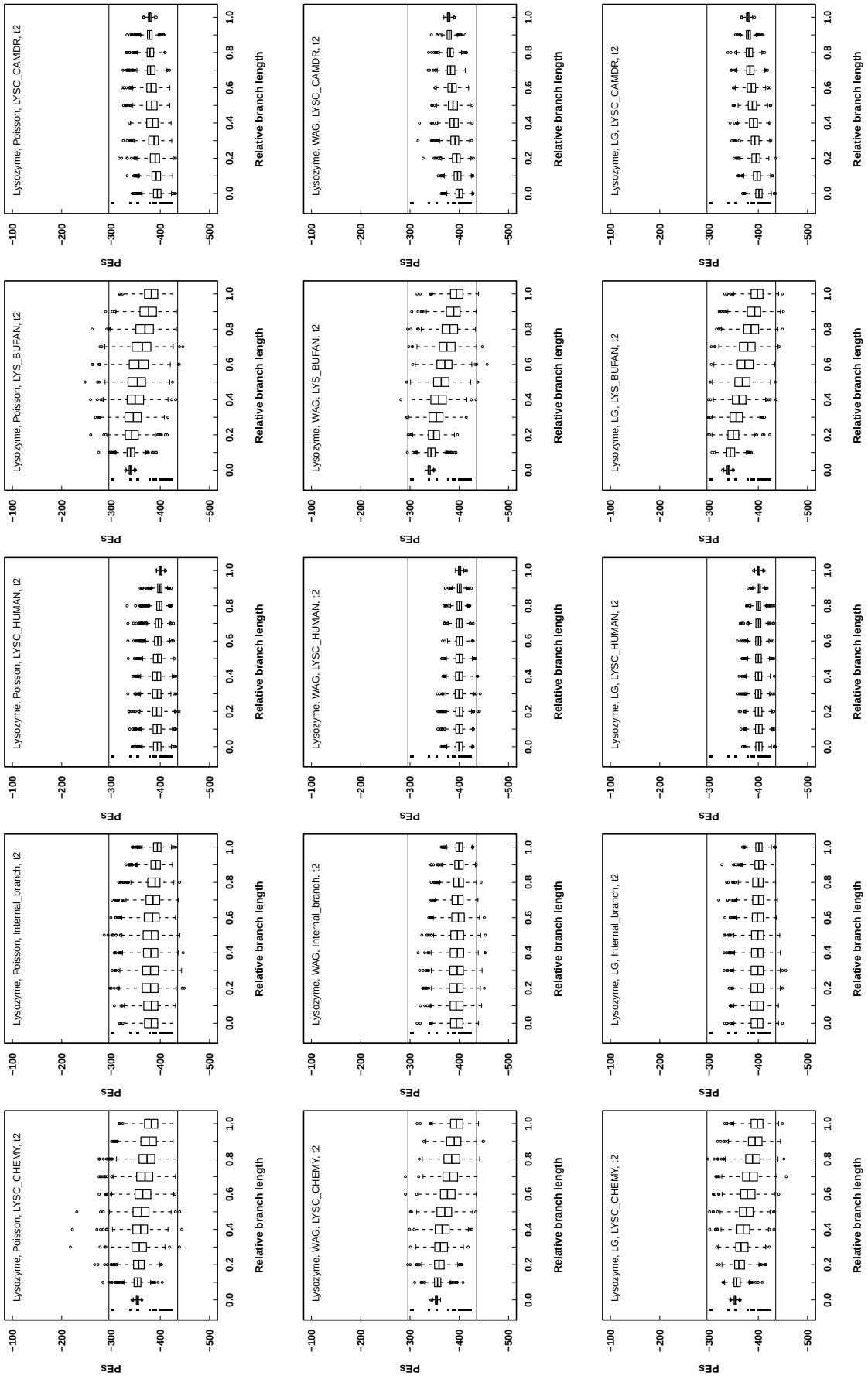

Figure 15. Lysozyme, PE_s, Tree 2

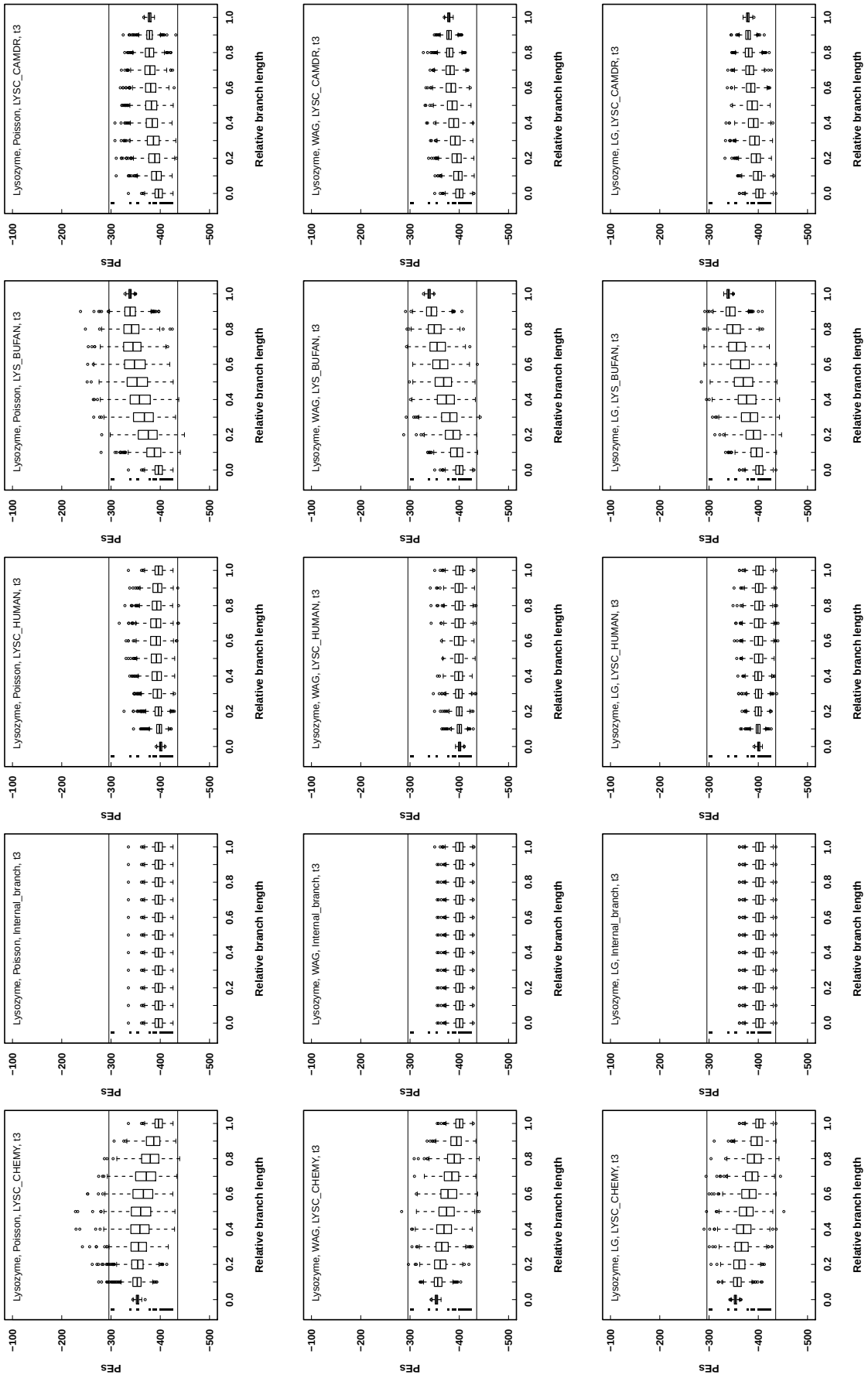

Figure 16. Lysozyme, PE_s, Tree 3

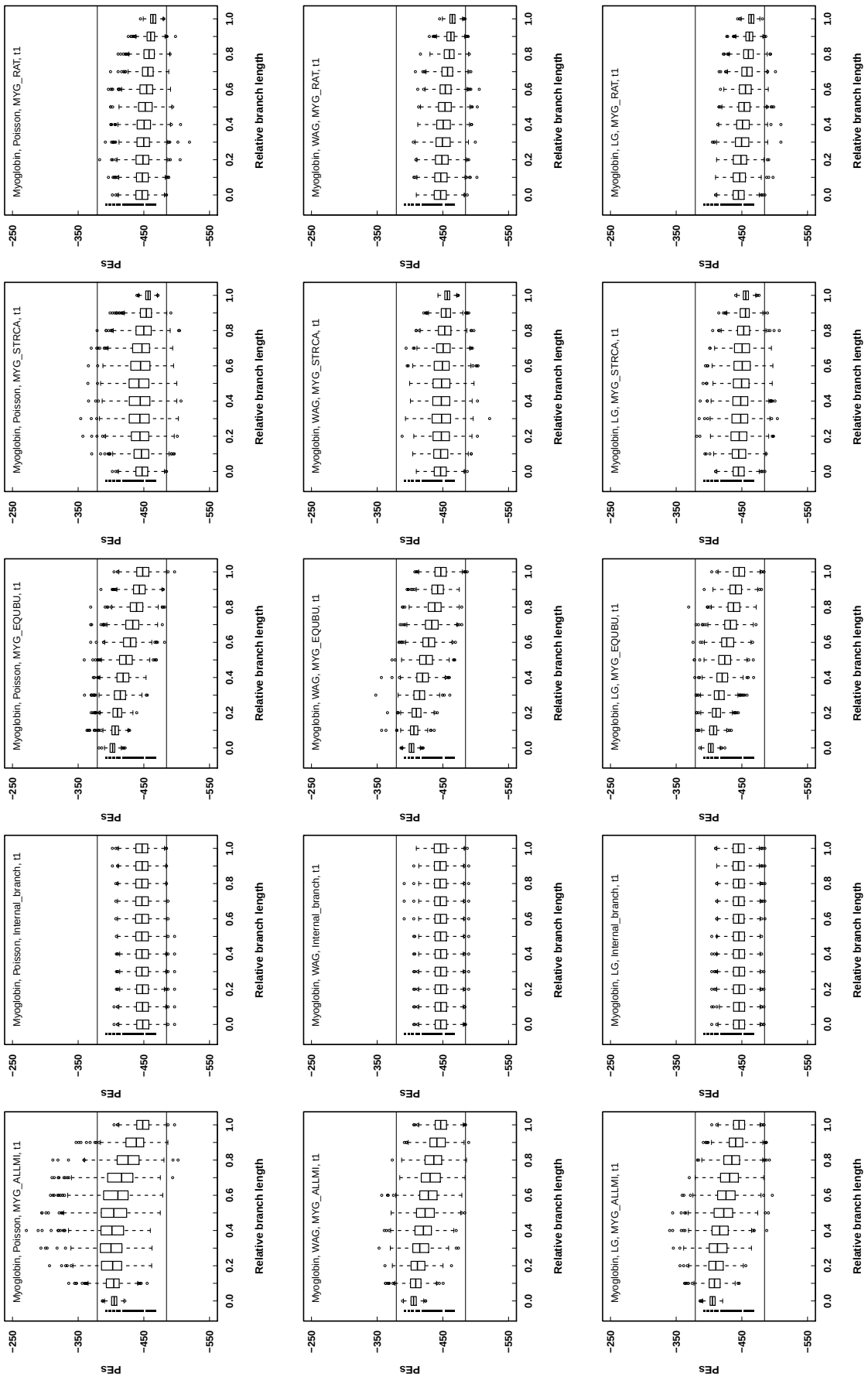

Figure 17. Myoglobin, PE_s, Tree 1

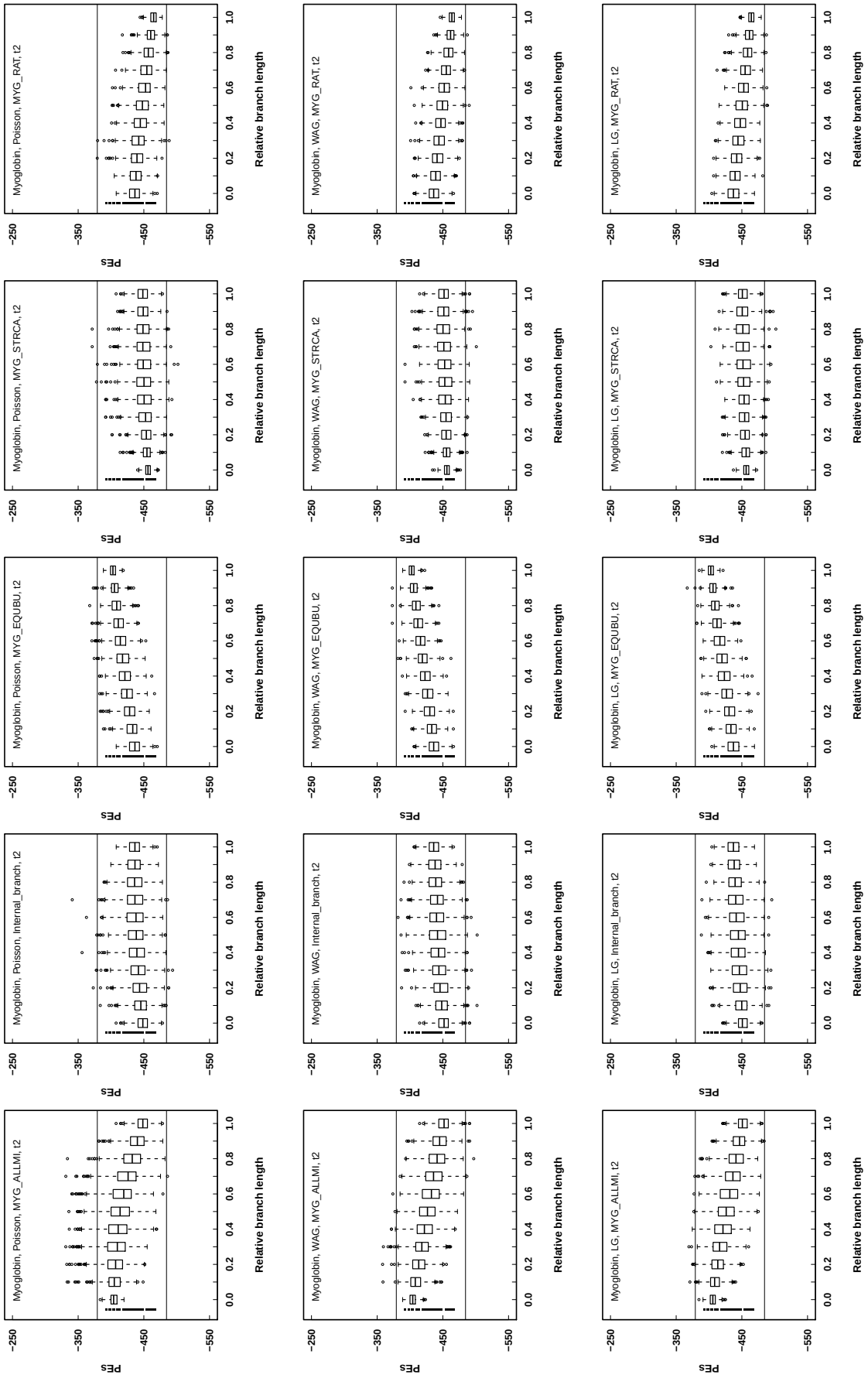

Figure 18. Myoglobin, PE_s, Tree 2

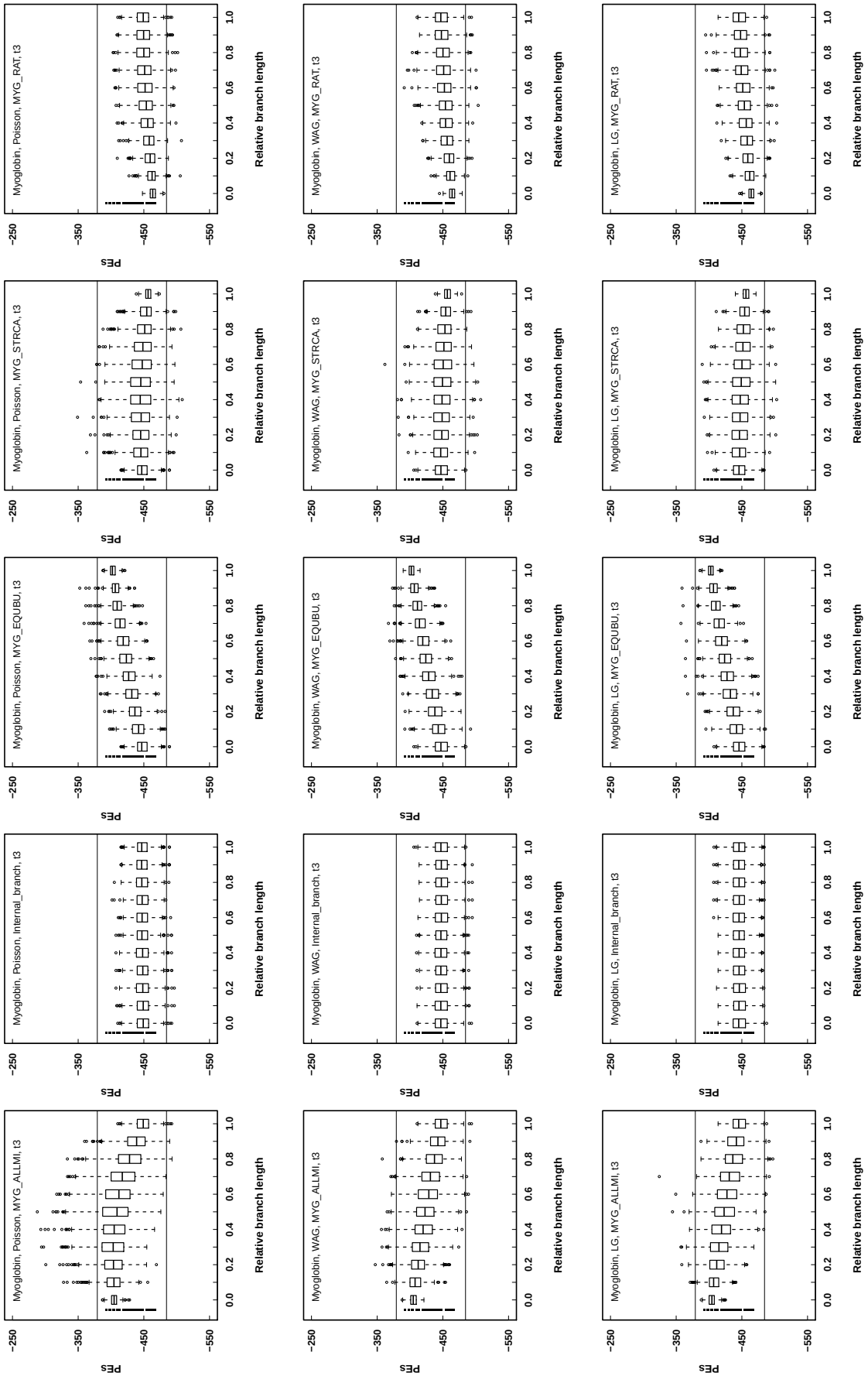

Figure 19. Myoglobin, PE_s, Tree 3

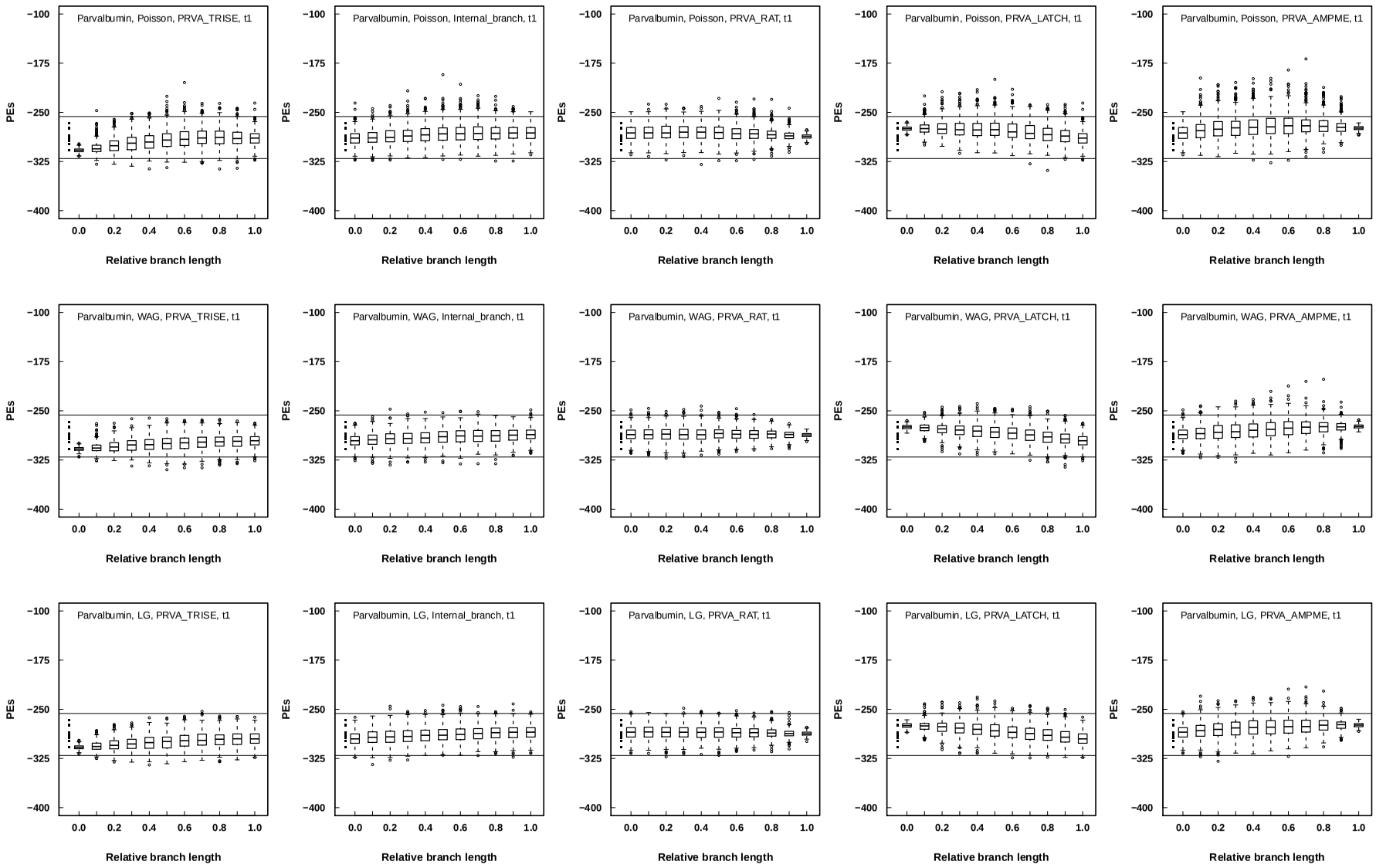

Figure 20. Parvalbumin, PE_s, Tree 1

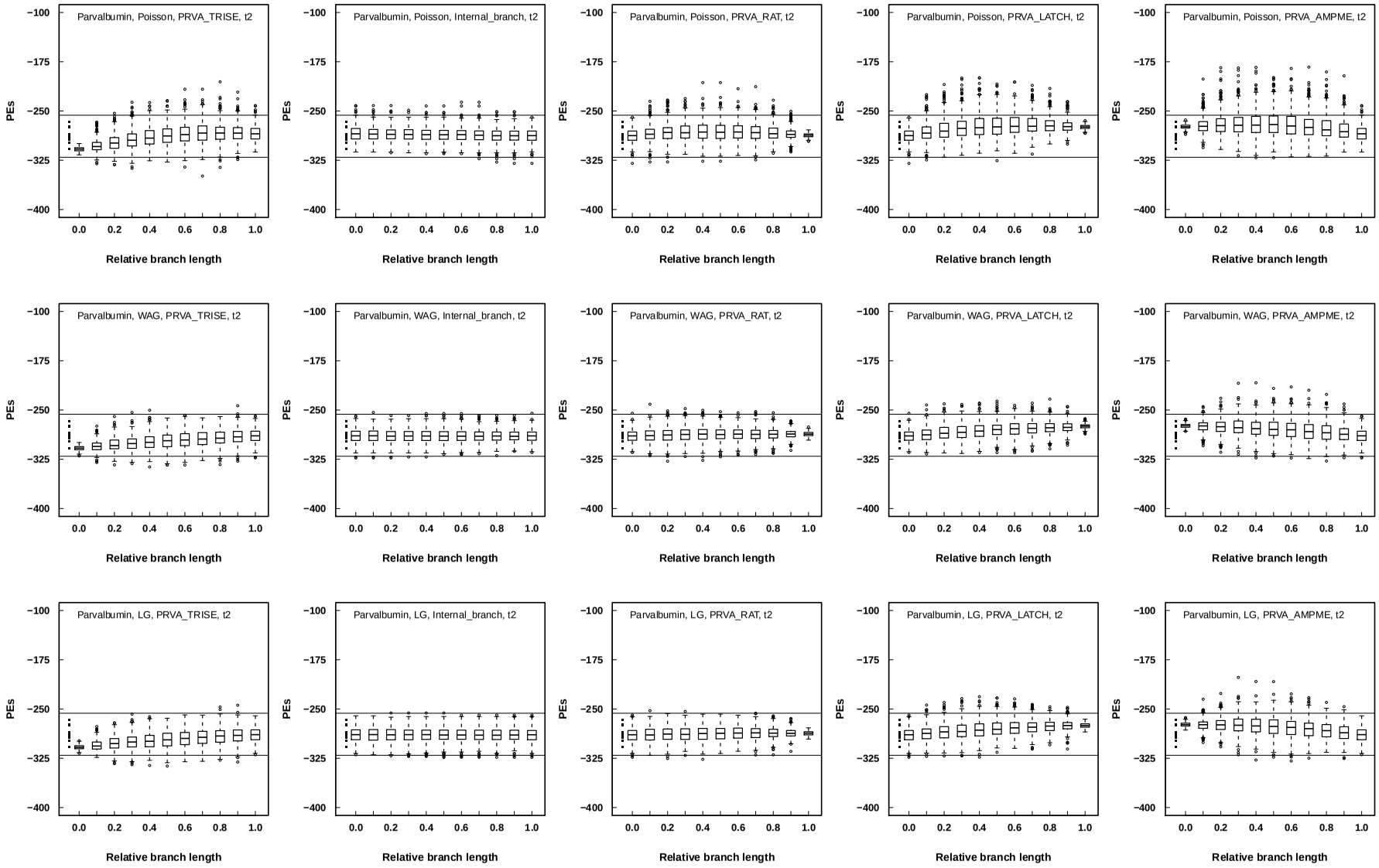

Figure 21. Parvalbumin, PE_s, Tree 2

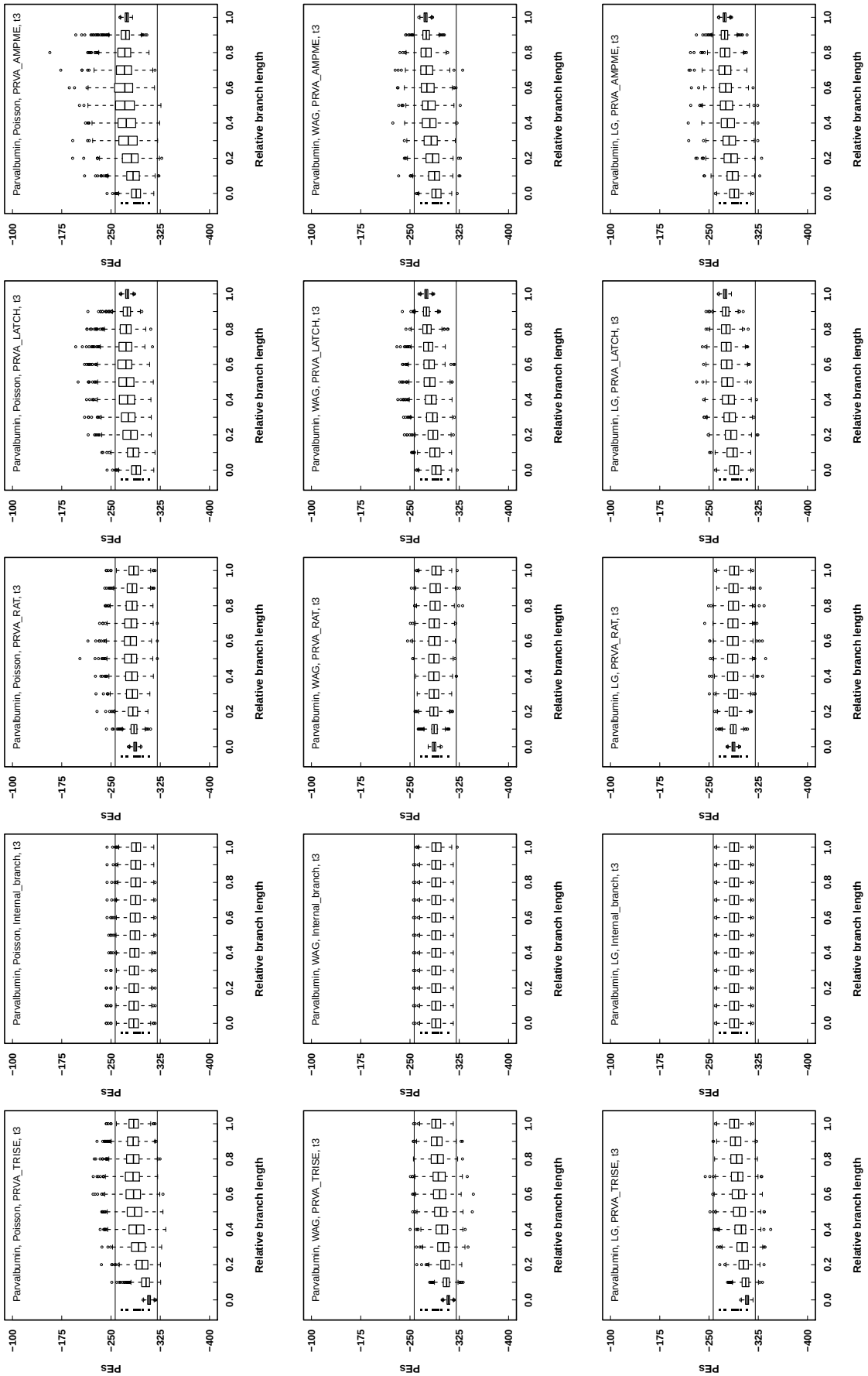

Figure 22. Parvalbumin, PE_s, Tree 3

b) Z_s

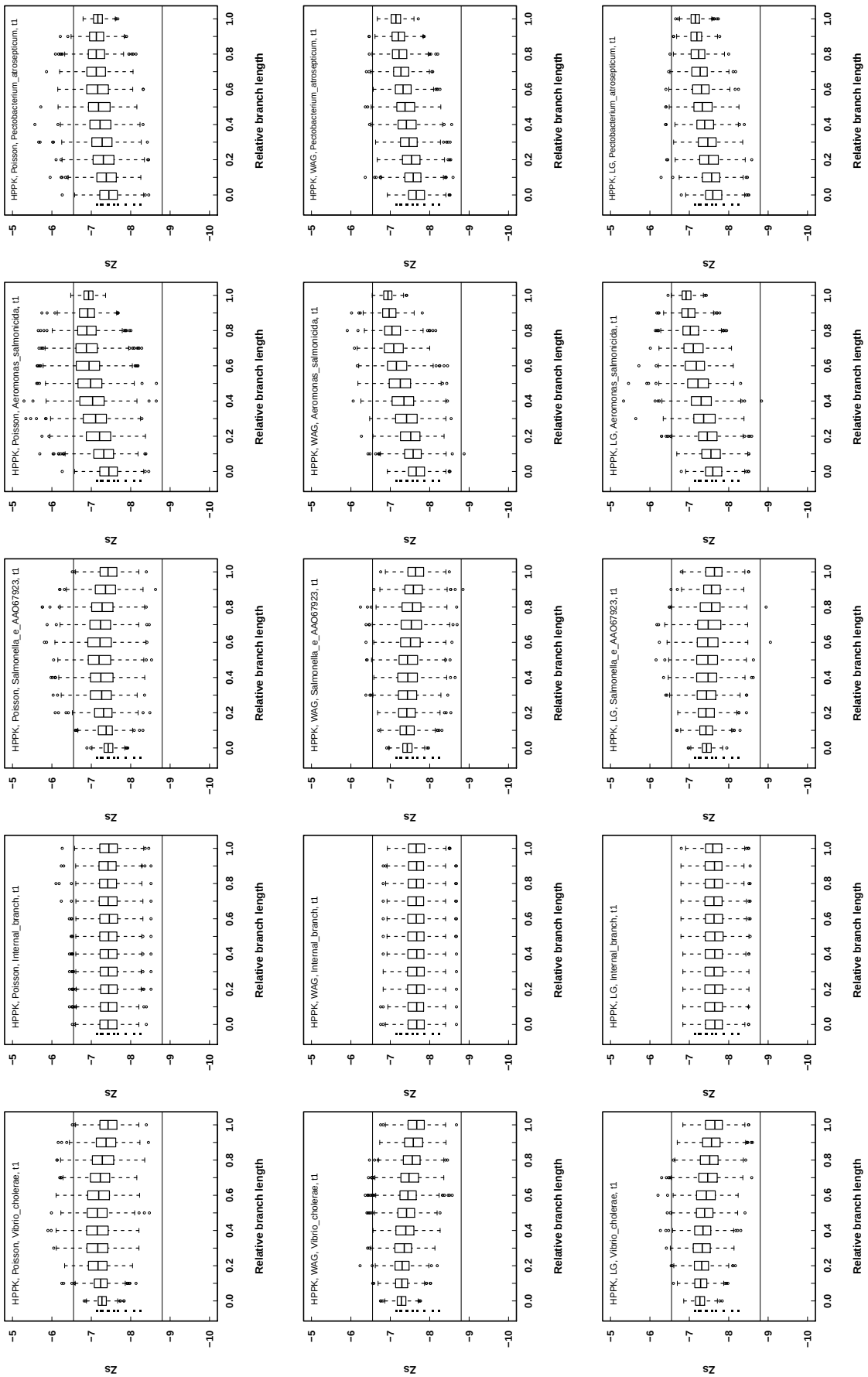

Figure 23. HPPK, Z_s, Tree 1

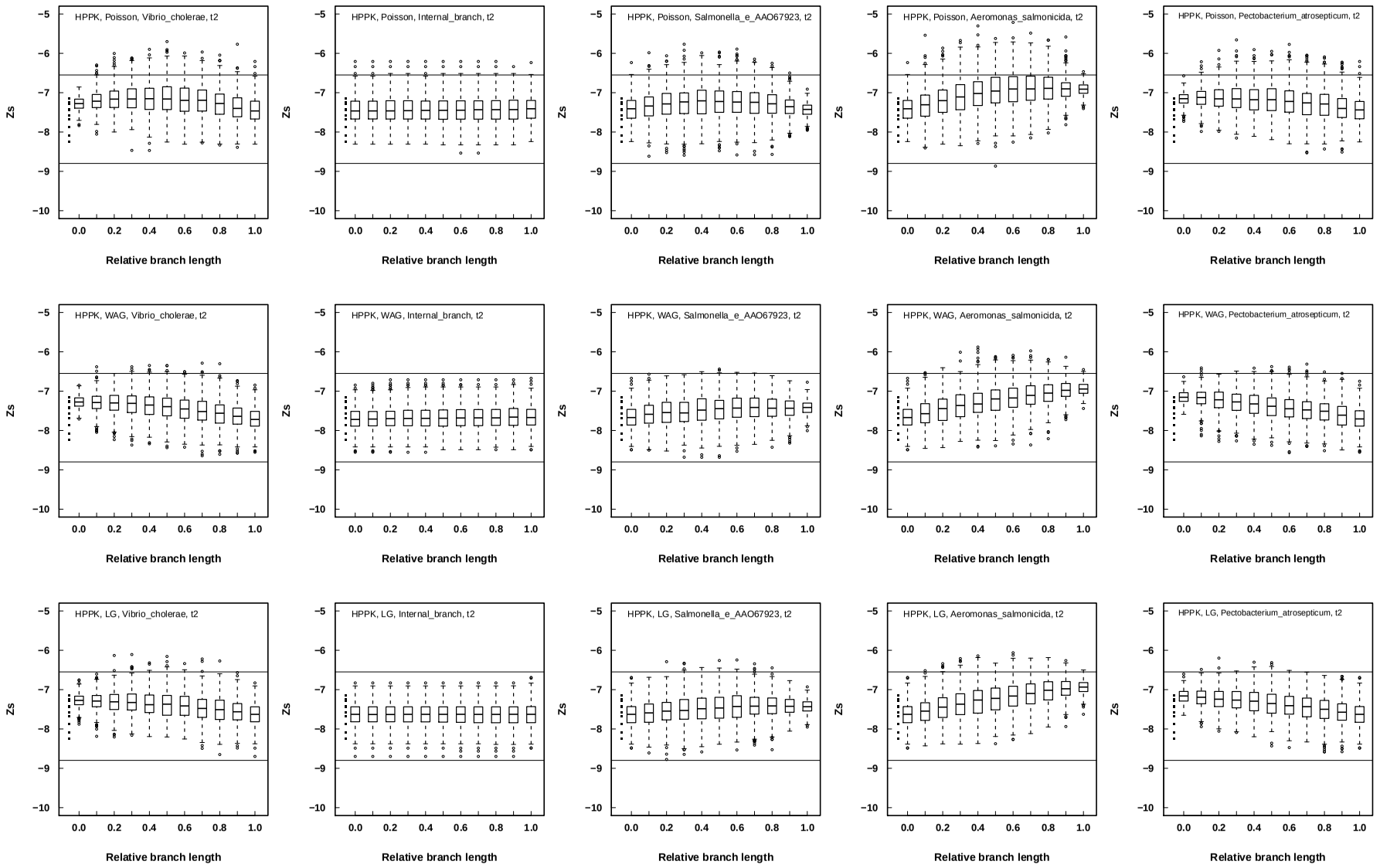

Figure 24. HPPK, Z_s, Tree 2

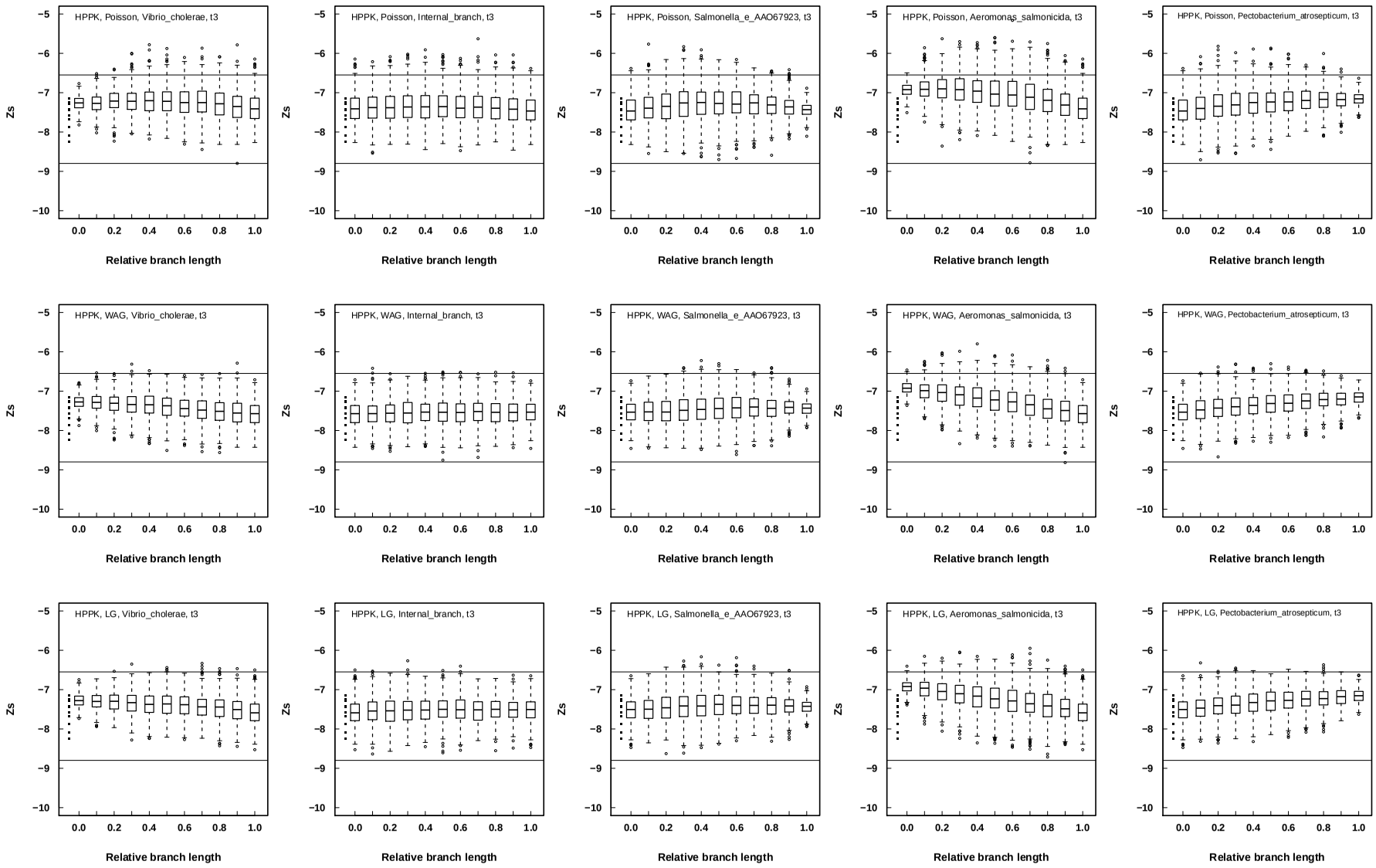

Figure 25. HPPK, Z_s, Tree 3

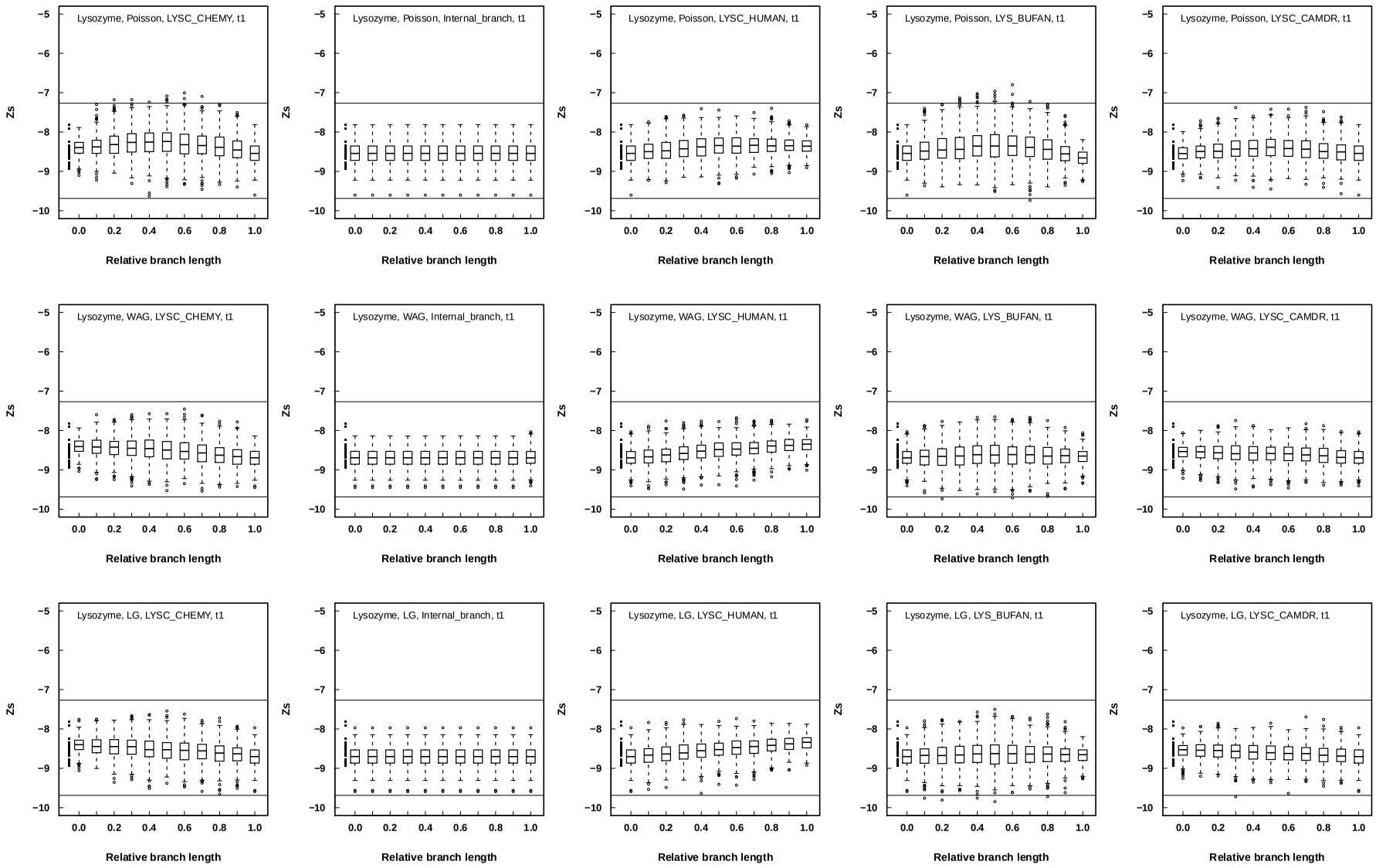

Figure 26. Lysozyme, Z_s, Tree 1

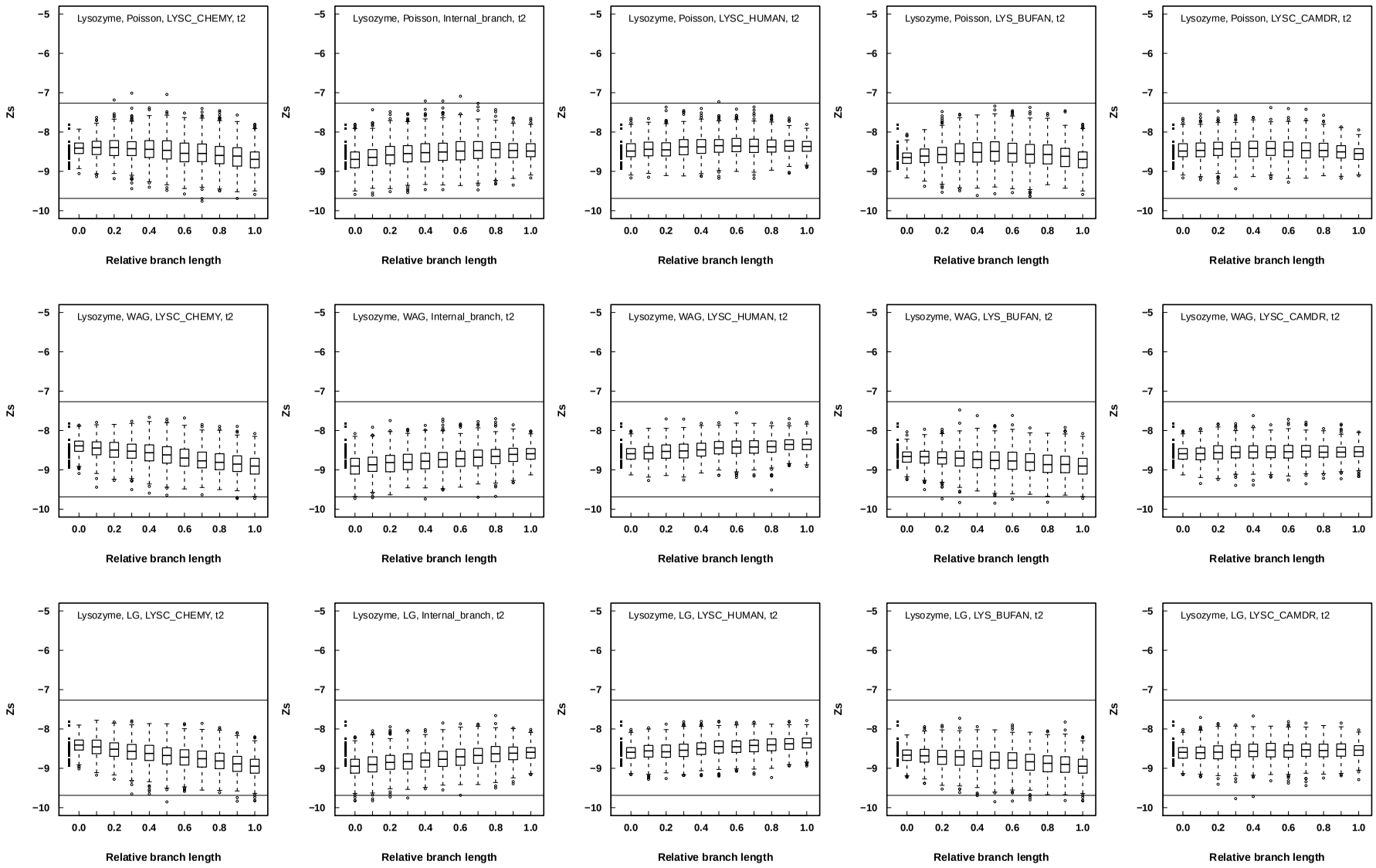

Figure 27. Lysozyme, Z_s, Tree 2

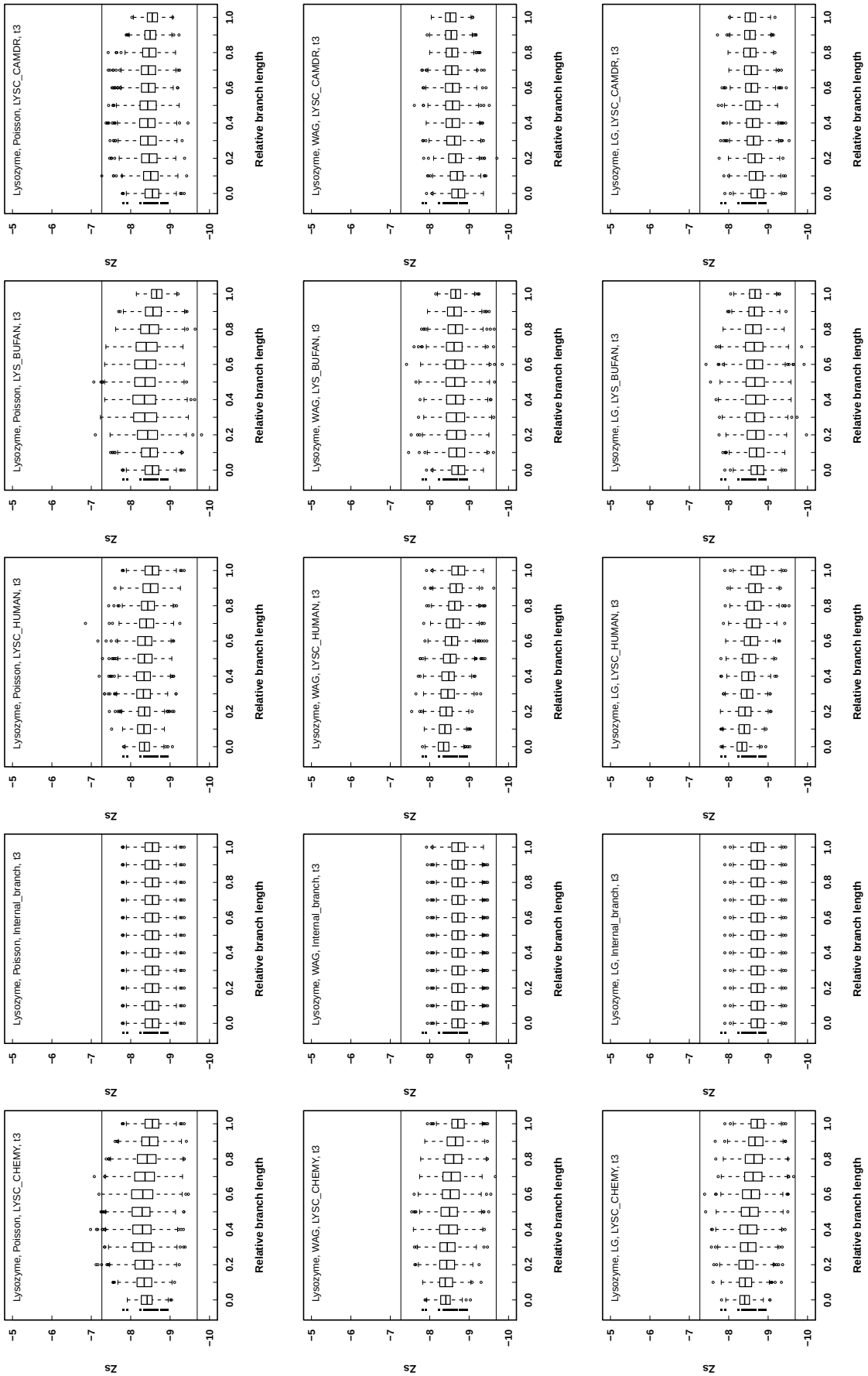

Figure 28. Lysozyme, Z_s, Tree 3

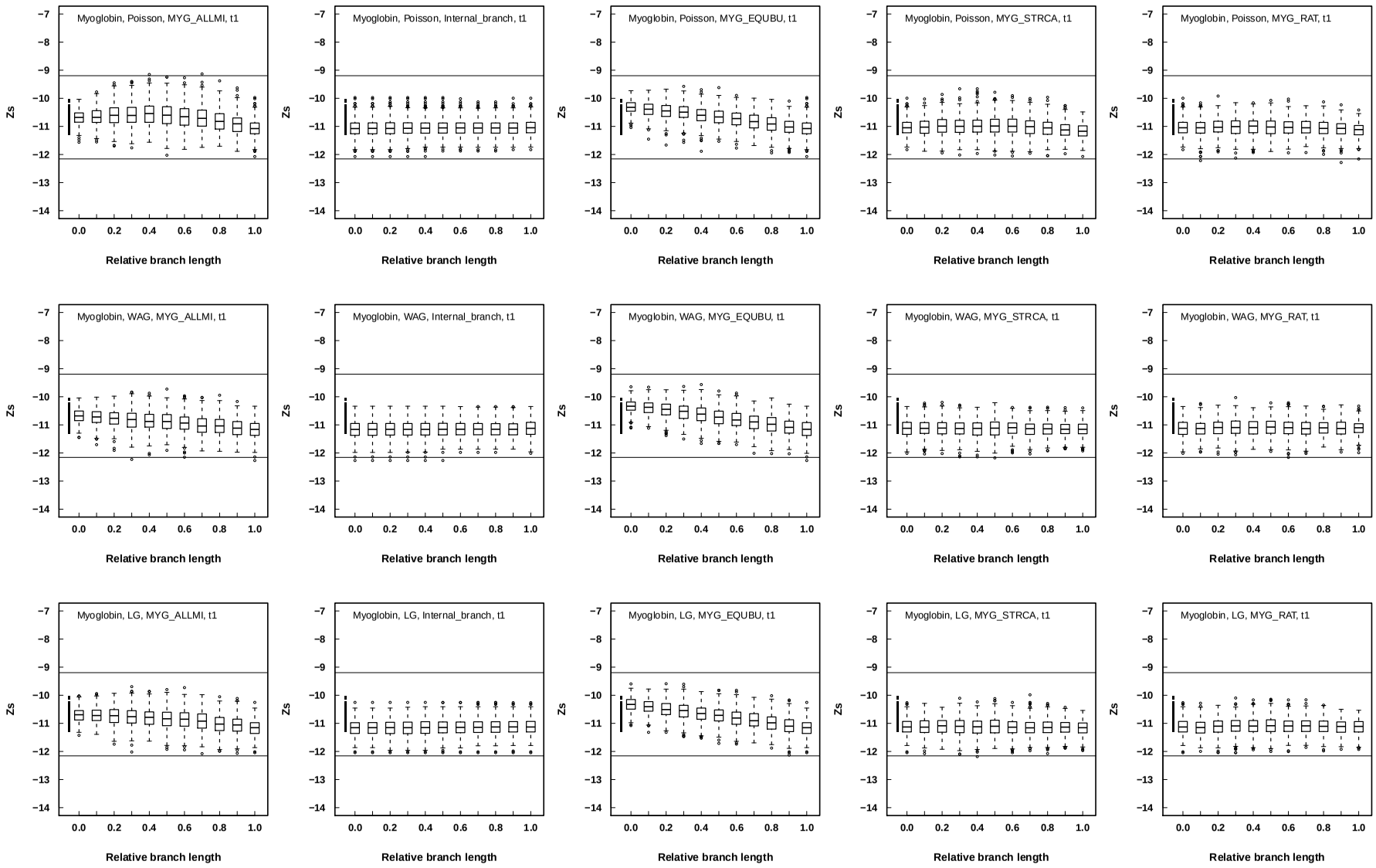

Figure 29. Myoglobin, Z_s, Tree 1

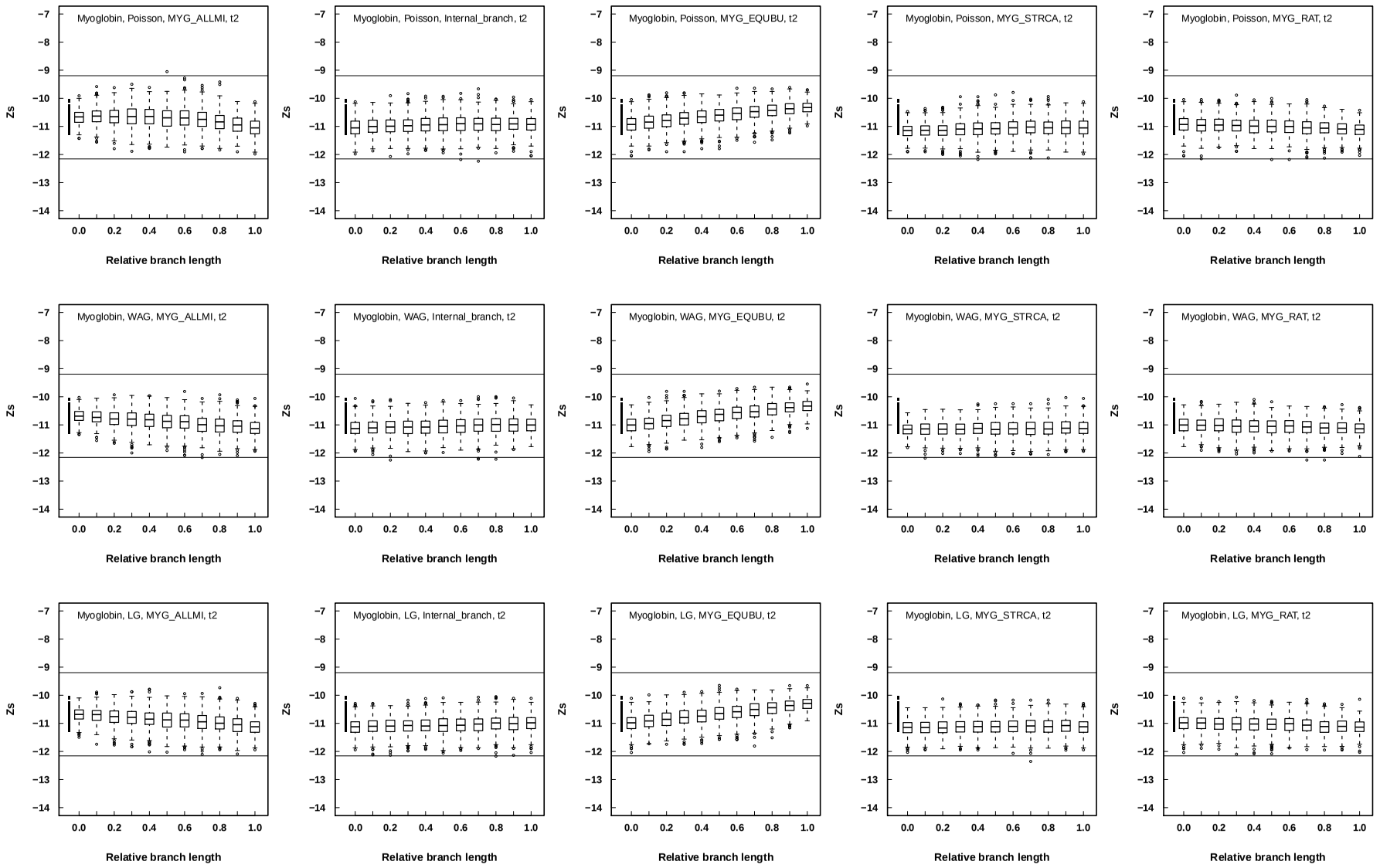

Figure 30. Myoglobin, Z_s, Tree 2

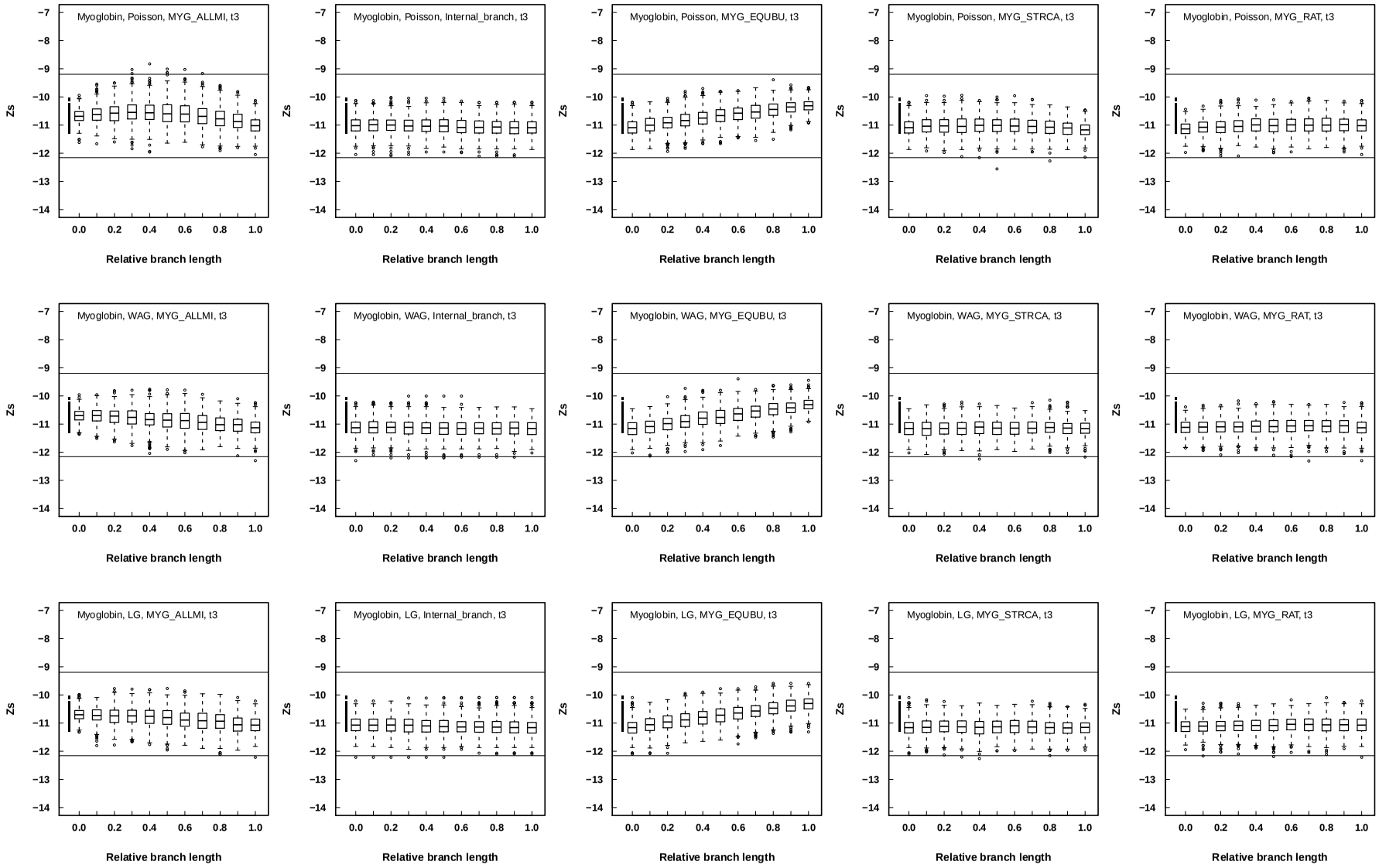

Figure 31. Myoglobin, Z_s, Tree 3

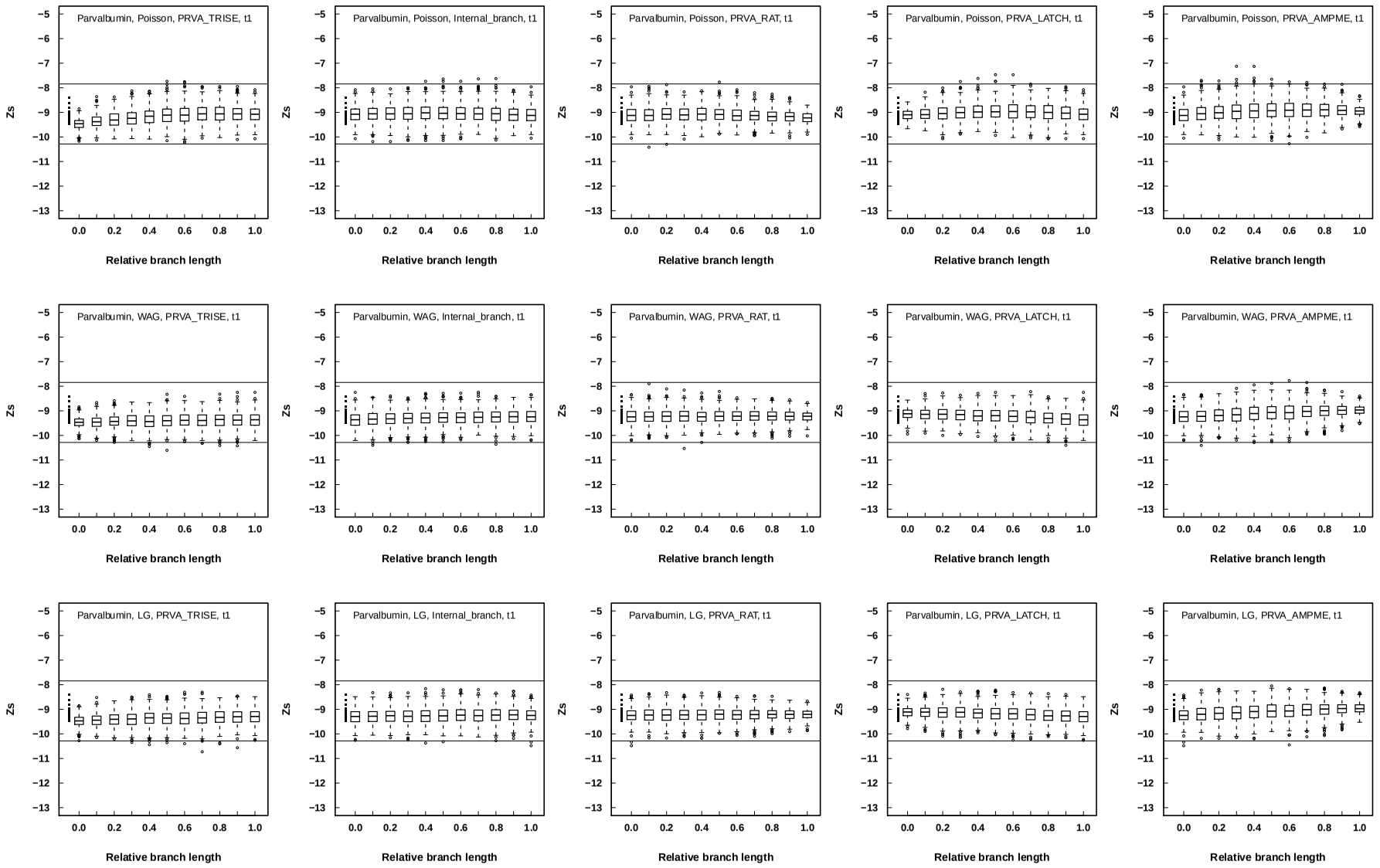

Figure 32. Parvalbumin, Z_s, Tree 1

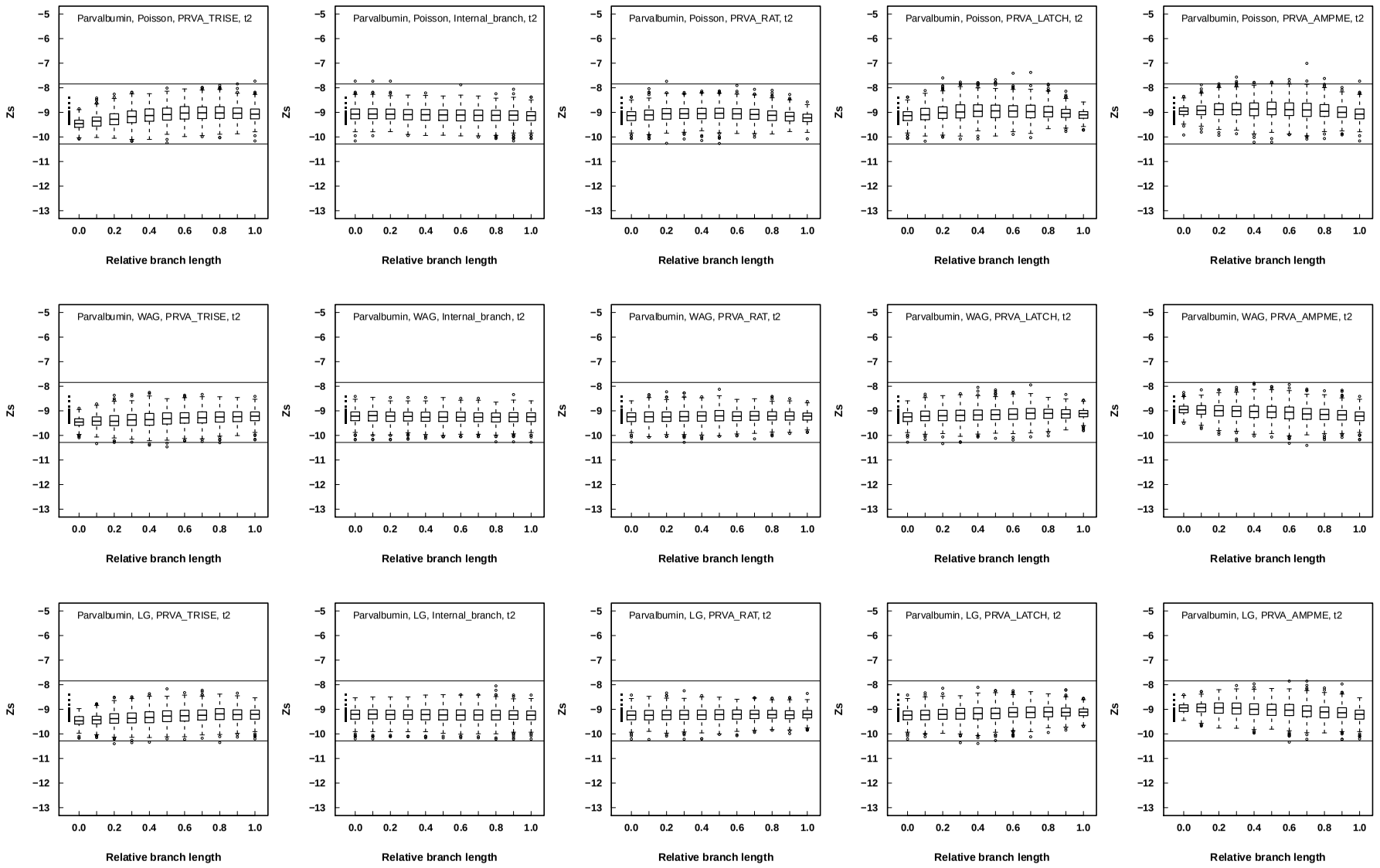

Figure 33. Parvalbumin, Z_s, Tree 2

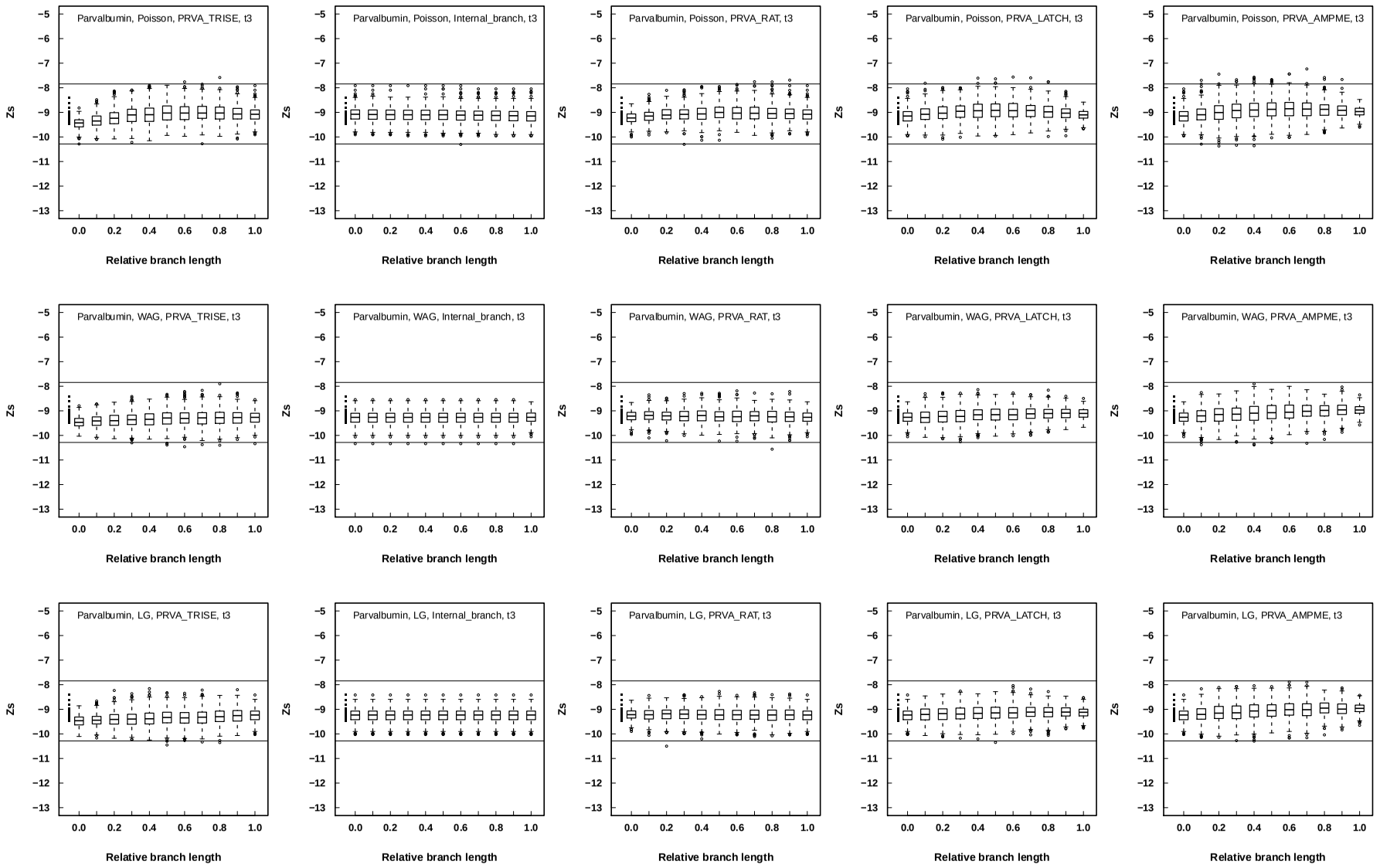

Figure 34. Parvalbumin, Z_s, Tree 3

Rate variation (discrete Γ)

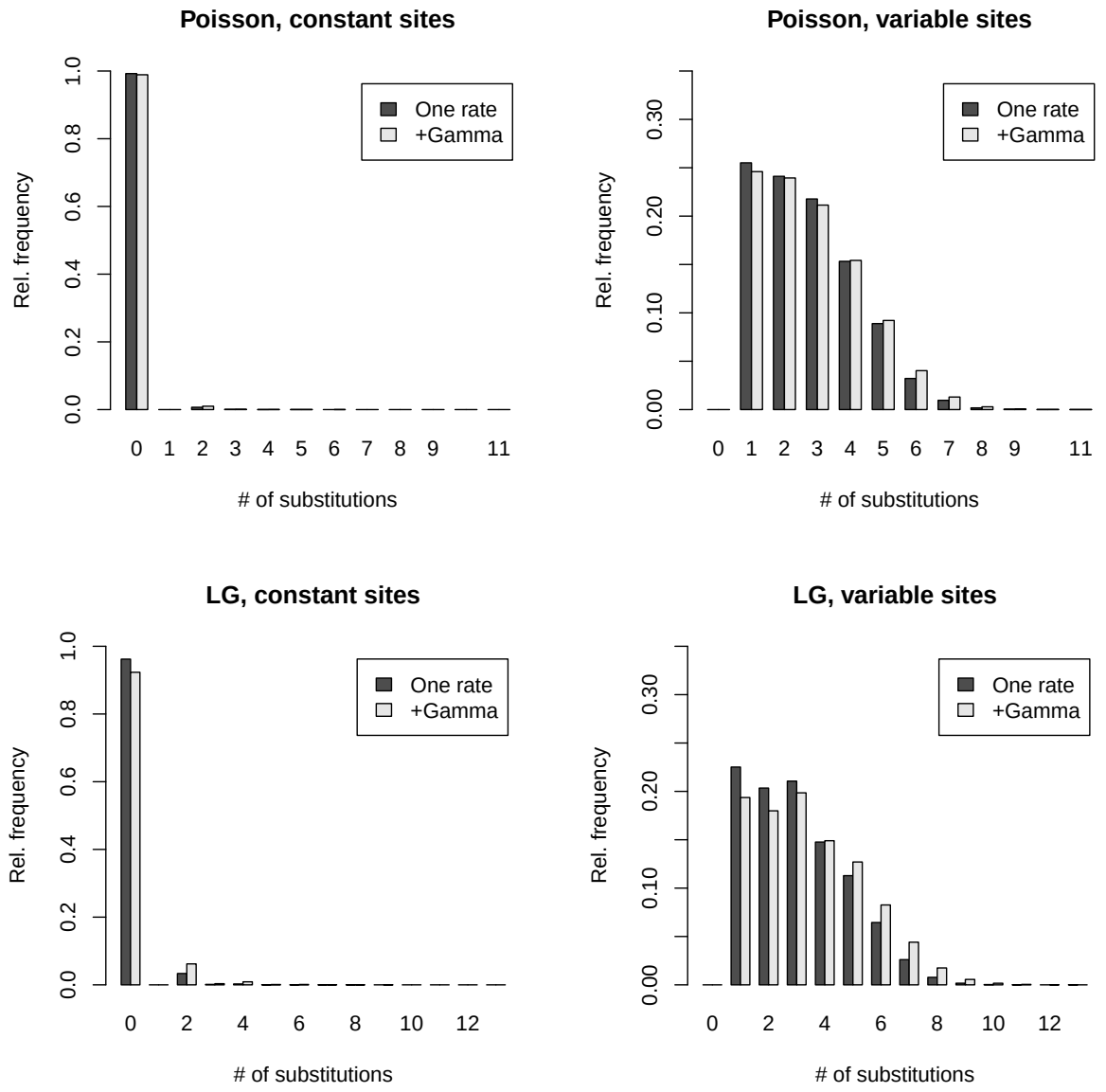

Figure 35. HPPK, 4 taxa, tree 3. Comparison of the number of substitutions by site type (conserved across all four taxa or variable) with and without rate heterogeneity. Paths were sampled with site-specific rates, estimated with PAML (using 30 categories, Yang, 2007). Rate variation led to increased branch lengths and more multiple substitutions. It was demonstrated in the paper that, for our data sets, decrease in structural compatibility was mainly associated with amino acids that were not observed in the terminals. This may be an explanation why we observed slightly more sequences with worse sequence:structure fit when rate heterogeneity was used. Results were similar for all trees and proteins.

References

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**(8), 1586–91.